Traitement du signal et des images

Traitement du signal et des images

Actes du 20^{e} colloque GRETSI

Louvain-la-Neuve (Belgique)

6–9 septembre 2005

Volume II

UCL PRESSES UNIVERSITAIRES DE LOUVAIN

© *Presses universitaires de Louvain, 2005*
Dépôt légal : D/2005/9964/21
ISBN : 2-87463-002-0
Imprimé en Belgique
Tous droits de reproduction, d'adaptation ou de traduction, par quelque procédé que ce soit, réservés pour tous pays, sauf autorisation de l'éditeur ou de ses ayants droit.

Couverture : Isabelle Sion

Diffusion : www.i6doc.com, l'édition universitaire en ligne

Sur commande en librairie ou à
Diffusion universitaire CIACO
Grand-Place, 7
1348 Louvain-la-Neuve, Belgique
Tél. 32 10 47 33 78
Fax 32 10 45 73 50
duc@ciaco.com

Table des Communications

Volume 1

Session Affiches : *P3.1 "Images : analyse et modèles"*

"Suivi de points robuste aux spécularités et aux changements d'éclairage" **1-4**

Michèle Gouiffès, Christophe Collewet, Christine Fernandez-maloigne, Alain Trémeau

"Suivi temps réel de personnes dans des séquences d'images couleur" **5-8**

Ghilès Mostafaoui, Catherine Achard, Maurice Milgram

"IDENTIFICATION PAR ANALYSE EN PAQUETS D'ONDELETTES DE L'IRIS ET TESTS DE ROBUSTESSE" **9-12**

Florence Rossant, Thomas Ea, Frédéric Amiel, Manuel Torres Eslava, Amara Amara

"Diagonalisation de matrices polynomiales quaternioniques : application à la séparation de mélanges convolutifs d'ondes polarisées" **13-16**

Nicolas Le Bihan

"Vers une classification non supervisée basée sur un nouvel indice de connectivité" **17-20**

Frédéric Blanchard, Michel Herbin, Philippe Vautrot

"LISSAGE ET SEGMENTATION D'IMAGES MULTI-PARAMETRIQUES ULTRASONORES PAR UNE APPROCHE « MEAN SHIFT »" **21-24**

Franck Davignon, Thomas Grenier, Chantal Revol-muller, Gérard Gimenez, Olivier Basset

"Représentation éparse d'images par ondelettes fovéales : principes et applications" **25-28**

Julien Ros, Christophe Laurent

"Filtrage Perceptuel Pyramidal" **29-32**

Azeddine Beghdadi, Razvan Iordache, Abdel-ouahab Abdel-ouahab

"Suivi Tridimensionnel en Stéréovision" **33-36**

Simon Conseil, Salah Bourennane, Lionel Martin

"Les representations "steerable" pour la compression et l'extraction de caracteristiques de bas niveau" **37-40**

Francois Tonnin, Patrick Gros, Christine Guillemot

Session Affiches : *P3.3 "Filtrage adaptatif et traitement d'antenne"*

"Filtrage adaptatif non-linéaire par méthodes de gradient stochastique court-terme à noyau" **41-44**

Cédric Richard

"Algorithmes du module constant en filtrage adaptatif :impact des progrès récents" **45-48**
Maurice Bellanger

"Evaluation de la précision des Algorithmes de Projection Affine en Virgule Fixe" **49-52**
Romuald Rocher, Daniel Ménard, Olivier Sentieys, Pascal Scalart

"MEAN SHIFT ADAPTATIF POUR LE LISSAGE D'IMAGES ULTRASONORES" **53-56**
Thomas Grenier, Franck Davignon, Chantal Revol-muller, Olivier Basset, Gérard Gimenez

"Decision Directed Algorithms for Blind Equalization Based on Constant Modulus Criteria" **57-60**
Carlos Alexandre Rolim Fernandes, João Cesar Moura Mota, Gérard Favier

"Etude des performances du modèle quaternionique en traitement d'antenne vectorielle" **61-64**
Sebastian Miron, Nicolas Le Bihan, Jérôme I. Mars

"Estimation des paramètres de propagation d'un canal MIMO par l'algorithme ESPRIT-Unitaire multidimentionnel" **65-68**
Steve Bartelmaos, Karim Abed-meraim, Valery Guillet, Jean-marie Chaufray

"Bornes de Cramer Rao de DOA pour signaux BPSK et QPSK en présence de bruit non uniforme" **69-72**
Habti Abeida, Jean Pierre Delmas

"Méthode Haute Résolution large bande pour la Localisation d'Objets Enfouis" **73-76**
Zineb Saidi, Salah Bourennane

"Estimation de directions d'arrivée en présence d'erreur de phase" **77-80**
Zineb Saidi, Salah Bourennane, Abdel-ouahab Boudraa, Délphine Dare-emzivat

"Blind Multiuser Equalization using a PARAFAC-Subspace Approach" **81-84**
André L. F. De Almeida, Gérard Favier, Joao C. M. Mota

"Conception et implémentation d'un système de localisation de source sonore large bande par formation de voie pour des applications robotiques" **85-88**
Sylvain Argentieri, Patrick Danes, Philippe Soueres

Session Affiches : *P4.3 "Détection"*

"Détection sous optimale en contexte alpha-stable : performances et algorithmes" **89-92**
Steeve Zozor, Jean-marc Brossier, Pierre-olivier Amblard, Cédric Duchêne

"Approche a contrario pour la détection de changements à partir d'images basse résolution" **93-96**
Amandine Robin, Lionel Moisan, Sylvie Le Hégarat-mascle

"Détection des zones de mouvement et régularisation : Application à la vidéo surveillance" **97-100**
Lionel Carminati, Jenny Benois Pineau

"Transformée de Radon Lissée et utilisation dans une procédure de détection en contrôle non destructif" **101-104**
Patrice Aknin, Patrice Simard, Hervé Cygan

"Segmentation Conjointe de Données Astrophysiques à l'aide d'une Analyse Bayésienne Hiérarchique" **105-108**
Nicolas Dobigeon, Jean-yves Tourneret

"DETECTION SOUS-OPTIMALES ET QUANTIFIEURS STOCHASTIQUES" **109-112**
Cédric Duchêne, Steeve Zozor, Pierre Olivier Amblard

"Algorithme de Synthèse d'Ouverture Basé sur un Détecteur à Sous Espace" **113-116**
Rémi Durand, Guillaume Ginolhac, Philippe Forster, Laetitia Thirion, Marc Lesturgie

"Utilisation des Statistiques d'Ordres Supérieurs en Contrôle Qualité de Détecteurs de Rayons X" **117-120**
Frédéric Maussang, Jocelyn Chanussot

"PROPRIETE CFAR-MATRICE DU DETECTEUR BORD - APPLICATION RADAR SUR SIGNAUX EXPERIMENTAUX NON-GAUSSIENS" **121-124**
Frédéric Pascal, Jean-philippe Ovarlez, Philippe Forster, Pascal Larzabal

"Un théorème limite et un test pour la détection non paramétrique de signaux dans un bruit blanc gaussien de variance inconnue" **125-128**
Dominique Pastor

Session Affiches : *P4.4 "Estimation"*

"Restauration d'images dans le cas d'un bruit Gaussien Poissonnien" **129-132**
Céline Theys, Henri Lantéri

"Inversion de données infrarouges issues du télescope SPITZER" **133-136**
Thomas Rodet, Alain Abergel, Hervé Dole, Alain Coulais

"Approche de Suivi d'Objet par Courbes de Niveau" **137-140**
Ali Ganoun, Raphael Canals, Sylvie Treuillet

"Une nouvelle approche des bornes Bayésiennes" **141-144**
Alexandre Renaux, Philippe Forster, Pascal Larzabal

"Estimation robuste de l'orientation locale par l'opérateur IRON" **145-148**
Franck Michelet, Christian Germain, Pierre Baylou

"Estimation semi-paramétrique robuste par minimisation de l'entropie des résidus, application en traitement d'images" **149-152**
Eric Wolsztynski, Eric Thierry, Luc Pronzato

"Régressions par machines à vecteurs supports pour la prédiction de séries chaotiques" **153-156**
Herwig Wendt, Patrick Flandrin, Patrice Abry

"Estimation parametrique multiresolution de mouvements transparents" **157-160**
Vincent Auvray, Patrick Bouthemy, Jean Lienard

"Estimation des données occultées : modèles statistiques à noyau" **161-164**
Barbara Romaniuk, Michel Desvignes, Maxime Bérar

"Estimation imprécise de densité de probabilité par transfert imprécis de comptage." **165-168**
Olivier Strauss

"Probabilité de décrochement d'un estimateur autodidacte du résidu de porteuse" **169-172**
Philippe Ciblat, Mounir Ghogho

"Estimation des dérivées d'un signal multidimensionnel avec applications aux images et aux vidéos" **173-176**
Michel Fliess, Cédric Join, Mamadou Mboup, Alexandre Sedoglavic

"Récepteur aveugle basé sur la décomposition PARAFAC pour des canaux de communication non linéaires" **177-180**
Alain Kibangou, Gérard Favier, Moha Hassani

Session Orale : *O6.1 "Applications : bio-ingénierie"*

"Sélection de segments pertinents dans des enregistrements SEMG en conditions expérimentales perturbées" **181-184**
Georges Farah, David Hewson, Jacques Duchêne

"Prédiction précoce de la syncope chez l'homme par réseaux de neurones" **185-188**
Mathieu Feuilloy, Daniel Schang, Jacques-olivier Fortrat, Sylvain Poggi, Elisabeth Bellard, Pascal Nicolas

"SEGMENTATION COMPETITIVE DE L'HIPPOCAMPE ET DE L'AMYGDALE A PARTIR DE VOLUMES IRM" **189-192**
Marie Chupin, Dominique Hasboun, Eric Bardinet, Sylvain Baillet, Line Garnero

"Compensation de déformations en tomographie dynamique" **193-196**
Laurent Desbat, Sébastien Roux, Pierre Grangeat

"Modélisation Markovienne pour l'Estimation Combinée de Forme et de Mouvement : Application au Coeur en Imagerie Scanner Multibarrette." **197-200**
Antoine Simon, Mireille Garreau, Jean-louis Coatrieux, Hervé Le Breton

Session Orale : *O4.4 "Filtrage particulaire"*

"A propos de l'utilisation des méthodes de Monte Carlo séquentielles pour l'estimation de paramètres dans les modèles de Markov cachés" **201-204**

Olivier Cappé, Eric Moulines

"Détection et suivi simultanés dans une séquence vidéo en couleur par filtres particulaires" **205-208**

Jacek Czyz, Branko Ristic, Benoit Macq

"Méthode de modélisation non supervisée du canal satellite mobile par chaînes de Markov" **209-212**

Clémence Alasseur, Lionel Husson, Sandro Scalise, Harald Ernst

"Filtrages de Kalman et particulaire dans les Chaînes de Markov Triplet" **213-216**

Boujemaa Ait-el-fquih, François Desbouvries

"Fusion de capteurs potentiellement défaillants par filtrage particulaire" **217-220**

François Caron, Manuel Davy, Emmanuel Duflos, Philippe Vanheeghe

---

Session Orale : *O3.1 "Analyse multirésolution"*

"Bases orthonormées d'ondelettes splines linéaires adaptées à une grille irrégulière" **221-224**

Najat Chihab, Anissa Zergaïnoh, Pierre Duhamel, Jean-pierre Astruc

"Multiscale analysis of multiple change points using unbalanced wavelets" **225-228**

Maarten Jansen

"Etude du bruit dans une analyse M-bandes en arbre dual" **229-232**

Caroline Chaux, Laurent Duval, Jean-christophe Pesquet

"Modélisation statistique multivariée des images dans le domaine de la transformée de curvelet" **233-236**

Larbi Boubchir, Jalal Fadili

"Analyse multirésolution L2-optimale : estimation par quasi-projections" **237-240**

Laurent Condat, Annick Montanvert

Session Spéciale : *S3.1 "Signal et Image en bioingénierie 1"*

"Sur les signaux électrophysiologiques : réflexion et quelques perspectives ouvertes" **241-244**

Lotfi Senhadji, Alfredo Hernandez, Fabrice Wendling, Guy Carrault

"Classification multivoie de signaux électromyographiques par ondelettes adaptées" **245-248**

Aude Maitrot, Marie-françoise Lucas, Christian Doncarli

"Etude de la cyclostationnarité du signal Doppler sanguin ultrasonore pour la détection de micro-emboles" **249-252**

Jean-marc Girault, Mathieu Biard, Denis Kouamé, Gregory Souchon, Aurore Bleuzzen, François Tranquart

"ANALYSE DES SIGNAUX POUR UN DISPOSITIF DE MESURE ET DE STIMULATION DU SYSTEME NERVEUX CENTRAL" **253-256**

Céline Moulin, Daniel Barbier, Blaise Ivert, Régis Guillemaud, Lionel Rousseau, François Goy

Session Spéciale : *S1.1 "Astro-statistique 1"*

"Détection de familles d'astéroïdes : un problème de segmentation non supervisée" **257-260**

Olivier J.j. Michel, Philippe Bendjoya, Pablo Rojo Guerra

"Géométrie des ensembles de chirps et détection des ondes gravitationnelles" **261-264**

Eric Chassande-mottin

"Analyse statistique de la détection de planètes par imagerie directe" **265-268**

André Ferrari, Jean Yves Tourneret, Eduard Serradel, Claude Aime

Session Spéciale : *S2.1 "Traitement du signal et télétrafic informatique 1"*

"Deux applications de processus ponctuels aux réseaux de communications" **269-272**

Olivier Dousse, Hung X. Nguyen, Patrick Thiran

"Traitement du signal et modélisation probabiliste du trafic Internet" **273-276**

Fabrice Guillemin

"Estimation en temps réel de la Matrice de Trafic par filtres de Kalman" **277-278**

Augustin Soule, Kavé Salamatian, Nina Taft

"Marginales non gaussiennes et longue mémoire : analyse et synthèse de trafic Internet" **279-282**

Antoine Scherrer, Patrice Abry

Session Spéciale : *S3.2 "Signal et Image en bioingénierie 2"*

"Quelques perspectives en imagerie médicale" **283-286**

Isabelle Bloch

"Modèle numérique d'estimation 2D de la déformation des tissus mous biologiques à partir d'images échographiques radiofréquence : Résultats sur simulations numériques et données expérimentales" **287-290**

Elisabeth Brusseau, Ghada Said, Patrick Clarysse

"Approximation de courbes de quaternions duaux par des courbes de Bézier et des splines" **291-294**

Gwenaël Guillard, Chafiaâ Hamitouche, Christian Roux

"Détection-estimation conjointe de l'activité cérébrale en imagerie fonctionnelle (IRMf)" **295-298**

Salima Makni, Philippe Ciuciu, Jérôme Idier, Jean-baptiste Poline

---

Session Spéciale : *S1.2 "Astro-statistique 2"*

"L'analyse des images dans les grands relevés d'astronomie visible/proche-infrarouge" **299-302**

Emmanuel Bertin

"Statistiques direction-multipôle pour la séparation de composantes astrophysiques dans le fonds de rayonnement cosmologique" **303-306**

Jean-françois Cardoso, Pierrick Abrial, Yassir Moudden, Jean-luc Stark, Jacques Delabrouille

"Deux méthodes de déconvolution de mélanges de composantes ; application à la reconstruction des amas de galaxies" **307-310**

Sandrine Anthoine, Elena Pierpaoli, Ingrid Daubechies

"Déconvolution avec contraintes de positivité et de support : sources ponctuelles + source étendue en astronomie" **311-314**

Jean-françois Giovannelli, Alain Coulais

Session Spéciale : *S2.2 "Traitement du signal et télétrafic informatique 2"*

"Revisiting the multiscaling hypothesis at medium timescales" **315-318**

Claudio Narduzzi, Paolo Attilio Pegoraro, Steve Uhlig

"Caractérisation et analyse du trafic Internet en fonction du type d'application" **319-322**

Nicolas Larrieu, Yu Zhang, Philippe Owezarski

"Détection d'attaques de "Déni de Services" : ruptures dans les statistiques du trafic" **323-326**

Pierre Borgnat, Nicolas Larrieu, Patrice Abry, Philippe Owezarski

---

Session Affiches : *P1.1 "Images"*

"Suivi des mouvements faciaux et de la pose 2D d'un visage" **327-330**

Soumya Hamlaoui, Franck Davoine

"Modélisation, Analyse et Fusion Markovienne Cachée de Multiples Séquences d'Evénements : Application en Cartographie IRM Fonctionnelle Cérébrale" **331-334**

Laurent Thoraval, Sylvain Faisan, Jean-paul Armspach

"Partitionnement optimal de l'image par minimisation d'un critère d'uniformité globale" **335-338**

Luc Klaine, Kacem Chehdi, Benoit Vozel

"Segmentation pyramidale et groupements perceptuels" **339-342**

Jérémy Huart, Pascal Bertolino

"Une méthode pour le pré-traitement et le calcul de matrices de crosstalk pour des images couleur issues d'une caméra tri-CCD" **343-346**

Alexandra Lathuilière, Yvon Voisin, Franck Marzani

"Traitement non linéaire uniforme pour la reconstruction d'une mosaïque chromatique répartie aléatoirement" **347-350**

Brice Chaix De Lavarène, David Alleysson, Jeanny Hérault

"Invariants intégraux pour l'indexation d'images omnidirectionnelles" **351-354**

Cyril Charron, Ouiddad Labbani-igbida, El Mustapha Mouaddib

Session Affiches : *P6.1 "Signal et images en bio-ingénierie 1"*

"Simulation d'images échographiques à partir de données scanner X" **355-358**

Soizic Laguitton, Jean-jacques Patard, Jean-louis Dillenseger

"Segmentation d'images échocardiographiques par contours actifs implicites : exploitation de descripteurs statistiques de régions" **359-362**

Olivier Bernard, Jan D'hooge, Denis Friboulet

"Estimation 3D du flux sanguin par imagerie ultrasonore haute fréquence : application à l'étude de la microcirculation" **363-366**

Walid Aoudi, Hervé Liebgott, Andrew Needles, Victor Yang, Stuart Foster, Didier Vray

"Evaluation de l'équilibre postural suite à la réalisation d'un pas selon trois méthodes d'estimation de l'exposant de Hurst" **367-370**

Hassan Amoud, Mohammed Abadi, Valérie Michel, David Hewson, Michel Doussot, Jacques Duchêne

"Reconstruction 3D de sténoses par modélisation markovienne à partir de plus de deux images d'angiographie rotative par rayons X" **371-374**

Moez Chakchouk, Sylvie Sevestre-ghalila, Christine Graffigne, Mériem Jaïdane, Hatem Rajhi, Radhi Hamza

"ICASENSE : une application de l'Analyse en Composantes Indépendantes à l'IRM parallèle" **375-376**

Gaël Le Bec, Kosai Raoof, Jean Paul Yonnet

"Modèle de connaissances pour l'interprétation d'un électrogramme" **377-380**

Antoine Defontaine, Alfredo Hernandez, Guy Carrault

"Proposition d'un indice de synchronisation pour la validation de réponses neuronales à des stimuli périodiques" **381-384**

Julien Huillery, Régine Le Bouquin Jeannès, Gérard Faucon, Pierre Dugué, Alexandre Guérin

"Mesure volumétrique à partir de vues stéréoscopiques non calibrées" **385-388**

Benjamin Albouy, Sylvie Treuillet, Yves Lucas

"Evaluation des déformations du myocarde sur des séquences temporelles d'images IRM par estimation contrainte du flot optique" **389-392**

Solofohery Randrianarisolo, Eric Deléchelle, Eric Petit, Yasmina Chenoune, Jérôme Garot, Alain Rahmouni

"Détection et caractérisation d'un bioprocédé fed-batch par l'analyse de l'exposant de Hölder" **393-396**

Sebastien Regis, Andrei Doncescu, Laetitia Faure, Jean-louis Uribelarea

---

Session Affiches : *P6.2 "Signal et images en bio-ingénierie 2"*

"Extraction des signatures de composés purs constituant la couche externe du grain d'orge à partir d'images de fluorescence" **397-400**

Abdelkamel Elhafid, Danielle Nuzillard, Marie-françoise Devaux, Nicolas Petrochilos, Fabien Belloir

"Application de modèles à dérivée non entière à la détection électrochimique sur biopuce" **401-404**

Alexandre Chibane, Pierre Grangeat, Laurent Desbat, Alina Voda

"Etude des parties codantes de l'ADN en fonction du sens lecture" **405-408**

Guillaume Sicot, Ramesh Pyndiah

"Caractérisation 3D d'images IRM pour l'analyse de l'ostéochondrite primitive de la hanche" **409-412**

Philippe Pouletaut, Isabelle Claude, Renaud Winzenrieth, Marie-christine Ho Ba Tho, Guy Sebag

"Caractérisation par analyse d'images de l'angiogenèse sur des coupes histologiques" **413-416**

Sophie Le Maire, Silvia Blacher, Jean-michel Foidart, Jacques Verly

"Réduction du bruit par pré-détection pour la localisation échographique radio fréquence de micro-outils dans les tissus biologiques" **417-420**

Jean Martial Mari, Christian Cachard

"Localisation distribuée des activités épileptiques par une analyse vectorielle : Approche expérimentale multi agent" **421-424**

Mamadou Ndiaye, Abel Kinié, Jean-jacques Montois, Fabrice Wendling

"Mesure de l'excitation et de l'inhibition dans le tissu neuronal en épilepsie par identification d'un modèle dynamique non linéaire d'activité EEG" **425-428**

Jean-jacques Bellanger, Alfredo Hernandez, Fabrice Wendling, Paul Frogerais

"Nouvelle approche de modélisation de milieux poreux. Application à l'os trabéculaire" **429-432**
Gabriel Aufort, Rachid Jennane, Rachid Harba, Alain Gasser, Damien Soulat, Claude Laurent Benhamou

"Etude des paramètres d'un modèle de génération de signaux EEG intracérébraux : synchronisation, conductivité cérébrale et surface corticale." **433-436**
Delphine Cosandier-rimélé, Jean-michel Badier, Fabrice Wendling

"Vers un codage numérique de l'activité Stéréo-EEG critique par une approche expérimentale multi-agent." **437-440**
Abel Kinie, Mamadou Ndiaye, Jean-jacques Montois, Fabrice Wendling, Yann Jacquelet

---

Session Affiches : *P6.3 "Applications"*

"Filtrage des échos ultrasonores en CND par la fonction de Cross Validation Généralisée (GCV)" **441-444**
Farid Tafinine, Boulem Ouazine, Amrane Houacine

"A novel approach in the Characterization of Correlated Measurements" **445-448**
Carlos Lopez-martinez, Eric Pottier

"Localisation de source par filtrage modal : application à des données petites échelles en ASM" **449-452**
Barbara Nicolas, Philippe Roux, Jean-louis Lacoume, Jérôme Mars

"Application d'opérateurs de fusion floue à la classification automatique de cibles réalisée par un radar polarimétrique" **453-456**
Emanuel Radoi, André Quinquis, Felix Totir

"Pollution de milieux poreux : identifiabilité et identification de modèles paramétriques de sources" **457-460**
Aurélien Hazart, Jean-françois Giovannelli, Stéphanie Dubost, Laurence Chatellier

"Calcul Rapide du Kurtogramme et Applications" **461-464**
Jérôme Antoni

"Apport du modele de Lambert generalisé à l'extraction de relief de surface rugueuse par stereovision photometrique : application aux revetements routiers" **465-468**
Anis Benslimane, Majdi Khoudeir, Jacques Brochard, Minh-tan Do

"Modélisation, évaluation et contrôle d'une chaîne de traitement d'images routières" **469-472**
Antonio Domingues, Yves Lucas, Driss Driouchi, Pierre Marché

"Détection de jingles dans les documents sonores" **473-476**
Julien Pinquier, Régine André-obrecht

"Estimation du signal glottique basée sur un modèle ARX" **477-480**
Damien Vincent, Olivier Rosec, Thierry Chonavel

"Classification d'Expressions Vocales Passives Versus Actives" **481-484**

Zakia Hammal, Baris Bozkurt, Laurent Couvreur, Devrim Unay, Alice Caplier, Thierry Dutoit

"Détection et classification des sons : application aux sons de la vie courante et à la parole" **485-488**

Dan Istrate, Michel Vacher, Jean François Serignat

"Extraction temps-réel de contours labiaux par segmentation vidéo robuste en vue d'animation 3D" **489-492**

Brice Beaumesnil, Franck Luthon, Marc Chaumont

"Influence de la Réduction des Couleurs sur la Détection des Changements de Plan dans les Films d'Animation" **493-496**

Bogdan Ionescu, Patrick Lambert, Didier Coquin, Vasile Buzuloiu

"Filtrage matriciel multicomposante et estimation de la polarisation" **497-500**

Caroline Paulus, Jérôme Mars

Session Orale : *O6.3 "Télédétection et spatial"*

"Analyse d'images SAR polarimétriques au moyen d'outils temps-fréquence" **501-504**

Paul Leducq, Laurent Ferro-famil, Eric Pottier

"Fusion de Méthodes pour la Classification de Zones Urbaines en Télédétection Satellitaire" **505-508**

Mathieu Fauvel, Jocelyn Chanussot, Jon Atli Benediktsson

"Détection de feux de forêt à partir d'images satellitaires IRT par analyse statistique d'évènements rares" **509-512**

Florent Lafarge, Xavier Descombes, Josiane Zerubia, Sandrine Mathieu-marni

"Variational approach to interpolate and correct biases in stereo" **513-516**

Gabriele Facciolo, Andrés Almansa, Alvaro Pardo

"RECONNAISSANCE AUTOMATIQUE DE FREQUENCES DE RESONANCE A PARTIR DES DONNEES SATELLITAIRES DE LA MISSION SPATIALE CLUSTER" **517-520**

Xavier Suraud, Jean-louis Rauch, Roger Lédée, Pierrette Décréau

---

Session Orale : *O4.1 "Segmentation d'images médicales"*

"Une méthode d'initialisation automatique pour contour actif paramétrique, application aux images échographiques" **521-524**

Clovis Tauber, Hadj Batatia, Alain Ayache

"Detection robuste et automatique des contours myocardiques en IRM cardiaque marquée" **525-528**

Aymeric Histace, Christine Cavaro-menard

"Segmentation guidée en imagerie CT 3D : Application à l'arborescence artérielle pulmonaire" **529-532**

Raphael Sebbe, Bernard Gosselin, Roger Lédée, Emmanuel Coche, Benoît Macq

"SEGMENTATION AUTOMATIQUE D'IMAGES A FAIBLE RESOLUTION ET FAIBLE CONTRASTE : APPLICATION A L'IRM CARDIAQUE DU PETIT ANIMAL" **533-536**

Roger Lédée, Rémy Leconge, Christine Cavaro-ménard

"Segmentation du contour de l'endocarde sur des séquences d'images d'échographie cardiaque" **537-540**

Jérôme Lebossé, François Lecellier, Marinette Revenu, Eric Saloux

---

Session Orale : *O1.3 "Multifractales : synthèse et analyse"*

"Synthèse de champs scalaires multifractals : application à la synthèse de texture" **541-544**

Pierre Chainais, Jian-jin Li

"Applications d'outils fractals pour la classification d'images microscopiques de matière grasse laitière" **545-548**

Antoine Echelard, Jacques Lévy-vehel, Ina Taralova

"Contributions à l'étude des performances statistiques des estimateurs multifractals" **549-552**

Stephane Roux, Bruno Lashermes, Patrice Abry, Stephane Jaffard

"Débruitage multifractal par évolution interactive" **553-556**

Evelyne Lutton, Jacques Levy Vehel, Pierre Grenier

Conférence invitée : *C2, M. Unser*

"Vers une théorie unificatrice pour le traitement numérique/analogique des signaux" **557-558**

Michael Unser

Session Affiches : *P1.2 "Non stationnaire et non linéaire"*

"Caractérisation des Fonds Marins par Décomposition Modale Empirique" **559-562**

Abdel-ouahab Boudraa, Claire Pollet, Jean-christophe Cexus, Zineb Saidi

"La Décomposition Modale Empirique à la recherche de l'Onde Circumpolaire Antarctique" **563-566**

Sylvie Roques, Frédérique Rémy

"Utilisation des contours actifs ou « snake » pour l'extraction des arêtes de représentations temps-fréquences : Application à l'analyse des contractions utérines" **567-570**
Jérémy Terrien, Catherine Marque, Guy Germain

"Analyse et représentation de signaux transitoires : application à la compression, au débruitage et à la détection de rupture" **571-574**
Michel Fliess, Cédric Join, Mamadou Mboup, Hebertt Sira-ramirez

"Synthèse temps-fréquence de signaux transitoires dans un contexte de tomographie acoustique océanique discrète" **575-578**
Lionel Cros, Cornel Ioana, André Quinquis, Gwladys Theuillon

"Amélioration de l'estimation spectrale par modélisation AR multidimensionnelle et découpage en sous-bandes" **579-582**
David Bonacci, Corinne Mailhes

"Vers une carte d'identité spectrale" **583-586**
Nadine Martin, Corinne Mailhes, Kheira Sahli, Gérard Lejeune

"Représentations temps-fréquence adaptées aux ondes guidées" **587-590**
Grégoire Le Touzé, Barbara Nicolas, Jérôme Mars, Jean-louis Lacoume, Dominique Fattaccioli

"Choix d'une covariance pour la prédiction par krigeage de séries chronologiques échantillonnées irrégulièrement" **591-594**
Emmanuel Vazquez, Eric Walter

"Amélioration par le bruit de la dynamique de capteurs à saturation" **595-598**
David Rousseau, François Chapeau-blondeau

Session Affiches : *P2.1 "Codage de canal et techniques turbo"*

"Optimisation des codes LDPC pour les communications OFDM" **599-602**
Valérian Mannoni, Guillaume Gelle

"Une méthode simple de turbo estimation de source en décodage conjoint source-canal" **603-606**
Marion Jeanne, Valérian Mannoni, Pierre Siohan

"Algorithme Simplifié pour le Décodage de Codes LDPC non-binaires" **607-610**
Adrian Voicila, David Declercq, Marc Fossorier, François Verdier

"Transmission prioritaire JPEG2000 sur lien sans fil" **611-614**
Benoit Parrein, Nicolas Normand, Patrick Le Callet

"ULTRA-WIDE-BAND RECEIVER BASED ON TURBO-EQUALIZATION" **615-618**
F. Benedetto, G. Giunta, A. Neri, D. Blasi

"Synchronisation de phase - Boucle à remodulation souple" **619-622**
Stéphanie Bay, Christophe Vanstraceele, Benoît Geller, Jean-pierre Barbot, Jean-marc Brossier

"Impact de l'a priori sur les performances de l'égaliseur MAP et distribution des LRV extrinsèques" **623-626**

Noura Sellami, Aline Roumy, Inbar Fijalkow

"Analyse de la convergence d'un système multi-antennes précodé itératif" **627-630**

Jérôme Le Masson, Charlotte Langlais

"Décodage de Turbo Codes Produit Ternaires" **631-634**

Loic Barnault, David Declercq, Jacques Eudes

Table des Communications

Volume 2

Session Affiches : *P3.4 "Reconstruction et débruitage d'images"*

"Reconstruction d'objets composés de plusieurs matériaux en tomographie micro onde" **635-638**
Olivier Féron, Bernard Duchêne, Ali Mohammad-djafari

"Restauration d'images par diffusion sélective" **639-642**
Aymeric Histace, Vincent Courboulay, Michel Menard, Christine Cavaro-menard

"Denoising and Migration Techniques for Target Identification from Ground Penetrating Radar 2D data ; a case study" **643-646**
Olga Lopera, Nada Milisavljevic, Bart Scheers, Benoît Macq

"Opérateurs tomographiques pour la reconstruction d'objets à symétrie de révolution" **647-650**
François Lavallou, Jean Michel Lagrange

"COMPARAISON ENTRE LES ALGORITHMES SEMI-QUADRATIQUES ET GRADIENTS CONJUGUES PRECONDITIONNES POUR LA RESTAURATION D'IMAGE PRESERVANT LES BORDS" **651-654**
Christian Labat, Jérôme Idier

"ECLATEMENT DES CONTRAINTES EN RECONSTRUCTION TOMOGRAPHIQUE" **655-658**
Thomas Capricelli, Patrick Combettes

"Restauration myope d'images pour l'inspection des ouvrages d'art par un système de vision embarqué sur drône" **659-662**
Benoît Vozel, Kacem Chehdi, Jean Dumoulin

"Scintigraphic Image Denoising using Fisz Transformation and Redundant Wavelet Packets" **663-666**
Nawres Kahlifa, Kamel Hamrouni, Noureddine Ellouze

"Restauration des images floues et bruitées multiplicativement" **667-670**
Luc Klaine, Benoit Vozel, Kacem Chehdi

Session Affiches : *P4.1 "Reconnaissance et classification"*

"Une nouvelle transformation pour la localisation des yeux dans une image de visage monochrome" **671-674**
Maurice Milgram, Lionel Prevost, Rachid Belaroussi

"Extraction d'un modèle 3D de visage en temps-réel et de manière robuste" **675-678**
Marc Chaumont

"Perception visuelle du geste de préhension" **679-682**

Pablo Augusto Negri, Xavier Clady, Maurice Milgram

"Détection de piétons par stéréovision et noyaux de graphes." **683-686**

Frédéric Suard, Alain Rakotomamonjy, Abdelaziz Bensrhair

"Apprentissages discriminants en reconnaissance de mots cursifs en-ligne" **687-690**

Emilie Caillault, Christian Viard-gaudin

"Analyse des Orientations pour la Caractérisation d'Images de Documents de la Renaissance" **691-694**

Nicholas Journet, Jean-yves Ramel, Rémy Mullot, Véronique Eglin

"Appariement couleur de planches de bois par inférence floue" **695-698**

Emmanuel Schmitt, Vincent Bombardier, Patrick Charpentier

"Reconnaissance et détection robuste d'objets couleur" **699-702**

Choubeila Maaoui, Hélène Laurent, Bruno Emile

"Quantification des images couleur par un algorithme de classification crédibiliste" **703-706**

Anne-sophie Capelle, Olivier Colot, Christine Fernandez

"Comparaison expérience/calcul par association de courbes" **707-710**

Jean-michel Lagrange, Roberto Ardon, Nicolas Villain

"Une Nouvelle Méthode de Classification en Grande Dimension pour la Reconnaissance de Formes" **711-714**

Charles Bouveyron, Stéphane Girard, Cordelia Schmid

"Localisation de défauts de capteur pour la validation des mesures hydrométriques dans les réseaux d'assainissement" **715-718**

Farida Hamioud, Claude Joannis, José Ragot

"Classification d'EEG pour les interfaces cerveau-machine" **719-722**

Alain Rakotomamonjy, Vincent Guigue, Gregory Mallet, Victor Alvarado

"Nouveau noyau de séquences pour la vérification du locuteur" **723-726**

Jérôme Louradour, Khalid Daoudi

"Espaces hybrides couleur-texture adaptés au comptage d'épis de blé" **727-730**

Frédéric Cointault, Dominique Guérin

Session Orale : *O2.1 "Théorie de l'information et systèmes de communication"*

"Codage entropique à base de règles de ré-écriture" **731-734**

Hervé Jégou, Christine Guillemot

"Allocation MC-CDMA : augmentation des débits sur les lignes de transmission" **735-738**
Jean-yves Baudais, Matthieu Crussière

"Égalisation fréquentielle sous-optimale d'un système mono ou multiporteuse" **739-742**
Carine Simon, Sylvie Mayrargue

"Un nouveau système de transmission basé sur les techniques MIMO SS-MC-MA avec saut de fréquence pour la voie montante des futurs réseaux cellulaires de quatrième génération" **743-746**
Laurent Cariou, Jean-françois Hélard

"PERFORMANCE OF HARQ SCHEMES WITH PACKET AND CODE COMBINING OVER MULTIPATH FADING CHANNELS FOR DS-SSS" **747-750**
M. Wissem El Bahri, Hatem Boujemâa, Mohamed Siala

Session Orale : *O1.4 "Représentation des signaux et des images"*

"Propriétés du second ordre de l'effet de grenaille" **751-754**
Bernard Picinbono

"Sur la Décomposition Modale Empirique des signaux échantillonnés" **755-758**
Gabriel Rilling, Patrick Flandrin

"Vers une nouvelle décomposition de matrice" **759-762**
Jean Jacques Fuchs

"Estimation de signaux par noyaux d'ondelettes" **763-766**
Vincent Guigue, Alain Rakotomamonjy, Stéphane Canu

"Une procédure générale d'estimation efficace de paramètres multiples avec application en localisation" **767-770**
Philippe Forster, Guillaume Ginolhac

Session Orale : *O5.1 "Architectures matérielles et logicielles"*

"Conception fonctionnelle d'une caméra intelligente pour le traitement vidéo temps réel" **771-774**
Barthélémy Heyrman, Michel Paindavoine, Renaud Schmit, Laurent Letellier

"Plate-forme de Prototypage et Conception pour les Systèmes de Radio Logicielle" **775-778**
Fabienne Nouvel, Arnaud Massiani, Pierre De Stéfano

"Synthèse Comportementale Sous Contraintes de Communication et de Placement Mémoire pour les composants du TDSI" **779-782**
Gwenole Corre, Philippe Coussy, Pierre Bomel, Eric Senn, Eric Martin

"Rétroprojection 2D sur plateforme SOPC" **783-786**

Stéphane Mancini, Nicolas Gac, Michel Desvignes

"Optimisation énergétique d'un système de communication dédié à un réseau de capteurs" **787-790**

Mickaël Cartron, Olivier Sentieys

Session Affiches : *P5.1 "Architectures matérielles et logicielles"*

"Un cluster de calcul hybride pour les applications de vision temps-réel" **791-794**

Joel Falcou, Jocelyn Sérot, Thierry Chateau, Frédéric Jurie

"Implantation d'un récepteur Multi-Utilisateurs appliqué à la liaison montante UMTS" **795-798**

Makram Touzri, Sylvie Kerouedan, Samir Saoudi, Patrick Adde

"Un Modèle Hiérarchique pour les Architectures Reconfigurables en Radio Logicielle" **799-802**

Jean Philippe Delahaye, Jacques Palicot, Pierre Leray

"Implantation du décodeur arithmétique de JPEG2000 basée sur la reconfiguration dynamique des FPGAs" **803-806**

Sophie Bouchoux, El-bay Bourennane

"Optimisation sur calculateur d'un modèle d'architecture de couche de cellules rétiniennes" **807-810**

François Devillard, Bernard Heit, Frédéric Grandidier

"Optimisation de la largeur des opérateurs arithmétiques en synthèse de haut-niveau" **811-814**

Nicolas Herve, Daniel Menard, Olivier Sentieys

"Une IP de scrutation et de codage entropique pour la normalisation d'un codeur MPEG4 temps réel" **815-818**

Hachicha Khalil, Garda Patrick

"Reconstruction 3D temps réel dans un VSIP" **819-822**

Tarik Graba, Bertrand Granado, Olivier Romain, Thomas Ea, Andrea Pinna, Patrick Garda

"Séquenceur Mémoire pour Applications Multimédias Temps Réel Gérant les Séquences d'Accès Indéterministes" **823-826**

Bertrand Le Gal, Emmanuel Casseau, Sylvain Huet, Caaliph Andriamisaina, Eric Martin

"Etude Spectrale de l'Echantillonnage par Traversée Niveaux" **827-830**

Fabien Aeschlimann, Emmanuel Allier, Laurent Fesquet, Marc Renaudin

Session Affiches : *P2.2 "Systèmes multi-utilisateurs et multi-capteurs"*

"Séquences de canalisation chaotiques pour les systèmes CDMA multicodes" **831-834**

Mourad Khanfouci, Sylvie Marcos

"Contrôle de puissance et détection multicode conjointe dans une liaison descendante à haut débit de l'UMTS" **835-838**

Bessem Sayadi, Stefan Ataman, Inbar Fijalkow

"Annulation d'interférence dans les systèmes optiques à accès multiple par répartition de code (O-CDMA)" **839-842**

Claire Goursaud, Anne Julien-vergonjanne, Christelle Aupetit-berthelemot, Jean-pierre Cancès, Jean-michel Dumas

"Méthode hybride de suppression d'interférence basée sur l'estimation par paquet des codes inconnus appliquée au HSDPA multicode" **843-848**

Ahmed Saadani, Jean-marie Chaufray

"Construction de filtres LPTV orthogonaux pour un système d'accès multiple par étalement spectral. Etude des effets d'une désynchronisation entre utilisateurs" **849-852**

Wilfried Chauvet, Bernard Lacaze, Daniel Roviras, Alban Duverdier

"Un nouveau décodeur MIMO pour des transmissions MAQ utilisant une approche géométrique : HISD" **853-856**

Amor Nafkha, Christian Roland, Emmanuel Boutillon

"Comparaison du précodeur dmin avec les OSTBC et le précodeur Max-SNR pour les systèmes MIMO" **857-860**

Baptiste Vrigneau, Philippe Rostaing, Koffi Yao

Session Affiches : *P2.3 "Systèmes de communications"*

"Construction d'une Modulation Codée" **861-864**

Sami Hadda, Slaheddine Jarboui

"Synthèse de la notion de facteur de crête et applications aux modulations monoporteuse" **865-868**

Yves Louët, Jacques Palicot

"TEB de système OFDM codé multi-utilisateurs" **869-872**

Adrien Renoult, Inbar Fijalkow, Marc Chenu-tournier

"ALGORITHMES A NORME CONSTANTE POUR LES SYSTEMES DE COMMUNICATION MIMO" **873-876**

Aïssa Ikhlef, Daniel Le Guennec, Jacques Palicot

"Etude théorique des performances d'un détecteur autodidacte d'entrelaceur" **877-880**

Guillaume Sicot, Sébastien Houcke

"Analyse du facteur de crête dans le contexte Radio Logicielle" **881-884**

Sidkièta Zabré, Jacques Palicot

Session Affiches : *P4.2 "Segmentation et indexation"*

"Critères multi-échelle d'évaluation de la segmentation" **885-888**

Sylvie Philipp-foliguet, Laurent Guigues

"Segmentation pyramidale localisée dans un ruban fermé" **889-892**

Pascal Bertolino, Jérémy Huart, Guillaume Foret

"Extraction de formes par approche agent dans des IRM" **893-896**

Yann Jacquelet, Jean-jacques Montois, Mamadou Lamine Ndiaye, Abel Kinié

"Modèles GMM et algorithme de Brandt pour la correction de la segmentation de la parole par HMM" **897-900**

Safaa Jarifi, Dominique Pastor, Olivier Rosec

"Segmentation en Caractères Individuels dans des Images de Scènes Naturelles" **901-904**

Céline Mancas-thillou, Matei Mancas, Bernard Gosselin

"Segmentation d'Images Multispectrales par Arbre de Markov Caché Flou" **905-908**

Pierre Lanchantin, Fabien Salzenstein

"Segmentation non supervisée d'images non stationnaires avec champs de Markov évidentiels" **909-912**

Dalila Benboudjema, Wojciech Pieczynski

"Fusion de segmentations d'images microscopiques par SVM dans différents espaces couleur selon la théorie de l'évidence" **913-916**

Christophe Charrier, Gilles Lebrun, Olivier Lézoray

"Fusion de critères pour l'évaluation de résultats de segmentation d'images" **917-920**

Sébastien Chabrier, Hélène Laurent, Christophe Rosenberger, Alain Rakotomamonjy

"Chaînes de Markov cachées multivariées à bruit corrélé non gaussien, avec applications à la segmentation du signal radar" **921-924**

Nicolas Brunel, Wojciech Pieczynski, Frédéric Barbaresco

"Représentation d'images par chaînes de symboles : Application à la recherche par le contenu" **925-928**

Jean-michel Jolion, Isabelle Simand

"Histogrammes spatiaux couleur optimisés pour l'indexation d'images par le contenu" **929-932**

Saïd Kharbouche, Patrick Vannoorenberghe, Christèle Lecomte, Pierre Miche

Session Orale : *O4.3 "Reconnaissance des formes"*

"Une approche de type filtrage pour la sélection de variables - Application à la reconnaissance automatique de poissons" **933-936**

Dahbia Semani, Carl Frélicot, Pierre Courtellemont

"AMELIORATION DE LA RECONNAISSANCE DE PARTITIONS MUSICALES PAR MODELISATION FLOUE ET INDICATION DES ERREURS POSSIBLES" **937-940**

Florence Rossant, Isabelle Bloch

"Combinaison de classifieurs pour la localisation de visage" **941-944**

Rachid Belaroussi, Lionel Prevost, Maurice Milgram

"Sélection de variables pour la reconnaissance de formes" **945-948**

Sébastien Gadat, Laurent Younes

"Classification de signaux invariante en translation" **949-952**

Vincent Guigue, Alain Rakotomamonjy, Stéphane Canu

Session Orale : *O2.2 "Codage canal et détection"*

"Treillis à complexité réduite pour le décodage de codes à longueur variable" **953-956**

Gholam-reza Mohammad-khani, Chang-ming Lee, Michel Kieffer, Pierre Duhamel

"Une comparaison de récepteurs source-canal conjoint utilisant des codes LDPC" **957-960**

Charly Poulliat, Catherine Lamy-bergot, Inbar Fijalkow

"Détection de phase à très faible rapport signal à bruit à l'aide d'un code de parité" **961-964**

Mathieu Dervin, Marie-laure Boucheret, Gilles Mesnager, Alain Ducasse

"Séparation aveugle de signaux de type DS-CDMA à l'aide de techniques algébriques" **965-968**

Castaing Joséphine, De Lathauwer Lieven

Session Orale : *O1.1 "Nouvelles tendances en temps-fréquence"*

"Reconnaissance des formes par méthodes à noyau dans le domaine temps-fréquence" **969-972**

Paul Honeiné, Cédric Richard, Patrick Flandrin

"Une Méthode à Noyaux pour l'Approximation Parcimonieuse des Représentations Temps-Fréquence Bilinéaires" **973-976**

Manuel Davy, Patrick Wolfe

"Estimation et reconstruction des signaux courts multicomposantes modules non lineairement à la fois en amplitude et en frequence" **977-980**

Meryem Jabloun, Michelle Vieira, Nadine Martin, François Léonard

"Algorithme de zéro-diagonalisation conjointe pour la séparation de sources déterministes" **981-984**
Elmostafa Fadaili, Nadège Thirion-moreau, Eric Moreau

"Suivi de trajectoires temps-fréquence par filtrage particulaire" **985-988**
Corentin Dubois, Manuel Davy

Conférence invitée : *C3, H. Vinet*

"Applications musicales du traitement de signal : synthèse et prospective" **989-996**
Hugues Vinet

Session Affiches : *P2.4 "Codage, compression, cryptage et tatouage"*

"Estimation de mouvement par maillages et lignes de rupture pour le codage de séquences vidéo" **997-1000**
Nathalie Cammas, Stéphane Pateux, Luce Morin, Benjamin Le Guen

"Codage vidéo par schéma lifting avec gestion des occlusions" **1001-1004**
Thomas André, Marc Antonini, Michel Barlaud

"Réduction du débit et de la complexité de décodage par l'analyse de scène pour la télé surveillance basée sur Motion JPEG 2000" **1005-1008**
Jérôme Meessen, Christophe Parisot, Xavier Desurmont, Jean-françois Delaigle

"Codage par descriptions multiples pour la transmission vidéo" **1009-1012**
Christophe Tillier, Teodora Petrisor, Beatrice Pesquet-popescu, Jean-christophe Pesquet

"COMPRESSION DE MAILLAGES 3D DE GRANDE RESOLUTION PAR TRANSFORMEE EN ONDELETTES AU FIL DE L'EAU" **1013-1016**
Akran Elkefi, Sami Abbadi, Marc Antonini, Chokri Ben Amar

"Diffusion géométrique pour le masquage d'erreurs de transmission et de quantification sur des images JPEG couleur" **1017-1020**
Pascal Bourdon, Bertrand Augereau, Christian Chatellier, Christian Olivier

"Interleaved S+P : Décomposition pyramidale entrelacée à contexte de prédiction enrichi" **1021-1024**
Marie Babel, Olivier Déforges, Joseph Ronsin

"Evaluation d'un codeur de parole à très bas débit" **1025-1028**
Marc Padellini, Geneviève Baudoin, François Capman

"Métriques de distorsion pour l'analyse comparative de schémas de filigranage 3D" **1029-1032**
Patrice Rondao Alface, Mathieu De Craene, Benoit Macq

"Un Algorithme de Tatouage d'Images Numériques reposant sur les Changements d'Horloge Périodiques" **1033-1036**
Vincent Martin, Marie Chabert, Bernard Lacaze

"Stéganographie d'images couleur dans l'espace L*a*b* par modulation de distance" **1037-1040**
Gael Chareyron, Dinu Coltuc, Alain Trémeau

"Insertion d'informations numériques dans un signal : Application à la classification de données expérimentales en anémométrie laser Doppler" **1041-1044**
Alain Le Duff, Bruno Gazengel

"Transfert sécurisée d'une ROI sans perte par une méthode d'insertion de données cachées robuste à la compression JPEG" **1045-1048**
Philippe Amat, William Puech

"Performances d'un cryptosystème numérique à base de chaos" **1049-1052**
Véronique Guglielmi, Michel Bonnefont, Danièle Fournier-prunaret, Pierre Pinel, Abdel Kaddous Taha

---

Session Affiches : *P3.2 "Séparation de sources et problèmes inverses"*

"Séparation de sources non-négatives par l'approche du maximum de vraisemblance" **1053-1056**
Saïd Moussaoui, David Brie, Cédric Carteret

"Séparation de mélanges convolutifs de modulations CPM suréchantillonnées" **1057-1060**
Pierre Jallon, Antoine Chevreuil, Philippe Loubaton, Pascal Chevalier

"Approximation en champs moyen pour la séparation de sources appliquée aux images hyperspectrales" **1061-1064**
Nadia Bali, Ali Mohammad-djafari

"Solution anlytique et généralisation d'un contraste à référence" **1065-1068**
Atman Jbari, Abdellah Adib, Driss Aboutajeddine

"Déflation parallèle avec des contrastes APF pour l'extraction aveugle de sources" **1069-1072**
Vicente Zarzoso, Ludwig Rota, Pierre Comon

"Méthodes séquentielles de séparation aveugle de sources en présence de bruit" **1073-1076**
Marc Castella, Jean-christophe Pesquet

"Simuler une distribution normale à support positif à partir de plusieurs lois candidates" **1077-1080**
Vincent Mazet, David Brie, Jérôme Idier

"Déconvolution impulsionnelle positive myope" **1081-1084**
Vincent Mazet, Jérôme Idier, David Brie

"Identification aveugle de modèle acoustique et application à la déréverbération de la parole" **1085-1088**
William Bobillet, Eric Grivel, Roberto Guidorzi, Mohamed Najim

"Le lissage H-infini : une alternative au filtrage de Kalman pour le rehaussement du signal de parole ?" **1089-1092**
David Labarre, Eric Grivel, Nicolai Christov, Mohamed Najim

"Recherche Linéaire Optimisée : Une Nouvelle Méthode pour Accélérer PARAFAC" **1093-1096**
Myriam Rajih, Pierre Comon

"Une méthode de sous-espace pour l'extraction aveugle d'une source cyclostationnaire" **1097-1100**
Roger Boustany, Jérôme Antoni

"Déconvolution aveugle dans le domaine fréquentiel fondée sur le taux d'information mutuelle" **1101-1104**
Anthony Larue, Jérome Mars, Christian Jutten

Session Affiches : *P1.3 "Modèles déformables, géométriques et cognitifs"*

"Définition d'une force d'attraction aux LPE pour la segmentation par contours actifs" **1105-1108**
Annabelle Gouze, Cédric De Roover, Benoît Macq

"Détection de signalisation routière par ajustement de formes prototypes" **1109-1112**
Guillaume Dutilleux, Pierre Charbonnier

"Utilisation des Gradients de Forme et des Contours Actifs Basés Régions pour la Segmentation des Vecteurs Mouvement" **1113-1116**
Stéphanie Jehan-besson, Stefan Duffner, Ariane Herbulot, Michel Barlaud, Gilles Aubert

"Extraction de la forme et de la perspective dans des textures artificielles et naturelles par modèles corticaux." **1117-1120**
Corentin Massot, Jeanny Hérault

"Recalage semi-rigide d'images de scanner aéroporté" **1121-1124**
Claude Cariou, Kacem Chehdi

"Contours actifs pour le suivi d'objet en temps-réel : multi-topologies et multi-résolutions" **1125-1128**
Sébastien Lefèvre, Nicole Vincent

Session Orale : *O4.2 "Méthodes à noyaux"*

"Influence des Vecteurs Caractéristiques en Stéganalyse par Séparateurs à Vastes Marges" **1129-1132**
Benoit Roue, Patrick Bas, Jean Marc Chassery

"Kernel machines : une nouvelle méthode pour l'optimisation de l'alignement des noyaux et l'amélioration des performances" **1133-1136**
Jean-baptiste Pothin, Cédric Richard

"Reconstruction par noyaux polynomiaux" **1137-1140**

Maxime Berar, Michel Desvignes, Barbara Romaniuk, Gérard Bailly, Yohan Payan

"Classification de végétaux" **1141-1144**

Michel Chapron, Louis Assemat

"Classification avec contraintes : problématique et apprentissage d'une règle de décision par SVM" **1145-1148**

Edith Grall-maës, Pierre Beauseroy, Abdenour Bounsiar

Session Orale : *O3.2 "Images : séquences et 3D"*

"Guidage du filtrage particulaire par une étude du mouvement" **1149-1152**

Elodie Bichot, Laurent Mascarilla, Pierre Courtellemont

"Recalage variationnel non rigide statistique référencérégion" **1153-1156**

Nicolas Rougon, Antoine Discher, Françoise Prêteux

"Décomposition de déformation pour l'estimation d'un mouvement de caméra" **1157-1160**

Jonchery Claire, Dibos Françoise, Koepfler Georges

"Agrégation, sélection et utilisation de l'information de mouvement issue d'un flux MPEG" **1161-1164**

Renan Coudray, Bernard Besserer

"Reconnaissance de sauts d'athlète dans les vidéos : fusion du mouvement de caméra et de la trajectoire de points caractéristiques" **1165-1168**

Emmanuel Ramasso, Denis Pellerin, Michèle Rombaut

Session Orale : *O6.2 "Radar"*

"Détection MV-TFAC dynamique dans un fouillis de mer hétérogène" **1169-1172**

Nicolas Bon, Ali Khenchaf, René Garello, Jean-michel Quellec, Myriam Chabah

"Imagerie d'objets mobiles à l'aide d'un radar bande étroite multistatique" **1173-1176**

Guillaume Ginolhac, Franck Daout, Françoise Schmitt

"Un algorithme de Gradient Conjugué modifié pour le traitement Spatio-Temporel adaptatif en radar aéroporté" **1177-1180**

Hocine Belkacemi, Sylvie Marcos, Hichem Semira

"Antibrouillage Radar en contexte de rotation d'antenne" **1181-1184**

Marc Oudin, Jean-pierre Delmas, Claude Adnet, Cécile Germond, Frédéric Barbaresco

"Analyse de la Texture dans les Images SAR Appliquée au Filtrage de Speckle" **1185-1188**

Olivier D'hondt, Laurent Ferro-famil, Eric Pottier

Session Orale : *O2.3 "Tatouage et stéganographie"*

"Quantification vectorielle algébrique modulée : codage de source/tatouage conjoints à débit variable" **1189-1192**

Ludovic Guillemot, Jean-marie Moureaux

"Amélioration de l'invisibilité par adaptation de la quantification aux données à insérer." **1193-1196**

Jean-luc Toutant, William Puech, Christophe Fiorio

"Steganographie d'images basées sur le theorème chinois des restes" **1197-1200**

Dinu Coltuc, Alain Tremeau

"Tatouage vidéo utilisant l'ACI et basé sur un treillis informé" **1201-1204**

Hussein Joumaa, Franck Davoine

"Mécanisme de synchronisation dans des systèmes de tatouage audio pour des perturbations désynchronisantes à forte dérive" **1205-1208**

Cléo Baras, Nicolas Moreau, Bassem Zayen

Session Orale : *O1.2 "Modèles déformables, géométriques et cognitifs"*

"ETUDE ET EXTENSION DES FLOTS DE RICCI, KÄHLER-RICCI ET CALABI DANS LE CADRE DU TRAITEMENT DE L'IMAGE ET DE LA GEOMETRIE DE L'INFORMATION" **1209-1212**

Frédéric Barbaresco

"Contraintes de forme invariante aux transformations affines pour les contours actifs orientés région" **1213-1216**

Alban Foulonneau, Pierre Charbonnier, Fabrice Heitz

"Formalisme et mesure de la perception : intérêt de la dynamique sensorimotrice chez un agent robotique" **1217-1220**

Mickaël Maillard, Philippe Gaussier, Laurence Hafemeister

"MODELE STATISQUE ET DESCRIPTION LOCALE D'APPARENCE NON LINEAIRE POUR LA DETECTION DES COUTOURS DES LEVRES" **1221-1224**

Pierre Gacon, Pierre-yves Coulon, Gérard Bailly

"MGS2 : Optimisation multicritères de contours actifs par algorithmes génétiques" **1225-1228**

Nicolas Cladel, Renaud Seguier

Session Orale : *O3.3 "Séparation de sources dans des mélanges sous-déterminés"*

"Séparation voix / musique à partir d'enregistrements mono : quelques remarques sur le choix et l'adaptation des modèles" **1229-1232**

Alexey Ozerov, Rémi Gribonval, Pierrick Philippe, Frédéric Bimbot

"Séparation aveugle sous-déterminée de sources audio par la méthode EMD (Empirical Mode Decomposition)" **1233-1236**

Abdeldjalil Aïssa-el-bey, Karim Abed-meraim, Yves Grenier

"Séparation de sources dans le cas sous-déterminé : comparaison de deux approches basées sur des décompositions parcimonieuses" **1237-1240**

Sylvain Lesage, Sacha Krstulovic, Rémi Gribonval

"Séparation d'un mélange sous-déterminé de sources par simulation stochastique" **1241-1244**

Éric Le Carpentier

"Construction d'estimateurs oracles pour la séparation de sources" **1245-1248**

Emmanuel Vincent, Rémi Gribonval

Reconstruction d'objets composés de plusieurs matériaux en tomographie micro-onde

Olivier FÉRON[1], Bernard DUCHÊNE[1], Ali MOHAMMAD-DJAFARI[1]

[1]Laboratoire des Signaux et Systèmes (L2S)
Supélec, Plateau de Moulon
3 rue Joliot Curie, 91192 Gif-sur-Yvette, France
feron,duchene,djafari@lss.supelec.fr

Résumé – Cet article propose une m´ethode de reconstruction d'objets compos´es de plusieurs mat´eriaux en tomographie micro-onde. Ce problème est connu pour être mal pos´e. L'id´ee consiste alors à utiliser l'approche bay´esienne pour introduire une information *a priori* : les diff´erents mat´eriaux constituant l'objet sont connus. La m´ethode propos´ee est bas´ee sur des techniques de MCMC, où une ´etape d'´echantillonage est remplac´ee par la minimisation d'un critère r´egularis´e. L'algorithme est test´e sur des donn´ees r´eelles issues de l'institut Fresnel de Marseille.

Abstract – We propose an inversion algorithm dedicated to the reconstruction of objects made of a few number of different materials in microwave imaging. Due to the ill-posedness of this problem the Bayesian approach is used to introduce *a priori* information : the knowledge of the different materials that compose the object. The proposed method is based upon MCMC techniques where a step of sampling is replaced by the minimization of a regularized criterion. This algorithm has been applied to real data collected at the Institut Fresnel (Marseille, France).

1 Introduction

Nous étudions ici l'interaction d'ondes électromagnétiques avec un objet inhomogène (constitué de plusieurs matériaux) plongé dans un milieu homogène de permittivité diélectrique ε_0, de conductivité σ_0 et de perméabilité magnétique μ_0.

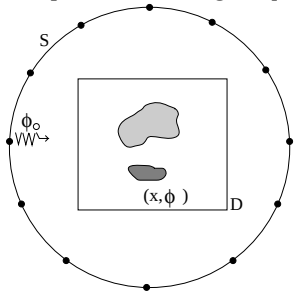


FIG. 1: *Configuration des mesures*

Un champ incident $\boldsymbol{\phi}_{0_{q,v}}$ est créé à plusieurs fréquences $q \in \{1,\ldots,Q\}$ et plusieurs angles d'incidence $v \in \{1,\ldots,N_v\}$. L'objet inhomogène $x(\boldsymbol{r})$ est supposé être contenu dans un domaine test D et est caractérisé par sa permittivité relative $\varepsilon_r(\boldsymbol{r})$ et sa conductivité $\sigma(\boldsymbol{r})$ et nous le chercherons sous la forme suivante :

$$x(\boldsymbol{r}) = \varepsilon_r(\boldsymbol{r}) - 1 + i\,\sigma(\boldsymbol{r}) = x_r(\boldsymbol{r}) + i\,x_i(\boldsymbol{r}) \qquad (1)$$

Cet objet crée, pour chaque fréquence et chaque angle d'incidence, un champ diffracté $\boldsymbol{y}_{q,v}(\boldsymbol{s})$ acquis aux points de mesures $\boldsymbol{s} \in \mathcal{S}$. En se plaçant dans une configuration 2D "transverse magnétique" (TM), le problème direct peut s'écrire sous forme vectorielle par l'intermédiaire d'une équation d'observation et d'une équation d'état, soit respectivement :

$$\begin{aligned} \boldsymbol{y}_{q,v} &= \boldsymbol{G}^S_{q,v}\boldsymbol{X}_q\boldsymbol{\phi}_{q,v}, \quad q=1,\ldots,Q, \quad v=1,\ldots,N_v \\ \boldsymbol{\phi}_{q,v} &= \boldsymbol{\phi}_{0_{q,v}} + \boldsymbol{G}^D_q\boldsymbol{X}_q\boldsymbol{\phi}_{q,v}, \qquad \boldsymbol{X}_q = diag(\boldsymbol{x}_q) \end{aligned} \qquad (2)$$

où $\boldsymbol{G}^S_{q,v}$ et $\boldsymbol{G}^D_q$ sont les matrices associées aux noyaux de Green, $\boldsymbol{\phi}_{q,v} = \{\boldsymbol{\phi}_{q,v}(\boldsymbol{r}), \boldsymbol{r} \in D\}$ est le champ total à l'intérieur D de l'objet, et $\boldsymbol{x}_q$ est défini par :

$$\boldsymbol{x}_q(\boldsymbol{r}) = \boldsymbol{x}_r(\boldsymbol{r}) + i\,\frac{\boldsymbol{x}_i(\boldsymbol{r})}{\Omega_q}, \qquad (3)$$

où Ω_q est une constante qui dépend de la fréquence q. Le problème inverse consiste à reconstruire les valeurs du contraste $\boldsymbol{x}$ à partir des données mesurées $\boldsymbol{y}_{q,v}$ du champ diffracté.
Plusieurs difficultés se présentent alors : (i) l'inversion de l'équation d'observation est connue pour être un problème extrèmement mal posé, (ii) le problème inverse est non linéaire, du fait que le champ total $\boldsymbol{\phi}$ à l'intérieur de l'objet est inconnu, et (iii) l'objet $\boldsymbol{x}$ recherché n'intervient dans les équations que par l'intermédiaire de $\boldsymbol{x}_q$ qui dépend de la fréquence.
Les méthodes classiques d'optimisation ont pour objectif d'estimer conjointement les variables $\boldsymbol{x}$ et $\boldsymbol{\phi}$ ([1]), ou $\boldsymbol{x}$ et $(\boldsymbol{w} = \boldsymbol{X}\boldsymbol{\phi})$ ([2]). C'est ce deuxième cas que nous nous proposons d'étudier dans ce papier en multipliant l'équation d'état par $\boldsymbol{X}_q$, de manière à obtenir :

$$\begin{aligned} \boldsymbol{y}_{q,v} &= \boldsymbol{G}^S_{q,v}\boldsymbol{w}_{q,v}, \\ \boldsymbol{w}_{q,v} &= \boldsymbol{X}_q\boldsymbol{\phi}_{0_{q,v}} + \boldsymbol{X}_q\boldsymbol{G}^D_q\boldsymbol{w}_{q,v}, \end{aligned} \qquad (4)$$

avec le changement de variable $\boldsymbol{w}_{q,v} = \boldsymbol{X}_q\boldsymbol{\phi}_{q,v}$. Une méthode classique de résolution de ces équations est la méthode d'inversion contraste source (CSI) qui consiste à minimiser le critère suivant :

$$\begin{aligned} J_{csi}(\boldsymbol{x},\boldsymbol{w}) = & \frac{\sum_{q,v}||\boldsymbol{y}_{q,v} - \boldsymbol{G}^S_{q,v}\boldsymbol{w}_{q,v}||^2_S}{\sum_{q,v}||\boldsymbol{y}_{q,v}||^2_S} \\ & + \frac{\sum_{q,v}||\boldsymbol{X}\boldsymbol{\phi}_{0q} - \boldsymbol{w}_{q,v} + \boldsymbol{X}\boldsymbol{G}^D_q\boldsymbol{w}_{q,v}||^2_D}{\sum_{q,v}||\boldsymbol{X}\boldsymbol{\phi}_{0q}||^2_D} \end{aligned} \qquad (5)$$

avec $\boldsymbol{w} = \{\boldsymbol{w}_{q,v},\ q=1,\ldots,Q,\ v=1,\ldots,N_v\}$, succesivement par rapport à $\boldsymbol{x}$ et $\boldsymbol{w}$.

Notons dans ce cas l'apparition de $\boldsymbol{X}_q$ au dénominateur du second terme, ce qui permet de faire varier le poids des deux termes. Il s'avère de plus qu'en pratique l'initialisation de $\boldsymbol{x}$ à des valeurs proches de zéro permet d'obtenir de meilleurs résultats et fait évoluer le terme $\sum_{q,v} ||\boldsymbol{X}\boldsymbol{\phi}_{0_{q,v}}||^2_D$ de manière toujours croissante. Dans cette approche aucune information *a priori* n'est utilisée. Le résultat obtenu peut être très sensible à l'initialisation de l'algorithme et au bruit sur les données.

2 Approche proposée

Nous supposons que l'objet est constitué de K matériaux différents (air, diélectrique,métal) connus à l'avance. L'objectif est donc d'utiliser les méthodes classiques d'optimisation en introduisant cette information *a priori*. Dans [1], une solution déterministe est proposée pour le cas d'objets binaires uniquement. Nous proposons ici une approche bayésienne comme une alternative pour l'introduction d'informations *a priori*.

2.1 Interprétation probabiliste de J_{csi}

Dans un premier temps nous identifions les hypothèses nécessaires pour interpréter la méthode d'inversion contraste source comme la maximisation d'une loi *a posteriori*. Ce type d'interprétation probabiliste a déjà été développé dans [3, 4] pour retrouver le critère du gradient modifié, en considérant un bruit gaussien centré iid sur chacune des deux équations couplées. Cependant la considération de ces deux bruits gaussiens entraine implicitement une corrélation entre ces erreurs et l'objet $\boldsymbol{x}$. Dans le cas de la méthode d'inversion contraste source, cette corrélation est directement visible. En effet, le critère J_{csi} se retrouve en ajoutant encore deux bruits gaussiens sur les deux équations :

$$\begin{aligned} \boldsymbol{y}_{q,v} &= \boldsymbol{G}^S_{q,v}\boldsymbol{w}_{q,v} + \boldsymbol{\varepsilon}_{q,v} \\ \boldsymbol{w}_{q,v} &= \boldsymbol{X}_q\boldsymbol{\phi}_{0_{q,v}} + \boldsymbol{X}_q\boldsymbol{G}^D_q\boldsymbol{w}_{q,v} + \boldsymbol{\eta}_{q,v} \end{aligned} \tag{6}$$

Supposons de plus que $\boldsymbol{\varepsilon}_{q,v} \sim \mathcal{N}(0, r^2 \sum ||\boldsymbol{y}_{q,v}||^2_S \, \boldsymbol{I}_S)$ et $\boldsymbol{\eta}_{q,v} \sim \mathcal{N}(0, r^2 \sum ||\boldsymbol{X}_q\boldsymbol{\phi}_{0_{q,v}}||^2_D \, \boldsymbol{I}_D)$, avec $\boldsymbol{I}_S$ et $\boldsymbol{I}_D$ les matrices identité dans les domaines $\mathcal{S}$ et D, alors nous obtenons

$$\begin{aligned} p(\boldsymbol{y}_{q,v}|\boldsymbol{w}_{q,v}) &\propto \exp\left\{-\frac{||\boldsymbol{y}_{q,v} - \boldsymbol{G}^S_{q,v}\boldsymbol{w}_{q,v}||^2_S}{2r^2\sum_{q,v}||\boldsymbol{y}_{q,v}||^2_S}\right\} \\ p(\boldsymbol{w}_{q,v}|\boldsymbol{x}) &\propto \exp\left\{-\frac{||\boldsymbol{w}_{q,v} - \boldsymbol{X}_q\boldsymbol{\phi}_{0_{q,v}} - \boldsymbol{X}_q\boldsymbol{G}^D_q\boldsymbol{w}_{q,v}||^2_D}{2r^2\sum_{q,v}||\boldsymbol{X}_q\boldsymbol{\phi}_{0_{q,v}}||^2_D}\right\} \end{aligned} \tag{7}$$

et l'estimation jointe de $\boldsymbol{x}$ et $\{\boldsymbol{w}_{q,v}\}_{q,v}$ au sens du maximum *a posteriori* (MAP) avec une loi *a priori* uniforme sur $\boldsymbol{x}$ nous ramène à minimiser le critère J_{csi}.

2.2 Choix d'un *a priori* approprié

Pour traiter ce problème inverse de diffraction, des méthodes de régularisation ont été développées ([1, 2, 3]) dans le but d'introduire l'information *a priori* que l'objet est constitué de zones homogènes (terme de régularisation de type markovien) ou de contours nets (terme de préservation de contours). Dans ce papier nous proposons une solution pour apporter une information supplémentaire : le fait que nous connaissons les différents matériaux qui peuvent constituer l'objet. Nous considérons pour cela un modèle de Markov caché (HMM) décrit dans [5] pour des variables réelles et étendu ici dans le cas de variables complexes. Nous introduisons donc une variable cachée discrète $z(\boldsymbol{r}) \in \{1, \ldots, K\}$ représentant une classification de l'objet $\boldsymbol{x}$ à reconstruire. Chaque classe k est caractérisée par une moyenne complexe $m_k = m_{kr} + i m_{ki}$ et deux variances σ^2_{kr} et σ^2_{ki} respectivement sur les parties réelles et complexes. De cette manière nous obtenons la loi *a priori* conditionnelle :

$$p(x(\boldsymbol{r})|z(\boldsymbol{r})) = \mathcal{N}\left(\begin{pmatrix} m_{kr} \\ m_{ki} \end{pmatrix}, \begin{pmatrix} \sigma^2_{kr} & 0 \\ 0 & \sigma^2_{ki} \end{pmatrix}\right), \tag{8}$$

où dans cette écriture $x(\boldsymbol{r})$ est considéré comme vecteur à deux dimensions $x(\boldsymbol{r}) = [x_r(\boldsymbol{r}), x_i(\boldsymbol{r})]^T$ (où T indique la transposition) en séparant sa partie réelle et sa partie imaginaire. Remarquons de plus que l'on peut utiliser le même raisonement pour écrire la loi *a priori* conditionnelle du vecteur $\boldsymbol{x} \in \mathbb{C}^{N_D}$ (où N_D est le nombre de pixels de D) considéré comme vecteur réel $\boldsymbol{x} = [\boldsymbol{x}_r, \boldsymbol{x}_i]^T \in \mathbb{R}^{2N_D}$ et on a alors :

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{z}) &= \mathcal{N}\left(\begin{pmatrix} \boldsymbol{m}_{z_r} \\ \boldsymbol{m}_{z_i} \end{pmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{z_r} & 0 \\ 0 & \boldsymbol{\Sigma}_{z_i} \end{bmatrix}\right) \\ &\propto \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m}_z)^T\boldsymbol{\Sigma}^{-1}_z(\boldsymbol{x} - \boldsymbol{m}_z)\right\} \end{aligned} \tag{9}$$

en définissant

$$\left.\begin{array}{ll} \boldsymbol{m}_{z_r}(\boldsymbol{r}) & = m_{kr} \\ \boldsymbol{m}_{z_i}(\boldsymbol{r}) & = m_{ki} \\ \boldsymbol{\Sigma}_{z_r}(\boldsymbol{r}) & = \sigma^2_{kr} \\ \boldsymbol{\Sigma}_{z_r}(\boldsymbol{r}) & = \sigma^2_{kr} \end{array}\right\} \text{ si } z(\boldsymbol{r}) = k$$

Le fait que nous connaissons les K différents matériaux qui peuvent constituer l'objet se traduit simplement par l'affectation de valeurs spécifiques aux moyennes m_{kr} et m_{ki}, et par des variances associées σ^2_{kr} et σ^2_{ki} plus ou moins faibles suivant la précision que l'on a sur les permittivités et conductivités des matériaux. Par exemple, l'étiquette 1 représentera le vide, et on prendra alors $m_{kr} = 0$, $m_{ki} = 0$, $\sigma^2_{kr} = 10^{-3}$ et $\sigma^2_{ki} = 10^{-3}$. Pour renforcer l'idée que l'objet est constitué de zones homogènes, nous considérons une corrélation *a priori* entre les pixels de l'image, par l'intermédiaire d'un champ de Potts à 4 voisins sur $\boldsymbol{z}$:

$$p(\boldsymbol{z}) = \frac{1}{Q(\alpha)}\exp\left\{\alpha\sum_{\boldsymbol{r}\in D}\sum_{\boldsymbol{s}\in\mathcal{V}(\boldsymbol{r})}\delta[z(\boldsymbol{r}) - z(\boldsymbol{s})]\right\}, \tag{10}$$

avec $\delta(0) = 1$ et $\delta(t) = 0$ si $t \neq 0$ et où $\mathcal{V}(\boldsymbol{r})$ est le voisinage de $\boldsymbol{r}$ (de 4 pixels), $Q(\alpha)$ est la fonction de partition et α représente le degré de dependance spatiale de $\boldsymbol{z}$.

2.3 Algorithme proposé

Si nous notons $\boldsymbol{y} = \{\boldsymbol{y}_{q,v};\ q = 1, \ldots, Q;\ v = 1, \ldots, Nv\}$ et $\boldsymbol{w} = \{\boldsymbol{w}_{q,v};\ q = 1, \ldots, Q;\ v = 1, \ldots, Nv\}$, l'objectif est donc d'estimer $(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z})$ à partir de la loi *a posteriori* jointe $p(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}|\boldsymbol{y})$, qui peut s'écrire de la manière suivante :

$$p(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{w})\, p(\boldsymbol{w}|\boldsymbol{x})\, p(\boldsymbol{x}|\boldsymbol{z})\, p(\boldsymbol{z}) \tag{11}$$

et

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{w}) &= \prod_{q,v} p(\boldsymbol{y}_{q,v}|\boldsymbol{w}_{q,v}) \\ p(\boldsymbol{w}|\boldsymbol{x}) &= \prod_{q,v} p(\boldsymbol{w}_{q,v}|\boldsymbol{x}) \end{aligned} \tag{12}$$

et les termes de ces deux équations sont définis par (7). En revanche la maximisation de cette loi *a posteriori* est difficile, notamment dû au fait que nous avons à la fois des variables discrètes et continues. Nous proposons alors d'utiliser des techniques de MCMC (Monte Carlo Markov chain) pour générer des échantillons successifs de $(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{z})$ suivant cette loi *a posteriori* et à partir desquels on peut calculer la moyenne *a posteriori*. L'algorithme proposé se rapproche des méthodes d'échantillonage de Gibbs :

- échantilloner $(\boldsymbol{x}, \boldsymbol{w}) \sim p(\boldsymbol{x}, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{z})$
- échantilloner $\boldsymbol{z} \sim p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w})$

Première étape : échantillonage de $(\boldsymbol{x}, \boldsymbol{w})$
En utilisant la règle de Bayes, nous obtenons la relation :

$$p(\boldsymbol{x}, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{z}) \propto p(\boldsymbol{y}|\boldsymbol{w})\, p(\boldsymbol{w}|\boldsymbol{x})\, p(\boldsymbol{x}|\boldsymbol{z}) \tag{13}$$

Du fait de la non linéarité du problème, cette loi *a posteriori* n'est pas une loi classique et l'échantillonage exact de $(\boldsymbol{x}, \boldsymbol{w})$ n'est pas possible. Une solution serait d'implémenter des méthodes de type Hastings-Metropolis ([6]) pour générer un échantillon suivant cette loi *a posteriori*. Dans ce papier, cette étape de génération d'un échantillon de $p(\boldsymbol{x}, \boldsymbol{w}|\boldsymbol{y}, \boldsymbol{z})$ est remplacée par l'échantillon à maximum *a posteriori* ([7]), obtenu en minimisant le critère régularisé :

$$J(\boldsymbol{x}, \boldsymbol{w}) = J_{csi}(\boldsymbol{x}, \boldsymbol{w}) + \lambda J_{reg}(\boldsymbol{x}) \tag{14}$$

où $\quad J_{reg}(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{m}_z)^T \boldsymbol{\Sigma}_z^{-1} (\boldsymbol{x} - \boldsymbol{m}_z)$

Cette minimisation s'inspire fortement du travail de van den Berg ([2]) sur la méthode d'inversion contraste source : l'étape de descente sur les $\boldsymbol{w}_{q,v}$ est en effet totalement identique. En revanche, comme dans ce papier nous considérons des objets mixtes (c'est à dire composés de métal et de diélectrique) et non binaires, nous proposons ici de décomposer la variable $\boldsymbol{x}$ en sa partie réelle et sa partie imaginaire. Dans ce cas, l'étape de descente sur $\boldsymbol{x}$, à l'itération (n), nécesite le calcul de deux gradients $\boldsymbol{g}_r^{(n)}$ et $\boldsymbol{g}_i^{(n)}$ respectivement pour les parties réelle et imaginaire de $\boldsymbol{x}$. Ces gradients sont approchés en considérant les dénominateurs constants et dépendant de $\boldsymbol{x}_q^{(n-1)}$. On obtient alors les relations suivantes :

$$\boldsymbol{g}_r^{(n)} = \frac{\Re e\left\{\overline{\boldsymbol{\phi}_{q,v}^{(n)}}(\boldsymbol{X}_q^{(n)}\boldsymbol{\phi}_{q,v}^{(n)} - \boldsymbol{w}_{q,v}^{(n)})\right\}}{\sum_{q,v} ||\boldsymbol{X}_q^{(n-1)}\boldsymbol{\phi}_{0_{q,v}}||_D^2} + \boldsymbol{\Sigma}_{z_r}^{-1}(\boldsymbol{x}_r - \boldsymbol{m}_{z_r})$$

$$\boldsymbol{g}_i^{(n)} = \frac{\frac{1}{\Omega_q}\Im m\left\{\overline{\boldsymbol{\phi}_{q,v}^{(n)}}(\boldsymbol{X}_q^{(n)}\boldsymbol{\phi}_{q,v}^{(n)} - \boldsymbol{w}_{q,v}^{(n)})\right\}}{\sum_{q,v} ||\boldsymbol{X}_q^{(n-1)}\boldsymbol{\phi}_{0_{q,v}}||_D^2} + \boldsymbol{\Sigma}_{z_i}^{-1}(\boldsymbol{x}_i - \boldsymbol{m}_{z_i})$$

A partir de ces gradients on calcule les directions conjuguées $\boldsymbol{d}_r^{(n)}$ et $\boldsymbol{d}_i^{(n)}$ de manière classique en utilisant la formule de Polak-Ribière. Les pas optimaux de descente $\alpha_r^{(n)}$ et $\alpha_i^{(n)}$ sont calculés conjointement en minimisant $J(\boldsymbol{x}+\alpha_r \boldsymbol{d}_r^{(n)} + i\,\alpha_i \boldsymbol{d}_i^{(n)})$. On peut en effet réécrire le critère J_{csi} de la manière suivante :

$$J_{csi} = \frac{a_r\alpha_r^2 + a_i\alpha_i^2 + 2d_{ri}\alpha_r\alpha_i + 2b_r\alpha_r + 2b_i\alpha_i + c}{A_r\alpha_r^2 + A_i\alpha_i^2 + 2D_{ri}\alpha_r\alpha_i + 2B_r\alpha_r + 2B_i\alpha_i + C},$$

avec, en notant

$$\boldsymbol{\rho}_{q,v} = \boldsymbol{w}_{q,v} - \boldsymbol{X}_q\boldsymbol{\phi}_{0_{q,v}} - \boldsymbol{X}_q\boldsymbol{G}_q^D\boldsymbol{w}_{q,v}$$
$$\boldsymbol{D}_{q,v} = \boldsymbol{X}_q\boldsymbol{\phi}_{0_{q,v}},$$

$$a_r = \sum_{q,v} ||\boldsymbol{\phi}_{q,v}\boldsymbol{d}_r||_D^2 \qquad b_r = \sum_{q,v} \Re e\langle\boldsymbol{\rho}_{q,v}, \boldsymbol{\phi}_{q,v}\boldsymbol{d}_r\rangle_D$$

$$a_i = \sum_{q,v} \frac{1}{\Omega_q^2}||\boldsymbol{\phi}_{q,v}\boldsymbol{d}_i||_D^2 \qquad b_i = \sum_{q,v} \frac{1}{\Omega_q^2}\Im m\langle\boldsymbol{\rho}_{q,v}, \boldsymbol{\phi}_{q,v}\boldsymbol{d}_i\rangle_D$$

$$c = \sum_{q,v} ||\boldsymbol{\rho}_{q,v}||_D^2 \qquad d_{ri} = \sum_{q,v} \frac{1}{\Omega_q^2}\Im m\langle\boldsymbol{\phi}_{q,v}\boldsymbol{d}_r, \boldsymbol{\phi}_{q,v}\boldsymbol{d}_i\rangle_D$$

$$A_r = \sum_{q,v} ||\boldsymbol{\phi}_{0q,v}\boldsymbol{d}_r||_D^2 \qquad B_r = \sum_{q,v} \Re e\langle\boldsymbol{D}_{q,v}, \boldsymbol{\phi}_{0q,v}\boldsymbol{d}_r\rangle_D$$

$$A_i = \sum_{q,v} \frac{1}{\Omega_q^2}||\boldsymbol{\phi}_{0q,v}\boldsymbol{d}_i||_D^2 \qquad B_i = \sum_{q,v} \frac{1}{\Omega_q^2}\Im m\langle\boldsymbol{D}_{q,v}, \boldsymbol{\phi}_{q,v}\boldsymbol{d}_i\rangle_D$$

$$C = \sum_{q,v} ||\boldsymbol{D}_{q,v}||_D^2 \qquad D_{ri} = \sum_{q,v} \frac{1}{\Omega_q^2}\Im m\langle\boldsymbol{\phi}_{0q,v}\boldsymbol{d}_r, \boldsymbol{\phi}_{q,v}\boldsymbol{d}_i\rangle_D$$

Si le critère n'est pas régularisé, alors la minimisation conjointe de $J(\boldsymbol{x}+\alpha_r^{(n)}\boldsymbol{d}_r^{(n)} + i\,\alpha_i^{(n)}\boldsymbol{d}_i^{(n)})$ implique la résolution numérique de deux équations de degré 3 et se ramène à la méthode d'inversion contraste source en séparant les parties réelle et imaginaire. Dans notre cas nous avons un terme de régularisation (J_{reg}), qui nous amène à la résolution numérique de 2 équations de degré 5. Ces deux méthodes seront comparées dans la prochaine section.

Deuxième étape : échantillonage de $\boldsymbol{z}$
En utilisant la formule de Bayes et la relation (11) on obtient :

$$p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{z})\, p(\boldsymbol{z}) \tag{15}$$

La loi $p(\boldsymbol{x}|\boldsymbol{z})$ étant séparable, la loi *a posteriori* $p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{x})$ reste un champ de Gibbs de voisinage égal à 4 pixels. L'échantillonage de $\boldsymbol{z}$ peut alors se faire en décomposant l'ensemble des pixels en pixels noirs et blancs formant un échiquier (équivalent des pairs et impairs en 1D). De cette manière, si l'on fixe les pixels noirs (respectivement blancs), alors les pixels blancs (respectivement noirs) sont indépendants et peuvent donc être échantillonés en même temps ([5]).

3 Résultats

Nous avons comparé notre méthode avec la méthode CSI sur deux jeux de données issus de l'institut Fresnel de Marseille (courtoisie de K. Belkebir et M. Saillard). Dans les deux cas le domaine de reconstruction est un carré de $17,5$ cm de côté, divisé en 51×51 pixels. Nous considérons quatre matériaux décrits dans le tableau 1. Le premier jeu de données provient

matériau	classe	moyenne		variance	
		m_{kr}	m_{ki}	σ_{kr}^2	σ_{ki}^2
vide	1	0	0	faible	faible
diél. $\varepsilon_r = 1,45$	2	$0,45$	0	faible	faible
diél. $\varepsilon_r = 3$	3	2	0	faible	faible
metal	4	0	$0,5$	forte	forte

TAB. 1: *Matéraiaux qui constituent l'objet inconnu*

d'un objet mixte composé d'un métal et d'un diélectrique ($\varepsilon_r = 1,45$). Les données acquises sont obtenues pour 18 angles de vues (tous les 20°) et pour 9 fréquences (de 3 à 11 GHz). La figure 2 montre les résultats obtenus. Notons de plus que dans le cas de notre méthode nous obtenons en même temps une

classification en régions correspondant à des matériaux distincts. Après convergence de l'algorithme l'objet reconstruit est constitué de zones quasiment constantes et la localisation des différents matériaux est bonne (fig. 2.g). En revanche on peut remarquer l'apparition d'un pic sur la partie réelle du métal pour les deux méthodes. Ce résultat est assez classique dans le cas d'un métal puisque l'onde ne le pénètre pas et la reconstruction de son intérieur en est difficile. Cependant l'information *a priori* introduite dans notre modèle améliore considérablement la reconstruction et la classification est très bonne, même si elle semble surestimer la taille des matériaux.

Cette dernière remarque se confirme avec le deuxième jeu de données qui provient d'un objet composé des deux diélectriques et du vide. Les données acquises sont obtenues pour 18 angles de vues et 9 fréquences (de 1 à 9 GHz). La figure 3 montre là encore une amélioration des résultats de reconstruction pour un objet compliqué, même si l'on peut observer une nette surestimation de la taille des objets. La figure 4 montre le résultat obtenu sur une ligne horizontale de l'image (à peu près au milieu) et montre alors que la détection et la localisation des matériaux est très bonne.

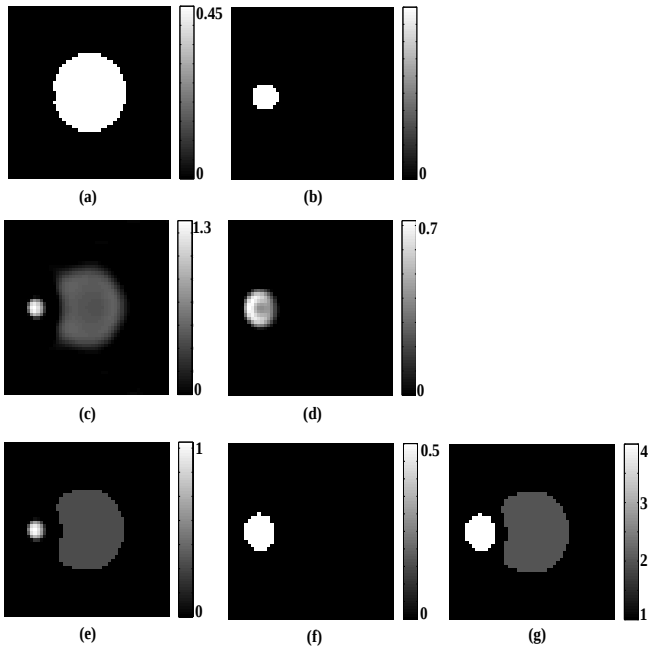


FIG. 2: *Résulat sur données réelles : (a,b) parties réelle et imaginaire du vrai contraste . (c,d) contraste reconstruit par la méthode d'inversion contraste source. (e,f) contraste reconstruit par la méthode proposée. (g) classification obtenue*

4 Conclusion

Nous avons proposé une méthode originale de reconstruction en tomographie micro-onde. Nous avons proposé une solution, dans le cadre d'une approche bayésienne, pour introduire une information *a priori* : le fait que nous connaissons les différents matériaux qui peuvent constituer l'objet. Nous avons alors montré les bonnes performances de notre méthode sur données réelles, en comparaison avec la méthode classique d'inversion contraste source sans terme de régularisation. Il est à noter que l'information *a priori* introduite est très forte. L'objectif dans la suite de ces travaux est d'une part de réduire cette information au fait que nous connaissons uniquement le nombre maximum de matériaux qui peuvent constituer l'objet, ce qui semble plus réaliste. D'autre part, l'étape de minimisation du critère régularisé estime conjointement les variables $\boldsymbol{w}_{q,v}$ et $\boldsymbol{x}$. Pour implémenter un algorithme d'échantillonage de Gibbs, il serait judicieux de se servir de la bilinéarité du modèle en ces variables pour effectuer un échantillonage exact de chacune d'entre elles indépendemment. Ces deux points consituent la suite de notre étude.

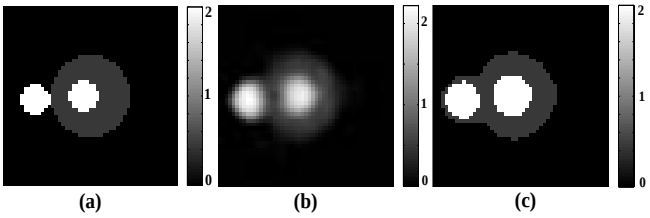


FIG. 3: *Résulat sur données réelles : (a) le vrai contraste. (b) contraste reconstruit par la méthode d'inversion contraste source. (c) contraste reconstruit par la méthode proposée.*

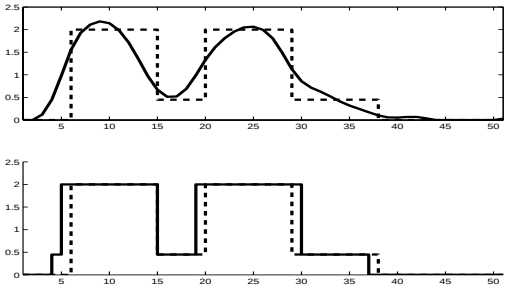


FIG. 4: *Coupe des résultats : comparaison entre le vrai gabarit (en pointillé) et le contraste reconstruit (en ligne pleine) par la méthode CSI (en haut) et par la méthode proposée (en bas)*

Références

[1] L. Souriau, B. Duchêne, D. Lesselier et R. E. Kleinman, « Modifi ed gradient approach to inverse scattering for binary objects in stratifi ed media », *Inverse Problems*, vol. 12, n° 4, pp. 463–481, août 1996.

[2] P. van den Berg, A. van Broekhoven et A. Abubakar, « Extended contrast source inversion », *Inverse Problems*, vol. 15, n° 5, pp. 1325–1344, octobre 1999.

[3] H. Carfantan, *Approche bay´esienne pour un problème inverse non lin´eaire en imagerie à ondes diffract´ees*, thèse de doctorat, Universit´e de Paris-Sud, Orsay, d´ecembre 1996.

[4] A. Baussard, K. Belkebir et D. Pr´emel, « A markovian regularization approach of modifi ed gradient method for solving a two dimensional inverse scattering problem », *Journal of Electromagnetic Waves and Applications*, vol. 17, pp. 989–1008, 2003.

[5] O. F´eron et A. Mohammad-Djafari, « Image fusion and unsupervised joint segmentation using HMM and MCMC algorithms », *accepted in Journal of Electronic Imaging*, novembre 2004.

[6] C. Robert, *M´ethodes de Monte-Carlo par chaînes de Markov*, Economica, Paris, 1996.

[7] O. F´eron et A. Mohammad-Djafari, « Reconstruction of piecewise homogeneous images from partial knowledge of their Fourier transform », in *MaxEnt04, Garching, Germany*, G. Erickson et Y. Zhai, Eds. août 2004, American Institute of Physics.

Restauration d'images par diffusion sélective

Aymeric HISTACE[1], Vincent COURBOULAY[2], Michel MÉNARD[2], Christine CAVARO-MÉNARD[1]

[1]Laboratoire d'Ingénierie des Systèmes Automatisés
62, avenue Notre Dame du Lac, 49000 Angers, France

[2]Laboratoire Informatique - Image - Interaction
Université de La Rochelle, Pôle Sciences et Technologie 17042 La Rochelle Cedex, France
histace@istia.univ-angers.fr, vcourbou@univ-lr.fr,
michel.menard@univ-lr.fr, christine.menard@univ-angers.fr

Résumé – La restauration anisotrope d'images par utilisation d'Équations aux Dérivées Partielles (EDP) a suscité et suscite toujours un intérêt important de par la possibilité qu'elle offre de lisser une image tout en préservant ses discontinuités. Les applications pratiques de ces méthodes de restauration sont nombreuses et touchent divers domaines (photographie, médical, etc.). Dans cet article, nous proposons une approche variationnelle pour la restauration d'images se fondant sur une fonctionnelle énergétique différente de celle rencontrée dans la littérature. L'EDP résultante se caractérise alors par la possible intégration dans le processus de diffusion d'information *a priori* sur des structures particulières de l'image que l'on désire restaurer. Nous présentons tout d'abord des résultats obtenus sur des images *ad hoc*, puis nous montrons la possible extension de ce processus de diffusion sélective à une application concrète : la restauration de séquences IRM cardiaques marquées.

Abstract – Anisotropic image restoration based on Partial Differential Equation (PDE) is of major interest for image regularization with preservation of the discontinuities. Applications are various and deal with numerous areas of image processing (photography, medical application, etc.). In this article we propose a variational approach for image restoration based on the minimization of an energetic function different from those encountered in the literature. The resulting PDE is characterized by the possible integration of *a priori* information on particular structures to restore during the diffusion process. We first present results obtained on *ad hoc* images, and then we show a possible extension of this research work to a particular application: the restoration of tagged cardiac MRI sequences.

1 Introduction

La restauration anisotrope d'images par approche variationnelle (EDP) a suscité et suscite toujours un intérêt important de par la possibilité qu'elle offre de lisser une image tout en préservant ses discontinuités. Depuis les travaux fondateurs de Perona et Malik [7], de nombreux enrichissements ont été apportés afin de répondre au mieux aux exigences de la restauration (élimination du bruit sans pour autant altérer les contours par exemple), en particulier pour les images en niveau de gris (ou scalaire) [1, 2, 4, 9] (entre autres).

D'une manière générale, si l'on note $\psi : \mathbb{R}^2 \times \mathbb{R}^+ \rightarrow \mathbb{R}$ la fonction luminance associée à une image scalaire, Diffuser l'information contenue dans l'image, c'est, entre autre, considérer un problème de minimisation d'une fonctionnelle E associée à ψ [4] et définie par :

$$E(\psi) = \int_{\Omega} \phi(\|\nabla\psi\|) d\Omega \qquad , \tag{1}$$

où $\phi : \mathbb{R} \rightarrow \mathbb{R}$ est une fonction croissante.

Dans le cas des images scalaires, la minimisation de E est équivalente à la juxtaposition de deux équations de la chaleur 1D orientées l'une dans la direction orthogonale au gradient de l'image, et l'autre dans la direction tangente :

$$\frac{\partial\psi}{\partial t} = \frac{\phi'(\|\nabla\psi\|)}{\|\nabla\psi\|}\psi_{\xi\xi} + \phi''(\|\nabla\psi\|)\psi_{\eta\eta} \qquad , \tag{2}$$

où $\eta = \nabla\psi/\|\nabla\psi\|$ et $\xi \perp \eta$.

Cette EDP se caractérise par un effet diffusif anisotrope effectué dans les directions privilégiées ξ et η et permet un débruitage de l'image. Cependant, elle ne permet pas la prise en compte de connaissances *a priori* sur d'éventuels motifs à préserver.

Nous nous proposons donc d'utiliser une fonctionnelle énergétique E différente, issue de la physique de la mesure [5] et récemment appliquée au traitement d'image [3]. En considérant les mêmes notations que précédemment, cette fonctionnelle est alors la suivante :

$$E(\psi) = \int_{\Omega}\int_{0}^{\infty} d\Omega dt \times \left[(\nabla - \mathbf{A})\,\psi.\,(\nabla - \mathbf{A})\,\psi + (\frac{\partial\psi}{\partial t})^2 - \psi^2\right]. \tag{3}$$

La particularité de cette fonctionnelle est d'intégrer un opérateur $\mathbf{A}$ associé à un potentiel vecteur permettant de contraindre la diffusion le long de lignes de niveaux données de l'image.

La minimisation de cette fonctionnelle permet l'obtention d'une EDP de diffusion particulière autorisant la prise en compte de connaissances *a priori* via le potentiel $\mathbf{A}$:

$$\frac{\partial\psi}{\partial t} = \frac{1}{2}(\nabla - \mathbf{A}).(\nabla - \mathbf{A})\psi \qquad . \tag{4}$$

C'est cette équation que nous nous proposons de paramé-

trer dans la section suivante, afin de montrer son potentiel sur des images mires dans un premier temps, puis dans le cadre particulier de la restauration d'images acquises en IRM cardiaques marquées.

2 Diffusion sélective

2.1 Paramétrisation de l'équation de diffusion

Afin d'obtenir un vecteur **A** susceptible de permettre une pondération de la diffusion, nous proposons de partir de la construction géométrique illustrée par la Figure 7 représentant, au pixel courant, le gradient associé ainsi que le vecteur **A**.

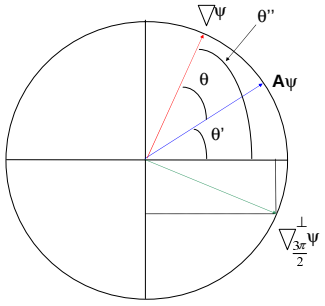


FIG. 1: *Construction géométrique de* **A** *et* $\nabla\psi$ *(à droite).*

L'idée sous-jacente est d'utiliser l'angle θ, caractéristique de la différence d'orientation entre le gradient local $\nabla\psi$ et **A**, comme facteur de pondération de la diffusion.

A partir de ces considérations géométriques, et compte-tenu des notations utilisées, nous pouvons donc exprimer **A** dans le repère global $(\mathbf{i}, \mathbf{j})$ sous la forme d'un opérateur particulier donné par :

$$\mathbf{A} = \nabla . \cos\theta + \nabla^{\perp}_{\frac{3\pi}{2}} . \sin\theta \qquad , \tag{5}$$

avec θ l'angle relatif entre $\mathbf{A}\psi$ et $\nabla\psi$.

Cette approche pour la construction de **A** permet donc bien de réaliser une diffusion sélective paramétrée par la différence d'orientation locale entre le gradient de l'image et **A** : ainsi, pour chaque pixel de l'image IRM marquée, si le gradient possède une orientation similaire au vecteur **A**, ce pixel ne sera pas diffusé. Dans le cas contraire, l'information sera lissée. Si maintenant, nous imposons l'orientation θ' de **A** de manière cohérente par rapport aux orientations des gradients à ne pas diffuser (du moins que l'on veut préserver), il sera alors possible de régulariser l'image sous ces conditions.

2.2 Exemple de diffusion sélective

Considérons une image représentant des lignes verticales, horizontales et diagonales orientées à 45°, le tout sur un fond uniforme de luminance moyenne. En imposant les valeurs de 135° et 315° pour l'orientation de **A**, il est possible, théoriquement, de diffuser uniquement les informations dont les gradients caractéristiques ne possèdent pas ces deux valeurs. La Figure 2 montre le résultat de diffusion sélective obtenu sur une telle image pour un pas temporel $dt = 0,2$ et pour 100 itérations.

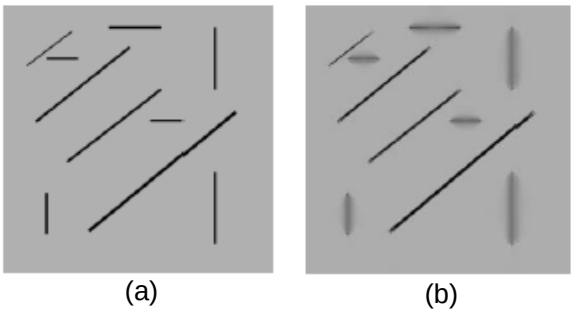


FIG. 2: *Diffusion d'une mire* ad hoc. *(a) image originale, (b) image diffusée.*

Nous constatons sur cette figure la disparition progressive des lignes non diagonales, celles étant orientées à 45° étant préservées. Un léger effet de bord se manifeste aux extrémités des lignes diagonales, mais ce phénomène est dû à la valeur de l'orientation locale des gradients de ces pixels qui n'est ni de 135°, ni de 315°.

Si le même procédé de diffusion, enrichi de deux orientations supplémentaires pour θ' (45° et 225°), est appliqué à une grille superposée sur une image uniforme bruitée par un bruit gaussien (Figure 3), il est alors possible de mettre en évidence l'effet débruitant du processus de diffusion anisotrope que nous proposons (Figure 4).

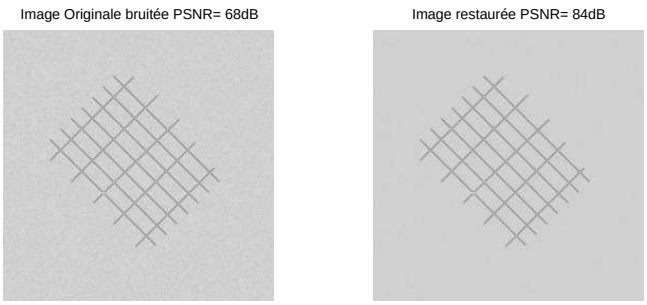


FIG. 3: *Diffusion d'une grille régulière.*

La grille n'est pas altérée par le processus de diffusion et le *PSNR* (*Peak Signal to Noise Ratio*) augmente de 68 dB à 84 dB : le bruit gaussien ajouté est donc, itération après itération, éliminé. Cet effet est manifeste sur deux profils lignes extraits respectivement de l'image bruitée et de l'image restaurée (Figure 5).

Ces deux exemples simples nous permettent de montrer le potentiel de la méthode de diffusion anisotrope. Ces exemples simple nous permettent de montrer le potentiel de la méthode de diffusion sélective.

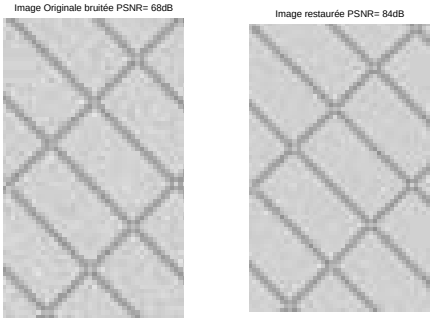


FIG. 4: *Diffusion d'une grille bruitée (agrandissement d'une région d'intérêt particulière).*

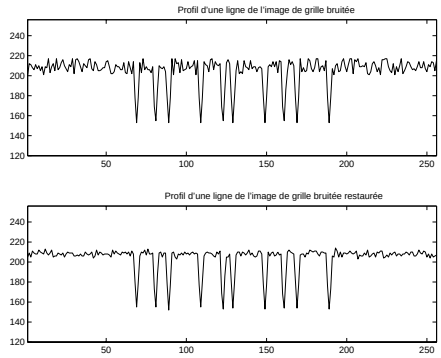


FIG. 5: *Comparaison de deux profils linéaires avant et après débruitage par l'équation de diffusion proposée.*

3 Application

Dans un cadre plus pratique, nous avons utiliser cette équation de diffusion afin de restaurer des images médicales issues d'une acquisition en IRM cardiaque marquée (Figure 6).

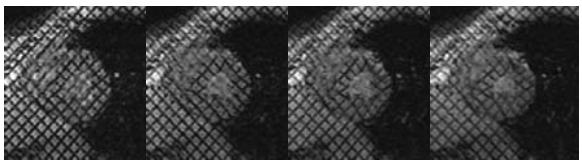


FIG. 6: *Quelques images d'une séquence Petit-Axe (PA) acquise en IRM cardiaque marquée (de la télédiastole à la télésystole).*

Le but est de régulariser les images des séquences tout en préservant le motif de marquage prenant la forme d'une grille orientée à 45 degrés. Cette particularité du motif de marquage, résultat du protocole d'acquisition défini au CHU d'Angers par les radiologues, nous amène à prendre en compte des caractéristiques précises quant à la restauration des IRM cardiaques marquées. En effet, les *tags* intègrent l'image de manière périodique, avec un espacement et une épaisseur connus, et surtout possèdent une orientation égale à $\pm$ 45°. Cette dernière propriété est particulièrement intéressante dans le sens où il est alors possible de caractériser chaque ligne de *tags* par deux gradients moyens qui leur sont propres : un premier orienté à $\theta_{tag} + \pi/2$ et l'autre à $\theta_{tag} + 3\pi/2$ (Figure 7).

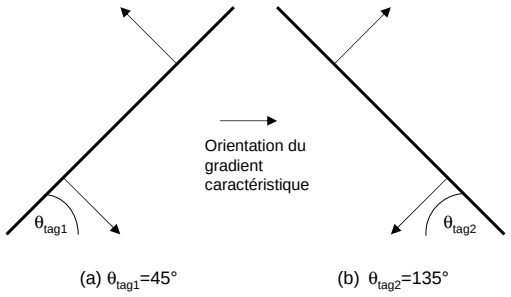


FIG. 7: *Gradients caractéristiques des lignes de* tags.

Nous proposons donc de prendre en compte l'orientation moyenne des gradients des *tags* afin de pondérer la diffusion localement en intégrant cette propriété dans la construction de l'opérateur **A**.

Un calcul local de **A** sur les images filtrées fréquentiellement (Figure 8) [10] et selon la méthode développée dans [8], nous a permis d'obtenir les résultats présentés Figures 9 et 10. Le pas temporel de diffusion est $dt = 0,2$ pour un nombre d'itérations égal à 60. Afin de montrer des résultats appréciables visuellement, nous avons choisi de ne traiter qu'une direction des *tags* sur les images extraites de séquences complètes.

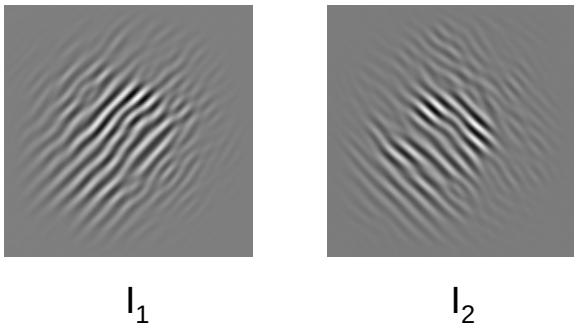


FIG. 8: *Images obtenues par filtrage dans le domaine de Fourier.*

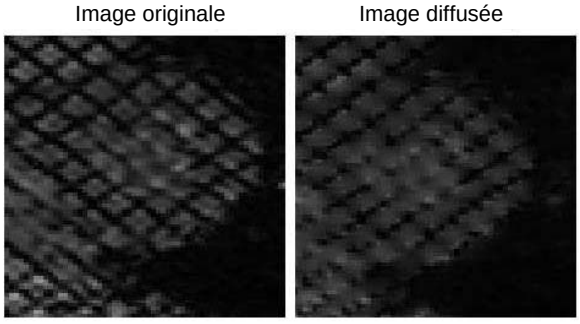


FIG. 9: *Préservation des* tags *orientés à* 45° *sur une IRM cardiaque marquée par le processus de diffusion proposé.*

Comme nous pouvons le voir sur la Figure 9, le schéma

de diffusion rend possible l'élimination progressive des artéfacts dus au bruit et également de l'information image non relative aux *tags* (*i.e.* comprise entre deux lignes de *tags* consécutives).

De plus, le fait que **A** soit calculé de manière locale sur un voisinage particulier rend effectivement possible l'application du processus pour des temps t différents de l'instant initial t_0 : la déformation locale étant prise en compte, le schéma de diffusion reste efficace durant toute la contraction systolique (Figure 10).

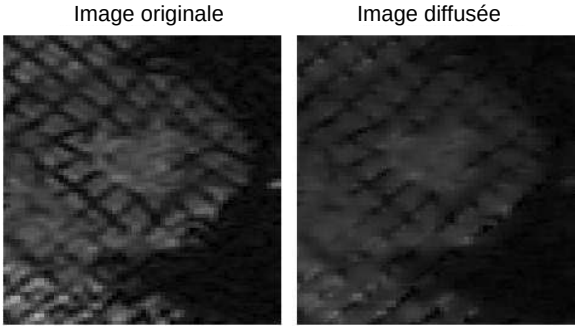
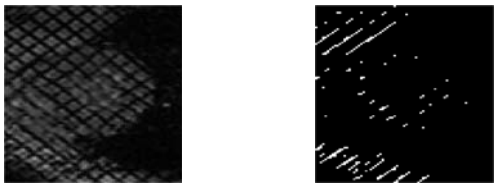
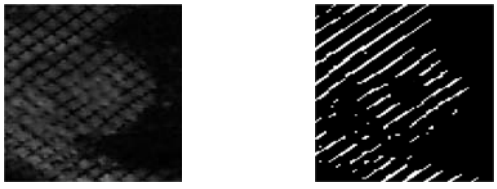


FIG. 10: *Diffusion d'une IRM marquée pour $t \neq t_0$.*

La réalisation d'une détection de vallées au moyen du calcul de la matrice hessienne locale dans le cadre de notre application en IRM marquée permet l'obtention de résultats qualitatifs intéressants : visuellement, si nous comparons une détection de vallée effectuée sur l'image *taggée* originale à la même détection réalisée sur l'image diffusée 60 fois pour un pas temporel $dt = 0.1$, il est possible de constater l'effet restaurateur de notre méthode de diffusion ; un plus large nombre de pixel appartenant aux *tags* orientés à 45 degrés sont détectés sur l'image restaurée (Figure 11).



(a) Image originale (b) Détection de vallées



(c) Image diffusée (d) Détection de vallées

FIG. 11: *Détections de vallées sur l'image originale et sur l'image rehaussée*

Au final, nous utilisons les images restaurées comme cartes de potentiels dans un modèle de contours actifs (sous forme de grille de *B-splines*) pour la détection et le suivi de la grille de marquage [6].

4 Conclusion

Cette technique de diffusion sélective présente donc un potentiel certain. Nous travaillons actuellement à généraliser son utilisation à des applications autres que l'imagerie médicale, comme en restauration de documents écrits par exemple.

Références

[1] F. Catté, T. Coll, P. Lions, and J. Morel. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM Journal of Applied Mathematics*, 29(1):182–193, 1992.

[2] G. Cottet and L. Germain. Image processing through reaction combined with non-linear diffusion. *Mathematics of Computation*, 61(204):659–673, 1993.

[3] V. Courboulay, M. Ménard, M. Eboueya, and P. Courtellemont. Une nouvelle approche du filtrage linéaire optimal dans le cadre de l'information physique extrême. In *RFIA 2002*, pages 897–905, Janvier 2002.

[4] R. Deriche and O. Faugeras. Les edp en traitements des images et visions par ordinateur. *Traitement du Signal*, 13(6):551–578, 1996.

[5] B.R. Frieden. *Physics from Fisher Information*. Cambridge University Press, 1998.

[6] A. Histace, V. Courboulay, C. Cavaro-Ménard, and M. Ménard. Détection et suivi des déformations intramyocardiques en irm cardiaque marquée. In *Actes du congrès GRETSI 2003*, volume 3, pages 316–320, 2003.

[7] P. Perona and J. Malik. Scale-space and edge detection using anistropic diffusion. *IEEE Transcations on Pattern Analysis and Machine Intelligence*, 12(7):629–639, 1990.

[8] O. Baylou P. Borda M. Terebes, R. Lavialle. Mixed anisotropic diffusion. In *Proceedings of the 16th International Conference on Pattern Recognition*, volume 3, pages 1051–4651, 2002.

[9] J. Weickert. Coherence-enhancing diffusion of colour images. In *Proceedings seventh National Symposium on Pattern Recognition and Image Analysis*, Barcelona, 1997.

[10] S. Zhang, M. Douglas, L. Yaroslavsky, R. Summers, V. Dilsizian, L. Fananapazir, and S. Bacharach. A fourier based algorithm for tracking SPAMM tags in gated magnetic resonance cardiac images. *Medical Physics*, 32(8):1359–1369, 1996.

Denoising and Migration Techniques for Target Identification from Ground Penetrating Radar 2D data; a case study

Olga LOPERA[1,2], Nada MILISAVLJEVIC[1], Bart SCHEERS[1], Benoît MACQ[3]

[1]Signal and Image Centre, Royal Military Academy
30 Av de la renaissance, 1000 Brussels, Belgium

[2]Dept. of Electrical and Electronic Engineering, University of Los Andes
Kr 1 Este 18A-10 Bogota, Colombia

[3]Telecommunications Laboratory, Catholic University of Louvain
2 Place du Levant, 1348 Louvain-la-Neuve, Belgium
olopera@elec.rma.ac.be

Résumé – Le radar de pénétration du sol (GPR) est une technique de télédétection employée pour obtenir l'endroit et la réflectivité spatiaux des objets enterrés. Puisque la plupart des antennes de GPR ne sont pas directives, les signaux dispersés enregistrés par le radar se prolongent au-dessus d'une grande ouverture latérale [1, 2]. Dans cette étude, des algorithmes de débruitage et de migration sont employés pour refocaliser les signaux dispersés de nouveau à leur point d'origine. Les données ont été prises pour différents scénarios. Afin de réaliser la séparation optimale de la signature de la cible de la réponse du soil, techniques de débruitage son utilisées sur les données 2D. La transformée de Hough randomisé est employé pour extraire des informations importantes [3]. Ces informations sont incluses dans un algorithme de migration [4], et la largeur aproximée de l'objet dans la direction du balayage est trouvée. Bien que les résultas sont pormetteurs, les algorithmes doivent toujours être validés dans différentes conditions.

Abstract – Ground-penetrating radar (GPR) is a remote sensing technique used to obtain the spatial location and reflectivity of subsurface objects. Since most GPR antennas have board beam widths, the scattered signals recorded by the radar extend over a large lateral aperture [1, 2]. In this study, denoising and migration algorithms are used to refocus the scattered signals back to their point of origin. The data were collected for different scenarios. In order to achieve optimal separation of the target signatures from the background, denoising techniques are used to filter the signal from the 2D data. The randomized Hough Transform is used to extract relevant features [3]. This extracted information is included in migration algorithms [4], and the approximated width of the target in the scan direction is found. Although the results are promising, algorithms still need to be validated under different conditions.

1 Introduction

Since the WWII, many military conflicts around the world have left more than 100 millions of antipersonnel mines (APM) landed across the globe [5]. In Colombia, four decades of civil war have left more than 100 thousand antipersonnel (AP) mines landed across 422 municipalities (40% of the national territory). These mines cause approximately 150 deaths per year, which a 31% are children [6]. These mines are American and Belgian, as well as mines that were made for the National Military Industry INDUMIL. However, a considerable amount was made in an artisan way by the guerrillas (*e.g.*, FARC and ELN[1]) . These last ones are called improvised explosive devises (IEDs), and are the principal problem in Colombian demining activities, due to their non-metallic content. In recent years, serious efforts for developing technologies that can help in landmine detection have been done [7]. Among these technologies, a great importance is given to the GPR because of the ability to detect metallic and low-metallic AP landmines by non-invasive subsurface sensing. Still, its main drawback is the complex nature of its data, and then their interpretation is usually limited in defining general "areas of interest" instead of accurately determining the shape and position of the target.

In this study, two-dimensional data-sets (B-scans) from a commercial time-domain GPR (800 MHz) are analyzed in order to extract relevant information. These data-sets are collected in different scenarios. In Section II, a description of the targets and the measurements setup are given. In order to achieve optimal separation of the target signatures from the background, the problem is viewed as an image-processing one as the collected two-dimensional data-sets represent image segments. Section III is devoted to the denoising techniques used to remove unwanted reflections and to extract relevant pieces of information as target position and depth, and the dielectric permittivity (real part) of the soil. Migration theory is then applied for focusing the scattered signals back in their point of origin. The Stolt migration is used and is introduced in Section IV. The analysis of the migrated data is presented in Section V. Finally, Section VI concludes our study.

2 Measurement Setup

The data are acquired at the test land of the Military School of Engineering, Colombia, for sandy and loamy soils (see Fig. 1). In this study we focus on the signatures of two types of

1. Fuerzas Armadas Revolucionareias de Colombia - FARC, Ejercito de Liberacion Nacional - ELN

FIG. 1: Test land at the Military School of Engineering, Colombia

AP mines: a Colombian AP mine and an improvised explosive devise made by the guerrillas. In the remainder of this paper these mine types will be referred to as APM A and B respectively. AP mine A has a plastic casing and low metallic content. AP mine B has a plastic casing and non-metallic content. Their dimensions, burial depths (from the top of the mine), and distance (from a reference point to the middle of the mine along the scanning direction) are summarized in Table 1.

For the acquisition we used a monostatic radar system in the time-domain developed by the Mala GeoScience Group [2], with a 800MHz antenna and a repetition frequency of 100kHz. All data have been recorded using a 10 ns time window of 512 samples corresponding to a time sampling interval of approximately 20ps. The resolution of each sample point is 16 bit. The data have been acquired along one surface direction and the antenna system was positioned above the soil surface. The distance was measured by an incremental encoder connected to the system and the antenna was moved whit a constant velocity. In Fig. 2, the collected data for the loamy soil is represented in function of the scanning distance and the time-of-flight of the signal (time to reach the reflection point and to come back to the receiver antenna) for the two types of mines.

TAB. 1: *Target features**

Target	Dimension (m)	Depth (m)	Distance (m)
APM A	d:0.075, h:0.07	0.01	0.60 (l), 0.60 (s)
APM B	d:0.120, h:0.07	0.10	0.60 (l), 0.60 (s)

* d: diameter, h: height, (l): loamy soil, (s): sandy soil

3 Ground Clutter Subtraction

To eliminate the clutter from an image we have first to define it. The electromagnetic waves transmitted by the radar are propagated through the air and, of course, to the ground. At any interface between any two media having different electromagnetic properties, part of the electromagnetic energy is reflected backwards to the receiver and a part is transmitted into the second media. The same phenomenon occurs when the transmitted signal comes across a buried object (of which electromagnetic parameters differ from those of the ground) and, unfortunately, across any other non-uniformities inside the ground (stones, layers having different electromagnetic parameters, etc). All these reflections which are not related to the object's own scattered field are called clutter [8].

Two denoising methods are used to filter the ground clutter level. These methods are implemented by convolving the image with a horizontal high pass filter. The derivative of a Gaussian represents such a filter. It is important to notice that the window width of the filter is crucial for the final result. The choice of the filter length depends on the size and resolution of the image. Several tests show that a too small window introduces noise, and a large window minimizes the contrast of the image [3]. We achieved the best results with a 10-pixels wide filter. Every change in the image resolution or image size needs a reevaluation of the ideal filter width. Note that a vertical noise was introduced to the image (because horizontal filtering is a derivative action). To reduce this effect, a vertical 1x5-pixels filter was created by a Gaussian function. This filter is used to average over 5 pixels in order to reduce the noise in the vertical direction. Once the image is filtered, we can see the hyperbolic response of the target. An algorithm to detect curves is applied for extracting relevant information from the equation of the hyperbola. The Randomized Hough Transform is used [3]. RHT randomly selects n pixels from an image and fits them to a parameterized curve. If the pixels fit within a tolerance, they are added to an accumulator with a score. Once a specified number of pixel sets are selected, the curve with the best score is selected from the accumulator and its parameters are used to represent a curve in the image. In RHT, if a curve in the accumulator is similar to the curves being tested, the parameters of the curves are averaged together and the new average curve replaces the curve in the accumulator. The extracted information form the hyperbola equation in included in the migration algorithm.

4 Phase-shift Migration

Phase-shift migration was first applied for seismic imaging by Stolt in 1978 [9]. The principle has been adopted for satellite SAR processing [10] and it has also been widely utilized for processing of GPR data in two dimensions [11]. Phase-shift migration is based on a transformation from the frequency domain ω to the wavenumber domain k. Consider the raw data set $b(x, z, t)$ collected from the radar, x being the B-scan distance, z the depth, and t the time-of-flight. Applying a 2D Fourier transform with respect to the spatial distance x and the time t to spatial frequency k_x, the result is an unfocused wavenumber data set

$$B(k_x, z, \omega) = \int\int b(x, z, t) e^{ik_x x - i\omega t} \mathrm{d}x\mathrm{d}t. \qquad (1)$$

The Fourier transformation along the x coordinate only makes sense if the propagation velocity does not vary in this direction. The method allows variations of the propagation velocity in the z direction. Defining the wavenumber vector k as the vector sum of k_x and k_z for one-way propagation, we have:

$$k = |\mathbf{k}| = \sqrt{k_x^2 + k_z^2} = \frac{\omega}{\upsilon} = \frac{2\pi}{\lambda}, \qquad (2)$$

2. http://www.malags.se

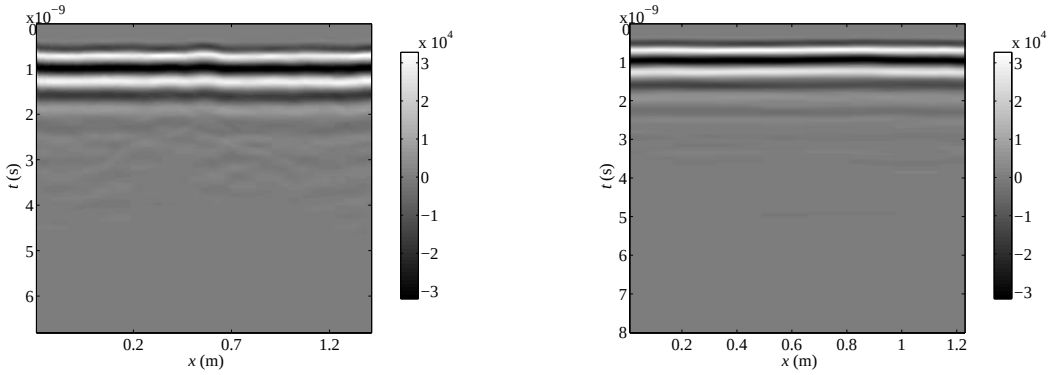


FIG. 2: *B-scans on loamy soil for APM A (a) and B (b).*

where v is the propagation velocity of the ground ($v = c/\sqrt{\varepsilon_r}$, being ε_r the relative dielectric permittivity, extracted from the equation of the hyperbola) and λ is the wavelength in the ground. The direction of the **k**-vector is identical to the traveling direction of a plane wave propagating from the target to the antenna. Assuming only upward coming waves, and by introducing k_z from (2) in (1), the Fourier transform of the wavefront at depth z is done by

$$B(k_x, z, \omega) = B(k_x, 0, \omega)e^{-ik_z z}. \quad (3)$$

The migrated data will be the inverse Fourier transform of (3) at time t=0:

$$\widehat{b}(x, z) = b(x, z, 0) = \int\int B(k_x, 0, \omega)e^{-i(k_x x + k_z z)}\mathrm{d}k_x\mathrm{d}\omega. \quad (4)$$

Equation (4) is the general representation of the phase-shift migration. The implementation of this method is computationally intensive, because of the number of floating point operations needed for migration [2]. For the reduction of the calculation time, a variant of the phase-shift migration for a constant propagation velocity is used. This variant was developed in [9]. In the special case where $v(z)$ is constant, (4) can be further developed by changing the variable $\mathrm{d}\omega$ to $\mathrm{d}k_z$. By replacing $\mathrm{d}\omega$ from (2), the data must be scaled by the Jacobian of the transformation from ω to k_z, $\frac{k_z v^2}{\omega}$. Hence, for the Stolt migration, (4) becomes

$$\widehat{b}(x, z) = v^2 \int\int \frac{k_z}{\omega} B(k_x, 0, \omega)e^{-i(k_x x + k_z z)}\mathrm{d}k_x\mathrm{d}k_z. \quad (5)$$

5 Results and Discussions

In Fig 2, the B-scans on loamy soil for AP mine A and B are shown. The reflection of the buried object is completely drowned in the ground reflection (between 0.5 ns and 1.5 ns). By applying denoising and segmentation techniques, relevant information can be extracted for the data. It is summarized in Table 2. It can be seen that the extracted values are quite close to real ones. The differences between the real values and the calculated ones are due to the three-dimensional form of the APM. The depth and distance were measured respectively from the top and from the middle of the mine. However, the backscattered signal received by the antenna comes from different points over the surface of the mine [2], and then different hyperbolas will be appear in the B-scan. By averaging these different hyperbolas, we will find a very good approximation of the position of the mine, but not the exact one.

TAB. 2: *Calculated depth and distance**

Target	Depth (m)	Distance (m)
APM A	0.0078	0.671 (l), 0.624 (s)
APM B	0.1056	0.652 (l), 0.636 (s)

* (l): loamy soil, (s): sandy soil

In Fig 3, results of applying the migration algorithm for the denoised data are shown. From this figure, the shape of the target on the B-scan direction can be seen. For the APM A, it appears at 0.5 ns. This points out that the AP mine is placed at the soil surface. For the AP mine B, it appears at 1.0 ns. This points out that the mine is buried close to the surface ($\sim$ 10 cm). Over the scanning axis, the position of the target (the shape of the target) appears quite close to the calculated and real position of the target in the soil.

6 Conclusions

In this paper, denoising and migration methods have been applied on real GPR data. The data are collected in two types of soil and for conventional and non-conventional APM. Namely, the major limit factor to detect APM with GPR is the ground clutter. To reduce the ground clutter an horizontal filtering is applied. Since the buried object signature on GPR 2D data are represented by a hyperbolic shape, the RHT is applied for the isolation of the object response. The relevant pieces of information that have been extracted from the equation of the hyperbola give us a very good approximation of the position of the buried object. The size of the object can be deduced from the migrated data. All this information can be used in data fusion architectures, involving other sensor data. The performance of these algorithms on real GPR data is encouraging. These algorithms are not computationally intensive. However, it depends on the

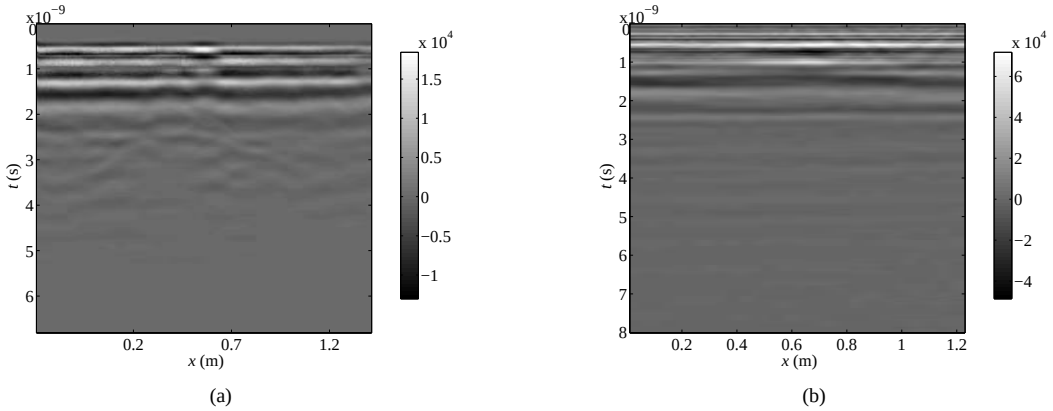


FIG. 3: *Results of applying denoising and migration techniques for B-scans on loamy soil for APM A (a) and B (b).*

size of the database. In order to make an objective analysis, the overall processing still needs to be validated in different conditions and for more than one object buried in the soil.

Acknowledgment

This research project is carried out at the Signal and Image Centre of the Royal Military Academy, Belgium, and at the Department of Electrical and Electronic Engineering of the University of Los Andes, Colombia, in collaboration with the Telecommunications Laboratory, Catholic University of Louvain, Belgium. It is supported by a research grant funded by the Ministry of Defence of Belgium, in the scope of the BEMAT project, and by a research grant from Colciencias (Colombian Institute for the Development of Science and Technology, Colombia). The authors would like to thank the Colombian Military Forces for their support, specially the personnel of the MARTE group and the BAMAI laboratory.

Références

[1] N. Milisavljevic, "Analysis and fusion using belief functions theory of multisensor data for close-range humanitarian mine detection," Ph.D. dissertation, Ecole Nationale Superieure des Telecommunications, France - Royal Military Academy, Belgium, 2001.

[2] B. Scheers, "Ultra-wideband ground penetrating radar with application to the detection of anti personnel landmines," Ph.D. dissertation, Catholic University of Louvain - Royal Military Academy, Belgium, 2001.

[3] O. Lopera, N. Milisavljevic, B. Macq, I. van den Bosch, S. Lambot, and A. Gauthier, "Analysis of segmentation techniques for landmine signature extraction from Ground Penetrating Radar 2D data," in *eProceedings of the II International IEEE Andean Region Conference, ISBN 958-33-6534-3*, IEEE, Ed., Bogota, Colombia, 2004.

[4] O. Lopera, S. Lambot, N. Milisavljevic, B. Scheers, and I. van den Bosch, "Background subtraction in the frequency domain for focusing ground-penetrating radar data," *Near Surface Geophysics*, vol. Special Issue, p. Submitted, 2005.

[5] GICHD, *A guide to Mine Action*. Geneva: GICHD., 2003.

[6] Human Rights Watch, *Landmine Monitor Report 2002, Toward a Mine-free World*. Whashington D.C.: ICBL, 2002.

[7] J. MacDonald, J. Loockwood, T. Altshuler, T. Broach, L. Carin, R. Harmon, C. Rappaport, W. Scott, and R. Weaver, *Alternatives for Landmine Detection*. USA: RAND, 2003.

[8] A. Yarovoy, V. Kovalenko, and A. Fogar, "Impact of clutter on buried object detection by Ground Penetrating Radar," in *International Geoscience and Remote Sensing Symposium*, France, 2003.

[9] R. Stolt, "Migration by fourier transform," *Geophysics*, vol. 43, pp. 23–48, 1978.

[10] C. Cafforio, C. Prati, and E. Rocca, "SAR data focusing using seismic migration techniques," *IEEE Trans. Aeroespace and Electronic Systems*, vol. 27, pp. 197–206, 1991.

[11] H. Yu, X. Ying, and Y. Shi, "The use of FK-techniques in GPR processing," in *6th International Conference on Ground Penetrating Radar*, Sendai, Japan, 1996, pp. 595–600.

Opérateurs tomographiques pour la reconstruction d'objets à symétrie de révolution

François Lavallou, Jean Michel Lagrange

CEA, centre DAM Ile de France
BP 12, 91680 Bruyères-le-Châtel, France
prenom.nom@cea.fr

Résumé – Cet article propose de nouveaux résultats pour la reconstruction d'objets à symétrie de révolution à partir de leurs projections parallèles en exploitant quelques propriétés d'une famille d'opérateurs tomographiques issue des opérateurs de Weyl. Notant f la fonction densité de l'objet et $\mathcal{Y}$ une projection perpendiculaire à l'axe de symétrie, nous donnons une nouvelle expression de l'inversion d'Abel et le moyen de calculer les moments de f en fonction de $\mathcal{Y}$. Nous étudions aussi les transformations à noyau intégral de f et déterminons une nouvelle relation permettant leur évaluation à partir de $\mathcal{Y}$ seulement. Nous illustrons ce procédé par le calcul numérique de la transformée en cosinus d'une fonction f à l'aide de sa projection $\mathcal{Y}$.

Abstract – This article provides new results for the reconstruction of radial objects from their parallel projections by introducing the properties of a particular class of tomographic operators stemming from the Weyl operators. Introducing f the density function of the object and $\mathcal{Y}$ the projection normal to its symmetry axis, we present a new formulation for the Abel inverse operator and indicate how to calculate the moments of f as a function of $\mathcal{Y}$. We also study the integral kernel transforms of f and determine a new relation allowing their evaluation from $\mathcal{Y}$ only. As a conclusion we apply our method to the numerical evaluation of the cosine transform of a function f given its projection $\mathcal{Y}$.

1 Introduction

Dans certaines expériences radiographiques réalisées au CEA/DAM, les contraintes liées au dispositif d'émission et de mesure d'un rayonnement X très dur sont telles que nous ne pouvons disposer que d'une unique vue des objets, recueillie sur un détecteur plat de type couple écran renforçateur/film ou bien sur une γ caméra. En imposant de fait une hypothèse de symétrie cylindrique (parfaitement adaptée à nos objets), ce problème de reconstruction tomographique en rayonnement parallèle peut être résolu et a d'ailleurs fait l'objet de nombreux travaux proposant des approches inverses (rétroprojection filtrée, Abel inverse [1], inversion de la matrice de projection [2]) ou directes (par minimisation d'un critère pénalisé [2] ou par déformation d'un modèle des objets [3] [4]).

Dans cet article, nous fournissons de nouveaux résultats sur la résolution du problème de reconstruction à partir des projections parallèles d'un objet à symétrie radiale en utilisant une famille d'opérateurs $\{A_n\}_{n\in\mathbb{N}^*}$ qui généralise l'opérateur d'Abel [1], noté $\mathcal{A}$ par la suite.

Dans la section 2, nous décrivons les principales propriétés de cette famille A_n, et en particulier la composition qui permet de faire le lien entre l'opérateur d'Abel inverse $\mathcal{A}^{-1}$ et la "reprojection des projections" d'un objet. Nous donnons également une expression de tous les moments de la solution du problème (la fonction densité f) à partir de ces projections parallèles itérées.

Dans la section 3, nous présentons une méthode d'intégration par parties qui permet de calculer des transformations à noyau intégral du type $\int K(x,y).f(y)dy$ à partir des projections de f. Pour illustration, nous en déduisons une expression de la transformée en cosinus ($K(x,y)=\cos(xy)$) de f qui constitue un résultat concurrent au théorème coupe-projection.

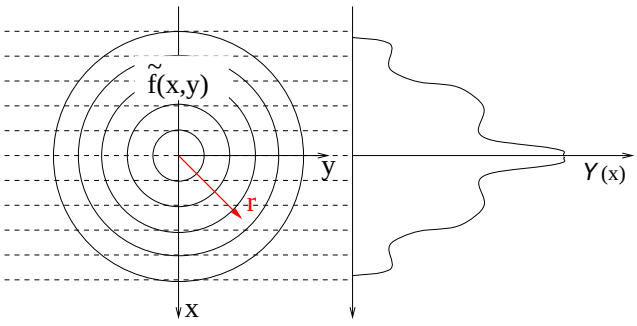


Fig. 1 – $\tilde{f}(x,y)$ représente la carte de densité de la coupe plane d'un objet à symétrie cylindrique et nous avons $f(r)=\tilde{f}(r\cos\theta, r\sin\theta), \forall\theta$.

La projection parallèle de cet objet (voir fig. 1) est notée $\mathcal{Y}$ avec :

$$\mathcal{Y}(x)=\int_{\mathbb{R}}\tilde{f}(x,y)dy$$

Dans le cas cylindrique, cette projection $\mathcal{Y}$ est la transformée d'Abel $\mathcal{A}$ de la fonction f :

$$\mathcal{Y}(x)=\mathcal{A}[f](x)=2\int_x^\infty \frac{r}{\sqrt{r^2-x^2}}f(r)dr$$

2 Opérateurs tomographiques

Pour toute fonction f de L^1 à support compact, nous définissons pour tout entier $n > 0$ les opérateurs d'Abel généralisés ou opérateurs tomographiques suivants, complétés, par prolongement analytique, par l'identité en composition pour $n = 0$:

$$A_n[f](x) = \frac{2}{\Gamma(\frac{n}{2})} \int_x^\infty u.f(u).(u^2 - x^2)^{-1+\frac{n}{2}} du \qquad (1)$$

Cette famille d'opérateurs, issue de l'étude de la tomographie divergente d'un objet à symétrie de révolution [6], est liée à une sous famille des opérateurs de Riesz [7][8], les intégrales de Weyl [9] notées $(W_\alpha)_{\alpha>0}$, dont nous utilisons les propriétés suivantes [10] :

$$W_\alpha \circ W_\beta = W_{\alpha+\beta}$$

$$\frac{d}{dx} W_{\alpha+1}[f](x) = -W_\alpha[f](x)$$

$$W_\alpha[f](x) = \frac{1}{\Gamma(\alpha)} \int_x^\infty f(u).(u - x)^{-1+\alpha} du$$

En notant $\tau_p[f](x) = f(x^p)$ et en définissant les opérateurs d'Erdélyi-Kober [11] :

$$\mathcal{W}_{\alpha,p} = p^{-\alpha} \tau_p \circ W_\alpha \circ \tau_{\frac{1}{p}},\ \alpha \in I\!R, p > 0$$

nous avons, en particulier pour $p = 2$:

$$A_n[f](x) = 2^{\frac{n}{2}} \cdot \mathcal{W}_{\frac{n}{2},2} = W_{\frac{n}{2}} \left[\tau_{1/2}[f]\right](x^2) \qquad (2)$$

avec $A_1[f](x) = \frac{1}{\sqrt{\pi}} \mathcal{A}[f](x) = \frac{1}{\sqrt{\pi}} \mathcal{Y}(x)$

Par l'intermédiaire de la relation (2), les propriétés des opérateurs d'Erdélyi-Kober se transposent directement aux opérateurs tomographiques A_n.

Propriété 1 *La composition entre opérateurs tomographiques est un homomorphisme additif :*

$$A_n \circ A_m = A_{n+m} \qquad (3)$$

et, en conséquence, l'opérateur A_1 composé n fois vaut A_n :

$$(A_1)^n \stackrel{def}{=} \underbrace{A_1 \circ A_1 \circ ... \circ A_1}_{n\ fois} = A_n$$

En recherchant des correspondances entre des décompositions de l'objet et de sa projection dans des bases, nous avons obtenu la propriété suivante :

Lemme 1 *L'opérateur adjoint de A_n étant donné par :*

$$A_n^*[f](x) = \frac{2 \cdot x}{\Gamma(\frac{n}{2})} \cdot \int_0^x f(u).(x^2 - u^2)^{-1+\frac{n}{2}} du$$

l'image de la base de Fourier par A_n^* est donnée par :

$$\begin{cases} A_n^*[\cos mx](x) = \frac{\sqrt{\pi}}{2}.m.J_{\frac{n-1}{2}}(mx).\left[\frac{2x}{m}\right]^{\frac{n+1}{2}} \\ A_n^*[\sin mx](x) = \frac{\sqrt{\pi}}{2}.m.\mathcal{H}_{\frac{n-1}{2}}(mx).\left[\frac{2x}{m}\right]^{\frac{n+1}{2}} \end{cases} \qquad (4)$$

avec J_n une fonction de Bessel de première espèce et $\mathcal{H}_n$ une fonction de Struve.

Cet opérateur adjoint, lié aux opérateurs de Riemann-Liouville (duaux de W_n), introduit, d'après (4), la transformée de Hankel TH_n d'ordre n par projection de l'opérateur tomographique A_{2n+1} sur une base trigonométrique :

$$(A_{2n+1}[f], \cos(mx)) = \left(f, A_{2n+1}^*[\cos(mx)]\right)$$

$$= \sqrt{\pi} \left(\frac{2}{m}\right)^n TH_n[x^n f(x)](m)$$

Ce résultat est établi par une autre voie dans [8] pour $n = 0$.

La propriété qui suit est fondamentale pour les opérateurs A_n et résulte d'une intégration par parties de la définition 1.

Propriété 2 *Pour tout entier n supérieur ou égal à 2, la dérivée de $A_n[f]$ s'obtient à partir de $A_{n-2}[f]$:*

$$A_n[f]'(x) = -2x A_{n-2}[f](x) \qquad (5)$$

Nous retrouvons, pour $n = 2$, une autre formulation de la reconstruction tomographique par inversion de l'opérateur d'Abel à partir des projections $\mathcal{Y}$. Il suffit de projeter un objet virtuel, à symétrie de révolution, de densité d égale à la projection $\mathcal{Y}$ (soit $d(r\cos\theta, r\sin\theta) = \mathcal{Y}(r)$) et d'évaluer :

$$A_2[f](x) = A_1\left[A_1[f]\right](x) = \frac{1}{\pi} \mathcal{A}[\mathcal{Y}](x)$$

et

$$f(x) = \frac{-A_2[f]'(x)}{2x} \qquad (6)$$

Les figures 2 et 3 illustrent la comparaison entre une reconstruction par inverse généralisé [2] et notre approche pour inverser la transformée d'Abel.

Propriété 3 *Tous les moments $\{M_n[f]\}_{n \in I\!N}$ de f se calculent à partir de la projection $\mathcal{Y}$:*

$$M_{n-1}[f] = \int_{I\!R^+} u^{n-1} f(u) du = \frac{\Gamma(\frac{n}{2})}{2} A_n[f](0) \qquad (7)$$

avec

$$A_n[f] = \frac{1}{\pi^{\frac{n}{2}}} (\mathcal{A})^{n-1} [\mathcal{Y}]$$

Numériquement, dans un cas discrétisé, il est plus aisé d'évaluer les opérateurs tomographiques selon leur parité à l'aide des relations 3 et 5 car ils se calculent de proche en proche par application successive de l'opérateur A_2 dont la forme matricielle, discrétisant l'expression

$$A_2[f](x) = 2. \int_x^{+\infty} u.f(u) du,$$

peut être donnée par :

$$\boldsymbol{A_2.f} = \begin{pmatrix} 1 & 3 & 5 & \dots & 2N-1 \\ 0 & 3 & 5 & \dots & 2N-1 \\ 0 & 0 & 5 & \dots & 2N-1 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 2N-1 \end{pmatrix} . \begin{pmatrix} f(1) \\ f(2) \\ f(3) \\ . \\ f(N) \end{pmatrix}$$

Nous pouvons, par exemple, calculer les moments de Legendre μ_i à partir des M_i (voir [5]) ce qui fournit une décomposition de la solution du type :

$$f = \sum_{i=0}^{\infty} L_i(x)\mu_i(\mathcal{Y})$$

avec L_i, le $i^{\text{ième}}$ polynôme de Legendre [5]. Sur la figure 4, nous illustrons le calcul des 12 premiers moments de Legendre de f (voir encadré de la figure 4) à partir de sa projection et les comparons avec leur valeur théorique. Les premiers ordres sont bien restitués mais, de manière classique, nous obtenons rapidement des divergences numériques pour les ordres élevés.

3 Calcul de la transformation en cosinus et autres noyaux à partir d'une projection

Dans cette section, nous considérons des transformations intégrales de noyau K et nous nous proposons de calculer la transformée d'une fonction f à support compact à partir de sa projection $\mathcal{Y}$.

Pour cela, nous utilisons le résultat suivant dont les éléments de preuve sont donnés en annexe :

Propriété 4 *Pour des noyaux K de classe C^2 sur $\mathbb{R}^{+2}$,*

$$\begin{aligned}\int_{\mathbb{R}^+} K(x,y).f(y)dy &= \int_{\mathbb{R}^+} \Phi(x,y)A_2[f](y)dy \qquad (8)\\ &+\frac{1}{2}\frac{\partial K}{\partial y}(x,0)A_2[f](0)+\frac{1}{2}K(x,0)\mathcal{Y}(0)\end{aligned}$$

avec

$$\Phi(x,y)=\frac{1}{2}\left[\frac{1}{y}\frac{\partial K}{\partial y}(x,y)+\frac{K(x,0)-K(x,y)}{y^2}\right]$$

Les contraintes imposées au noyau K, suffisantes mais non nécessaires, permettent de traiter un bon nombre de noyaux K classiques ; ainsi, en prenant $K(x,y)=\cos(xy)$, nous obtenons une expression de la transformée en cosinus $T_c[f]$ de la fonction f en fonction de sa projection $\mathcal{Y}$:

$$T_c[f](x)=\int_{\mathbb{R}^+}\Phi_c(x,y)\mathcal{A}[\mathcal{Y}](y)dy+\frac{1}{2}\mathcal{Y}(0) \qquad (9)$$

avec

$$\Phi_c(x,y)=\frac{1}{2}\left(-\frac{x\sin(xy)}{y}+\frac{1-\cos(xy)}{y^2}\right)$$

Contrairement au théorème coupe projection qui nous fournit une approximation de la transformée de Fourier 2D de $\tilde{f}$, via l'incontournable interpolation polaire→cartésien, nous obtenons directement l'expression exacte de la transformée de Fourier d'une méridienne de $\tilde{f}$.

Nous illustrons, sur la figure 5, le calcul de $T_c[f]$ de la fonction de densité valant 1 entre 0 et 200 et dont la projection parallèle est donnée sur le graphe Gr.a. Malgré une discrétisation "brutale" de l'équation (9), la transformée estimée est très proche de la valeur exacte.

4 Travaux futurs

Nous étudions actuellement la sensibilité au bruit des différentes reconstructions ainsi que d'autres transformations à noyaux dans le cadre de la propriété 4. Parmi les nombreuses propriétés des opérateurs tomographiques, nous cherchons de plus à établir des relations entre les coefficients d'une fonction densité et de sa projection dans des bases couplées par un opérateur dual A_n^*, comme illustré dans le lemme 1.

Annexe : Eléments de preuve de la propriété 4

Notons Δ l'opérateur "taux d'accroissement" et Q l'opérateur dérivée :

$$\Delta f(x)=\frac{f(x)-f(0)}{x} \qquad ; \qquad Qf(x)=f'(x)$$

Puisque, d'après (6), $f(x)=-\frac{1}{2x}QA_2[f]$, nous obtenons par intégration par parties :

$$\begin{aligned}(g,f) &= \int_{\mathbb{R}^+} g(u).f(u)du\\ &= g(0).M_0[f]+\int_{\mathbb{R}^+}[g(u)-g(0)].f(u)du\\ &= g(0).M_0[f]-\frac{1}{2}\int_{\mathbb{R}^+}\Delta g(u).QA_2[f](u)du\\ &= g(0).M_0[f]+g'(0).M_1[f]\\ &\quad +\frac{1}{2}.\int_{\mathbb{R}^+}(Q\circ\Delta)g(u).A_2[f](u).du\end{aligned}$$

En prenant $g=K(x,.)$ et en utilisant la relation 7, nous établissons la propriété 4.

Notons que l'équation 8 est la version avec reste intégral de la décomposition en série d'une transformation obtenue en réitérant le processus d'intégration par parties :

$$\begin{aligned}T[f](x) &= \frac{1}{2}\cdot\sum_{n\geq 0}\frac{K^{(0,2n)}(x,0)}{2^{2n}n!}\cdot A_{2n}[\mathcal{Y}](0)\\ &+ \frac{1}{2}\cdot\sum_{n\geq 0}\frac{n!.K^{(0,2n+1)}(x,0)}{(2n+1)!}\cdot A_{2n+2}[f](0)\end{aligned}$$

avec $K^{(m,n)}=\frac{\partial^{n+m}}{\partial x^n.\partial y^m}\cdot K$.

Références

[1] R. N. Bracewell. The Fourier transform and its applications. MCGraw-Hill editions.

[2] J. M. Dinten. Tomographie à partir d'un nombre limité de projections : régularisation par des champs markoviens. Thèse de doctorat, Université Paris Sud, centre d'Orsay, 1990.

[3] J. M. Lagrange. Reconstruction tomographique à partir d'un petit nombre de vues. Thèse de doctorat de l'ENS Cachan soutenue le 16 Décembre 1998.

[4] I. et R. Abraham. An active curve approach for tomographic reconstruction of binary radially symmetric objects. Accepté à Inverse Problems.

[5] G. Talenti. Recovering a function from a finite number of moments. Inverse Problems, vol. 3, pp. 501-517, 1987.

[6] F. Lavallou. Réflexions sur le monde conique. Rapport interne CEA 2004.

[7] J. Hadamard. Le problème de Cauchy et les équations aux dérivées partielles linéaires hyperboliques. Paris, Herman, 1932.

[8] M. Riesz. L'intégrale de Riemann-Liouville et le problème de Cauchy pour l'équation des ondes. CRIM de Juillet 1937, Paris, Gauthiers-Villars, 1939.

[9] A. D. Poularikas. The transforms and applications handbook. CRC Press, 1996.

[10] A. Erdelyi et al. Tables of integral transforms. Vol. II, chap. 13, Mc GrawHill, NY 1954.

[11] C.A.M. van Berkel, S.J.L. van Eijndhoven Fractional calculus, gegenbauer transformations, and Integral Equations J. of Math. Analysis and Applications, Vol.194, pp.126-146, 1995.

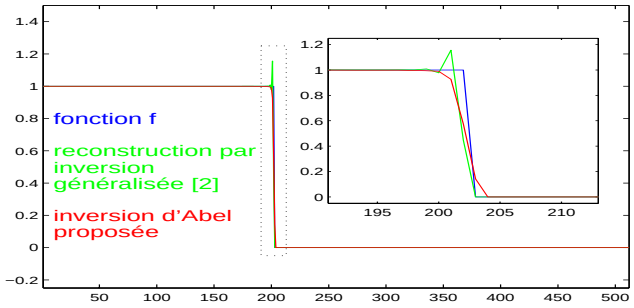


Fig. 2 – *Comparaison de la reconstruction d'une densité f constante par inverse généralisé et par "reprojection". L'atténuation de l'oscillation (voir zoom) est due au lissage inhérent au mode de calcul de la dérivée de $A_2[f]$ (demie somme de la dérivée à droite et à gauche) qui produit un front de montée indépendant du rayon du cylindre.*

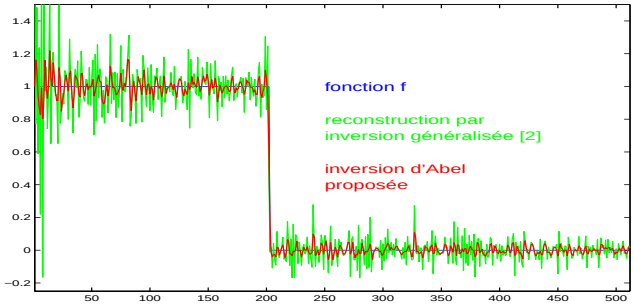


Fig. 3 – *Pour les conditions de calcul de la figure 2, le bruit sur une reconstruction dépend de la méthode numérique utilisée pour l'inversion de la transformée d'Abel. La moins grande sensibilité au bruit de la "reprojection" provient du lissage intégral de $\mathcal{Y}$ dans l'équation 6.*

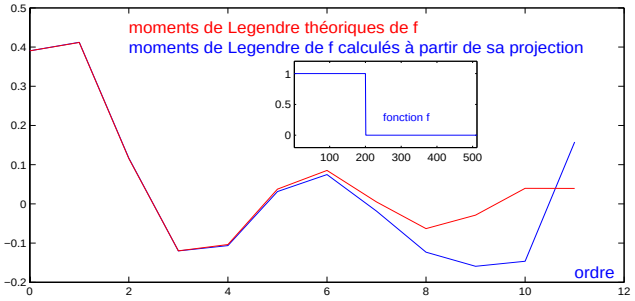


Fig. 4 – *Le calcul des itérés de la projection de f permet de calculer les moments de la fonction densité en appliquant 7. Nous en déduisons la valeur des moments de Legendre de cette fonction densité (voir encadré) qui illustre un corps cylindrique de densité constante. Nous constatons un bon accord jusqu'à l'ordre 6, au-delà le cumul des erreurs numériques rend cette évaluation inopérante.*

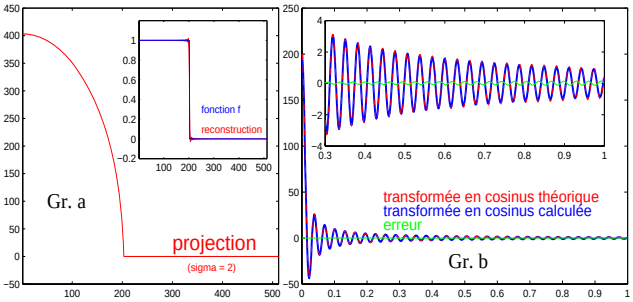


Fig. 5 – *Le graphe Gr.a fournit la projection (en rouge) d'un objet cylindrique de densité constante (la fonction f en bleu). En superposition, est representée la reconstruction obtenue par la transformée en cosinus inverse. Le graphe Gr.b illustre le calcul de la transformée en cosinus de la fonction f à partir de sa projection $\mathcal{Y}$ par discrétisation "brutale" de l'équation (9).*

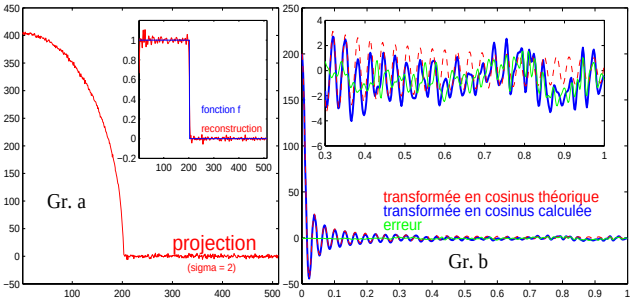


Fig. 6 – *Avec la projection de la figure 5 bruitée ($\sigma = 2$), nous constatons que l'erreur sur l'estimation de $T_c[f]$ croît avec la fréquence, mais est très inférieure à celle que l'on obtiendrait par application du théorème coupe-projection.*

Comparaison entre les algorithmes semi-quadratiques et gradients conjugués préconditionnés pour la restauration d'image préservant les bords

Christian LABAT, Jérôme IDIER

IRCCyN-CNRS UMR 6597
ECN, 1 rue de la Noë, BP 92101, 44321 Nantes Cedex 3, France
prénom.nom@irccyn.ec-nantes.fr

Résumé – Notre contribution concerne la restauration d'image. L'approche adoptée consiste à minimiser un critère des moindres carrés pénalisés. Ici, nous nous intéressons à la recherche d'algorithmes efficaces pour mener une telle tâche. Plus précisément, nous étudions des liens entre algorithmes semi-quadratiques (SQ) et gradients conjugués préconditionnés (GCP) pour la minimisation de critères pénalisés. Nous proposons un théorème de convergence pour l'algorithme GCP associé à une recherche de pas spécifique. Enfin, nous menons une comparaison expérimentale entre algorithmes SQ et GCP dans le cadre de la déconvolution montrant la grande efficacité de ce dernier.

Abstract – Our contribution deals with image restoration. The adopted approach consists in minimizing a criterion of penalized least squares. Here, we are interested in research of effective algorithms to carry out such a task. More precisely, we study links between half-quadratic (HQ) algorithms and preconditioned conjugate gradient (PCG) algorithm for the minimization of penalized criteria. We propose a theorem of convergence for the PCG algorithm associated with a specific line search. Lastly, we carry out an experimental comparison between algorithms HQ and PCG within the framework of deconvolution showing the great effectiveness of this last.

1 Introduction

Notre contribution concerne la restauration d'image où une image solution $\hat{\boldsymbol{x}} \in \mathbb{R}^N$ est estimée à partir d'une observation bruitée $\boldsymbol{y} \in \mathbb{R}^M$. Suivant [1], nous proposons de minimiser un critère des moindres carrés pénalisé avec un terme préservant les bords de l'image:

$$\hat{\boldsymbol{x}} = \min_{\boldsymbol{x}\in\mathbb{R}^N} \mathcal{J}(\boldsymbol{x}) \tag{1}$$

$$\mathcal{J}(\boldsymbol{x}) = \mathcal{Q}(\boldsymbol{x}) + \lambda\Phi(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^N \tag{2}$$

où $\mathcal{Q}(\boldsymbol{x}) = \|\mathbf{A}\boldsymbol{x} - \boldsymbol{y}\|^2$, avec $\mathbf{A} \in \mathbb{R}^{M\times N}$ la matrice d'observation, λ un hyperparamètre de régularisation et

$$\Phi(\boldsymbol{x}) = \sum_{c\in\mathcal{C}} \phi(\boldsymbol{v}_c^{\mathrm{t}}\boldsymbol{x}),$$

avec $\mathcal{C}$ ($|\mathcal{C}| = C$) un ensemble de cliques sur $\mathcal{S}$ ($|\mathcal{S}| = N$), chaque vecteur $\boldsymbol{v}_c \in \mathbb{R}^N$ ayant un support restreint à la clique c. Pour illustrer, $\mathcal{C}$ peut correspondre à l'ensemble des paires de pixels voisins $c = \{r, s\}$, avec $\boldsymbol{v}_c^{\mathrm{t}}\boldsymbol{x} = x_r - x_s$. La fonction potentiel $\phi : \mathbb{R} \mapsto \mathbb{R}$ peut être C^2 convexe et issue du domaine de l'estimation robuste préservant les bords, c'est-à-dire qu'elle a un comportement affine vers l'infini [2]. Le critère $\mathcal{J}$ est alors lui-même convexe, ce qui assure qu'il n'y a pas de multiples minima locaux. Le coût de calcul d'un tel estimateur peut être un facteur limitant l'utilisation de ce type de méthode, c'est pourquoi nous nous focalisons ici sur la recherche d'algorithmes rapides pour minimiser $\mathcal{J}$.

Soit $\mathbf{Q}_0 = 2\mathbf{A}^{\mathrm{t}}\mathbf{A}$ et $\mathbf{V} = [\boldsymbol{v}_1^{\mathrm{t}}|...|\boldsymbol{v}_C^{\mathrm{t}}]$. Dans la suite, on suppose que la matrice $\mathbf{Q}_0 + \lambda\mathbf{V}^{\mathrm{t}}\mathbf{V}$ est inversible.

2 Les algorithmes semi-quadratiques

2.1 Formulation "gradient-reliée"

La minimisation de (1) peut être conduite selon une approche semi-quadratique (SQ), ce qui fournit alors les constructions de Geman & Reynolds (GR) et Geman & Yang (GY) [1]. Dans [3], il est démontré que les algorithmes SQ GR et GY s'identifient à des méthodes *gradient-reliées* à *pas constant unitaire*.

La forme GR se définit par

$$\boldsymbol{d}_{\mathrm{GR}}^k = -(\mathbf{B}_{\mathrm{GR}}(\boldsymbol{x}^k))^{-1}\nabla\mathcal{J}(\boldsymbol{x}^k) \quad \text{(direction GR)} \tag{3}$$

$$\boldsymbol{x}^k = \boldsymbol{x}^k + \boldsymbol{d}_{\mathrm{GR}}^k \quad \text{(mise à jour GR)}$$

où l'opérateur normal $\mathbf{B}_{\mathrm{GR}}(\boldsymbol{u})$ est définie par

$$\boldsymbol{b}(\boldsymbol{u}) := \mathrm{Vect}\left[\frac{\phi'(|[\mathbf{V}\boldsymbol{u}]_c|)}{|[\mathbf{V}\boldsymbol{u}]_c|}\right],$$

$$\mathbf{B}_{\mathrm{GR}}(\boldsymbol{u}) := \mathbf{Q}_0 + \lambda\mathbf{V}^{\mathrm{t}}\,\mathrm{Diag}\{\boldsymbol{b}(\boldsymbol{u})\}\mathbf{V},$$

et la forme GY se définit par

$$\boldsymbol{d}_{\mathrm{GY}}^k = -(\mathbf{B}_{\mathrm{GY}}^a)^{-1}\nabla\mathcal{J}(\boldsymbol{x}^k) \quad \text{(direction GY)} \tag{4}$$

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \boldsymbol{d}_{\mathrm{GY}}^k \quad \text{(mise à jour GY)}$$

où $a = 1/\phi''(0)$ et la matrice normale $\mathbf{B}_{\mathrm{GY}}^a$ est définie par

$$\mathbf{B}_{\mathrm{GY}}^a := \mathbf{Q}_0 + \frac{\lambda}{a}\mathbf{V}^{\mathrm{t}}\mathbf{V}.$$

Des résultats de convergence de ces algorithmes SQ se trouvent notamment dans [1].

2.2 Inversion des matrices normales semi-quadratiques

Le calcul exact de la direction de descente des algorithmes SQ (3)-(4) est généralement trop coûteux. [2, 4, 5] proposent

une version de cette direction de descente approchée par gradients conjugués (GC) ou GC préconditionnés (GCP) tronqués, dans le sens où un nombre de sous-itérations plus faible que la dimension de l'espace solution est utilisé. Ce type d'implémentation a pour avantage d'être efficace, mais au prix de la perte des résultats de convergence associés aux calculs des directions de descentes SQ exactes.

La partie suivante présente en quelque sorte une version duale de cette implémentation: au lieu d'utiliser les GCP au sein des SQ, nous étudions une forme d'algorithme GCP non linéaires où se retrouve une version scalaire des SQ tronqués.

3 L'algorithme des gradients conjugués préconditionnés non linéaire

3.1 La forme préconditionnée de Polak-Ribiere

La méthode des gradients conjugués (GC) est une mise à jour itérative de la forme $\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \alpha_k \boldsymbol{d}^k$, où les directions de descentes $\left\{\boldsymbol{d}^k\right\}$ sont mutuellement conjuguées dans le cas d'un problème purement quadratique. Suivant [6], nous nous limitons ici à la forme de Polak-Ribiere des GC préconditionnée (PRGCP) [7]:

$$\begin{aligned}
\boldsymbol{p}^k &= -(\mathbf{M}^k)^{-1}\nabla\mathcal{J}(\boldsymbol{x}^k) \quad \text{(préconditionnement)} \qquad (5)\\
\beta_k &= \begin{cases} 0 & \text{si } k = 0 \\ \dfrac{\left(\nabla\mathcal{J}(\boldsymbol{x}^k) - \nabla\mathcal{J}(\boldsymbol{x}^{k-1})\right)^{\mathrm{t}} \boldsymbol{p}^k}{\left(\nabla\mathcal{J}(\boldsymbol{x}^{k-1})\right)^{\mathrm{t}} \boldsymbol{p}^{k-1}} & \text{si } k > 0 \end{cases}\\
\boldsymbol{d}^k &= \boldsymbol{p}^k + \beta_k \boldsymbol{d}^{k-1} \quad \text{(direction PRGCP)}\\
\boldsymbol{x}^{k+1} &= \boldsymbol{x}^k + \alpha_k \boldsymbol{d}^k \quad \text{(recherche du pas et mise à jour)}
\end{aligned}$$

L'algorithme PRGCP possède deux degrés de liberté: le pas α_k, et la matrice de préconditionnement définie positive $\mathbf{M}^k$.

3.2 Recherche du pas avec les formes scalaires des algorithmes semi-quadratiques

Contrairement aux algorithmes SQ, les méthodes GC requièrent usuellement une recherche du pas non triviale afin de déterminer un pas α_k admissible. Dans [6], une procédure itérative de recherche du pas est présentée. Elle a pour avantage d'être simple et convergente dans le sens où la suite générée $\alpha_k^1, \ldots, \alpha_k^i$ converge vers un minimum local de $f(\alpha) = \mathcal{J}(\boldsymbol{x}^k + \alpha \boldsymbol{d}^k)$. Cette procédure s'identifie en fait à la version scalaire (1D) de l'algorithme GR. Dans le même esprit, une alternative utilisant l'algorithme GY scalaire peut être envisagée. En considérant la formulation 1D gradient-reliée de ces algorithmes SQ appliqués à la minimisation de $f(\alpha)$, ces algorithmes s'écrivent

$$\begin{cases} \alpha_k^0 = 0 \\ \alpha_k^{i+1} = \alpha_k^i - \frac{(\boldsymbol{d}^k)^{\mathrm{t}} \nabla\mathcal{J}\left(\boldsymbol{x}^k + \alpha_k^i \boldsymbol{d}^k\right)}{(\boldsymbol{d}^k)^{\mathrm{t}} \mathbf{Q}_i^k \boldsymbol{d}^k}, \quad 0 \leqslant i \leqslant I-1 \\ \alpha_k = \alpha_k^I \end{cases} \qquad (6)$$

avec $\mathbf{Q}_i^k = \mathbf{B}_{\mathrm{GY}}^a$ pour GY1D et avec $\mathbf{Q}_i^k = \mathbf{B}_{\mathrm{GR}}(\boldsymbol{x}^k + \alpha_k^i \boldsymbol{d}^k)$ pour GR1D. Lorsque $I = 1$, la formule obtenue est la même que celle proposée dans [8], pour une vaste famille de méthodes des gradients conjugués.

3.3 Théorème de convergence

Le théorème suivant concerne la convergence de l'algorithme PRGCP avec la recherche du pas (6). La preuve se trouve dans [9].

Théorème 1. *Supposons que*

$$\begin{cases} \phi : \mathbb{R} \mapsto \mathbb{R} \textit{ est } \mathcal{C}^2\textit{, coercive et paire} \\ \forall t > 0,\ \phi'(t) > 0 \\ \psi(t) = \phi(\sqrt{t}) \textit{ est concave pour } t > 0 \\ \sup_{t\in\mathbb{R}} |\phi''(t)| < \infty \end{cases}$$

Alors, la forme de Polak-Ribiere préconditionnée par une suite de matrices constantes $\mathbf{M}^k = \mathbf{M}_0$ *est convergente dans le sens où*

$$\liminf_{k\to\infty} \|\nabla\mathcal{J}(\boldsymbol{x}^k)\| = 0$$

lorsque la recherche du pas (6) *est employée avec* $\mathbf{Q}_i^k = \mathbf{B}_{\mathrm{GY}}^a$ *ou bien* $\mathbf{Q}_i^k = \mathbf{B}_{\mathrm{GR}}(\boldsymbol{x}^k + \alpha_k^i \boldsymbol{d}^k)$.

4 Liens entre algorithmes

4.1 Comparaison des pas

Si on choisit $\mathbf{Q}^k = \mathbf{M}^k$ et si on utilise une stratégie sans conjugaison, alors la formule du pas (6) avec $I = 1$ se simplifie en $\alpha_k = 1$. Les algorithmes GY et GR correspondent alors aux choix particuliers des matrices normales SQ pour $\mathbf{M}^k$. Par conséquent, la formule du pas (6) peut être interprétée comme une généralisation naturelle du pas constant lorsqu'on passe des algorithmes gradients-reliés aux algorithmes des gradients conjugués.

4.2 Intérêt du choix de la matrice de préconditionnement

La matrice de préconditionnement des algorithmes semi-quadratiques est imposée ($\mathbf{B}_{\mathrm{GR}}(\boldsymbol{x}^k)$ et $\mathbf{B}_{\mathrm{GY}}^a$), ce qui entraîne en général des difficultés d'inversion. Par contre, l'algorithme PRGCP permet de choisir une matrice de préconditionnement définie positive quelconque et constante (afin que le théorème 1 s'applique). Ainsi, il est possible d'utiliser une matrice de type circulante bloc circulante (CBC) comme préconditionneur lorsque le Hessien de $\mathcal{J}$ est approximativement CBC, sous l'hypothèse classique de régularité de l'image recherchée. On peut alors tirer profit du fait que les matrices CBC se diagonalisent facilement dans la base de Fourier.

5 Comparaison expérimentale de la vitesse de convergence

5.1 Les problèmes traités

Nous considérons comme problème de restauration celui de la déconvolution en présence de bruit. Deux cas sont étudiés: convolution à partir d'un noyau gaussien (écart-type 2.24, taille 17×17) avec l'hypothèse de bords zéros à l'extérieur (Dirichlet), et débruitage qui n'est qu'un cas limite de déconvolution. L'image d'origine, de taille $N = 512 \times 512$ à valeur dans $[0, 255]$, est la même pour ces deux problèmes.

La pénalisation retenue utilise pour $\mathcal{C}$ l'ensemble des paires de pixels voisins et pour potentiel la fonction convexe hyperbolique $\phi(t) = \sqrt{\delta^2 + t^2}$, où δ correspond à un seuil entre le régime quadratique et affine de ϕ [1].

Pour chacun des deux problèmes, le critère $\mathcal{J}$ convexe ainsi défini dépend des hyperparamètres λ et δ: nous avons choisi de les régler de manière à avoir l'image restaurée la plus proche visuellement de l'image d'origine. Ainsi, nous avons retenu $\delta = 13$ pour les deux problèmes, $\lambda = 10$ pour le débruitage et $\lambda = 0.2$ pour la déconvolution.

Finalement, nous avons deux critères à minimiser pour lesquels le même critère d'arrêt global est utilisé

$$\eta = \|\nabla \mathcal{J}(\boldsymbol{x}^k)\|/N < 10^{-6}. \quad (7)$$

(a) Image bruitée: RSB 20 (b) Image convoluée: RSB 40

(c) Image débruitée (d) Image déconvoluée

FIG. 1: *(a)-(b) Images simulées. (c)-(d) Restauration par minimisation de critère pénalisé.*

5.2 Les versions algorithmiques testées

A présent, nous comparons les vitesses de convergence des familles d'algorithmes précédemment décrites, où la notation {P} signifie avec ou sans préconditionnement CBC:

1 Les algorithmes GR et GY approchés avec inversion des matrices normales par une forme tronquée des GC ou GCP linéaires: GR+GC{P} et GY+GC{P}.

2 L'algorithme PRGC ou PRGCP non linéaire avec détermination du pas

  2.1 à l'aide de la recherche scalaire `fminbnd` de Matlab: GC{P}+fminbnd

  2.2 à l'aide des formes scalaires de GR ou GY (6): GC{P}+GR1D et GC{P}+GY1D.

Toutes ces familles d'algorithmes présentent la même structure itérative: des itérations globales au sein desquelles des sous-itérations sont effectuées, que ce soit des GC linéaires pour la première famille d'algorithmes ou bien une recherche du pas pour la deuxième famille d'algorithmes. Le critère d'arrêt (7) porte sur les itérations globales, mais il nous faut aussi définir un critère d'arrêt pour les sous-itérations.

Pour la première famille d'algorithmes, nous reprenons la proposition de [4]: les GC{P} linéaires sont utilisés avec le critère d'arrêt $\|r^i\|/\|r^0\| < 10^{-6}$, où r^i est le résidu de l'équation normale après i itérations.

La fonction `fminbnd` (nombre d'or et interpolation parabolique) de Matlab est utilisée avec son critère d'arrêt par défaut.

Les formes scalaires de GR ou GY (6) sont utilisées avec I itérations, où I est fixé au préalable.

5.3 Résultats

Nous comparons expérimentalement l'efficacité des familles d'algorithmes 1, 2.1 et 2.2 pour la minimisation des critères définis précédemment, ceci selon deux types d'évaluations complémentaires:

- le nombre d'itérations globales et le nombre moyen de sous-itérations qui donnent une idée de la complexité de l'algorithme considéré,
- le temps mis pour effectuer la minimisation qui dépend bien sûr de la plateforme d'implémentation.

Les Tab. 1 et 2 récapitulent, pour un algorithme donné, le premier critère d'évaluation sous la forme (It1, It2), où:

- It1 est le nombre d'itérations globales,
- It2 est le nombre moyen de sous-itérations.

Les expériences ont été menées avec Matlab 6.5 sur un PC P4 2.8GHz RAM 1Go. Les algorithmes ont été rangés par ordre de durée de minimisation croissant.

Nous avons testé la famille d'algorithme 2.2 avec $I \in \{1, 2, 3, 4, 5, 10, 15, 20\}$. Seules les versions $I = 1$ ont été systématiquement reportées dans Tab. 1 et 2, les autres versions n'ayant été reportées que si elles sont plus rapides.

	Sans precond.		Avec precond.	
	Iter.	Tps (s)	Iter.	Tps (s)
GC+GR1D	(11, 2)	9.2	(9, 2)	10.62
”	(12, 1)	7.95	(13, 1)	12.57
GC+GY1D	(11, 4)	12.19	(9, 4)	13.17
”	(14, 1)	9.67	(20, 1)	19.5
GC+fminbnd	(12, 11.8)	20.85	(10, 12.9)	21.35
GR+GC	(10, 21.5)	62.76	(10, 12.3)	82.23
GY+GC	(20, 20.5)	91.83	(20, 7.3)	86.14

TAB. 1: *Comparaison des algorithmes sur le problème de débruitage.*

	Sans precond.		Avec precond.	
	Iter.	Tps (s)	Iter.	Tps (s)
GC+GR1D	(90, 1)	140.1	(43, 1)	79.16
GC+GY1D	(89, 2)	151.3	(45, 2)	89.91
”	(93, 1)	145.7	(53, 1)	98.85
GC+fminbnd	(94, 9.3)	220.1	(47, 10.4)	129.4
GR+GC	(13, 1135.5)	20485	(13, 161.8)	3598
GY+GC	(38, 498.7)	26490	(38, 66.9)	4331

TAB. 2: *Comparaison des algorithmes sur le problème de déconvolution.*

L'analyse des Tab. 1 et 2 fait ressortir en premier lieu que la famille d'algorithmes 2 surpasse largement la famille d'algorithmes 1 tant du point de vue du nombre d'itérations que

du temps de calcul. Si le préconditionnement CBC est très favorable pour le problème de déconvolution, il devient presque systématiquement contre productif pour le problème de débruitage. Ceci est une conséquence du meilleur conditionnement de ce dernier. Cependant, l'utilisation d'un préconditionneur à base de transformée en cosinus pourrait être plus efficace que celui utilisé à base de transformée de Fourier.

Au sein de chaque famille d'algorithmes, les versions GR font systématiquement mieux que les versions GY. Ainsi, par souci de clarté nous avons exclu les versions GY de la Fig. 2.

Enfin, il apparaît clairement qu'un nombre très faible de réitérations des formes scalaires SQ fournit à chaque fois la version la plus rapide parmi la famille d'algorithmes 2.2, le cas non réitéré ($I = 1$) arrivant même souvent en tête.

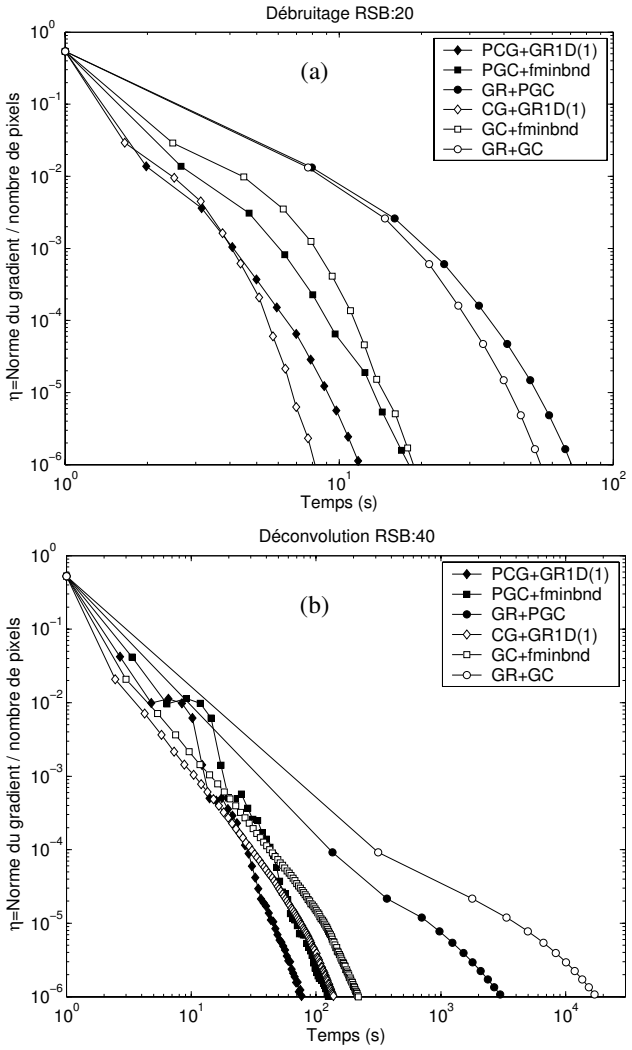


FIG. 2: *Comparaison des taux de convergence:* $\|\nabla \mathcal{J}(\boldsymbol{x}^k)\|/N$ *en fonction du temps. (a) Problème de débruitage. (b) Problème de déconvolution.*

6 Conclusion

La famille d'algorithmes PRGC{P} non linéaires avec un pas déterminé par une forme scalaire SQ présente plusieurs intérêts majeurs par rapport aux algorithmes SQ avec inversion tronquée par GC{P} linéaires:

- la convergence de la première famille est assurée par le Théorème 1, alors que la convergence de la seconde famille est à l'heure actuelle un problème ouvert, et
- la première famille apparaît expérimentalement comme largement plus efficace à tel point qu'il nous semble raisonnable de penser que pour les problèmes étudiés, les ordres de grandeurs des écarts temporels observés seront préservés indépendemment de l'implémentation.

Pour finir, le Théorème 1 assurant la convergence de la famille d'algorithme PRGC{P} non linéaires pour un pas déterminé par une forme scalaire SQ présente aussi des avantages par rapport aux résultats de convergence existants portant sur une recherche du pas qui vérifie des conditions telles que celles de Wolfe:

- d'efficacité, car il permet d'éviter de faire des sous-itérations inutiles puisque la convergence est assurée pour un nombre de sous-itérations I quelconque, donc aussi petit que souhaité; et
- de simplicité, car il dépend du seul paramètre naturel I alors que le réglage des paramètres des conditions de Wolfe est délicat.

[1] J. Idier, « Convex half-quadratic criteria and interacting auxiliary variables for image restoration », *IEEE Trans. Image Processing*, vol. 10, n° 7, pp. 1001–1009, July 2001.

[2] P. Charbonnier, L. Blanc-Féraud, G. Aubert et M. Barlaud, « Deterministic edge-preserving regularization in computed imaging », *IEEE Trans. Image Processing*, vol. 6, n° 2, pp. 298–311, Feb. 1997.

[3] M. Allain, J. Idier et Y. Goussard, « On global and local convergence of half-quadratic algorithms », to appear, *IEEE Trans. Image Processing*, 2005.

[4] M. Nikolova et M. Ng, « Fast image reconstruction algorithms combining half-quadratic regularization and preconditioning », in *Proc. IEEE ICIP*, Thessaloniki, Greece, Oct. 2001, pp. 277–280.

[5] M. Nikolova et M. Ng, « Analysis of half-quadratic minimization methods for signal and image recovery », to appear, *SIAM Journal on Scientific computing*, 2005.

[6] J. A. Fessler et S. D. Booth, « Conjugate-gradient preconditionning methods for shift-variant PET image reconstruction », *IEEE Trans. Image Processing*, vol. 8, n° 5, pp. 668–699, May 1999.

[7] W. H. Press, S. A. Teukolsky, W. T. Vetterling et B. P. Flannery, *Numerical recipes in C, the art of scientific computing*, Cambridge Univ. Press, New York, 2nd edition, 1992.

[8] J. Sun et J. Zhang, « Global convergence of conjugate gradient methods without line search », *Ann. Operations Research*, vol. 103, pp. 161–173, 2001.

[9] C. Labat et J. Idier, « Extended convergence properties of preconditioned conjugate gradient methods without line search and application to image restoration », Rapport technique, IRCCyN, 2005.

ÉCLATEMENT DES CONTRAINTES EN RECONSTRUCTION TOMOGRAPHIQUE

Thomas D. CAPRICELLI, Patrick L. COMBETTES

Laboratoire Jacques-Louis Lions
Université Pierre et Marie Curie – Paris 6
75005 Paris, France
capricelli@ann.jussieu.fr, plc@math.jussieu.fr

Résumé – Le problème de la reconstruction tomographique est formulé en tant que problème de programmation convexe. Une fonction-objectif convexe est minimisée sur l'intersection de convexes construits à partir des connaissances a priori et des observations. Les ensembles convexes sont associés à des contraintes sur l'image originale et sur son sinogramme. L'algorithme proposé pour résoudre ce problème procède par éclatement des contraintes et permet de les traiter par blocs en parallèle. Des résultats numériques sont fournis.

Abstract – The tomographic reconstruction problem is formulated as a convex programming problem. An objective function is minimized over the intersection of convex sets arising from a priori knowledge and the observations. The convex sets are associated with constraints on the original image and its sinogram. The basic principle of the algorithm we propose to solve this problem is to decompose the constraints and to process them by blocks in parallel. Numerical results are provided.

1 Introduction

Le problème de la tomographie bidimensionnelle est de reconstruire l'image $\overline{x}$ d'un objet à partir de mesures de ses intégrales rectilignes suivant des directions $(\theta_i)_{1\leq i\leq q}$. Mathématiquement, il s'agit donc d'estimer une fonction $\overline{x}$ à partir d'une observation partielle et bruitée de sa transformée de Radon [3, 11, 13].

En tomographie assistée par ordinateur, l'image originale $\overline{x}$ est discrétisée et représentée par un point dans l'espace euclidien $\mathbb{R}^N$, N étant le nombre total de pixels. Désignons par s_i ($1 \leq i \leq q$) la i^{e} observation, i.e., la mesure des intégrales rectilignes discrètes au travers de $\overline{x}$ dans la direction θ_i. Chaque mesure s_i est discrétisée sur M points et est donc représentée par un vecteur de $\mathbb{R}^M$. Les vecteurs $(s_i)_{1\leq i\leq q}$ sont liés à $\overline{x}$ par les relations

$$(\forall i \in \{1,\ldots,q\}) \;\; s_i = L_i\overline{x}, \tag{1}$$

où L_i est la restriction de la transformée de Radon à la direction θ_i ; cette transformation linéaire s'exprime en fonction de θ_i [3, 11]. Dans la pratique, le modèle idéal (1) doit être remplacé par

$$(\forall i \in \{1,\ldots,q\}) \;\; s_i = L_i\overline{x} + w_i, \tag{2}$$

où w_i représente un bruit modélisant divers phénomènes physiques. Il est notoire que le problème de la reconstruction de $\overline{x}$ à partir de l'observation partielle $y = (s_i)_{1\leq i\leq q}$ de son sinogramme est mal posé et que, pour obtenir des solutions fiables, il faut incorporer dans la méthode de reconstruction autant d'informations a priori que possible. En prenant comme cadre l'optimisation convexe, le problème peut se mettre sous la forme

$$\text{Trouver } \; x \in S = \bigcap_{i=1}^{m} S_i \quad \text{tel que } \; J(x) = \inf J(S), \tag{3}$$

où la fonction-objectif $J\colon \mathbb{R}^N \to \left]-\infty,+\infty\right]$ est une fonction convexe et les ensembles-contrainte $(S_i)_{1\leq i\leq m}$ sont des convexes fermés de $\mathbb{R}^N$. Ces contraintes proviennent des connaissances a priori et des mesures. Sans perte de généralité, nous pouvons supposer que les ensembles $(S_i)_{1\leq i\leq m}$ sont de la forme

$$(\forall i \in \{1,\ldots,m\}) \;\; S_i = \big\{x \in \mathbb{R}^N \mid f_i(x) \leq \delta_i\big\}, \tag{4}$$

où $(f_i)_{1\leq i\leq m}$ est une famille de fonctions convexes de $\mathbb{R}^N$ dans $\mathbb{R}$ et $(\delta_i)_{1\leq i\leq m} \in \mathbb{R}^m$ des paramètres réels. On supposera que le problème est bien formulé au sens où $S \neq \varnothing$.

Ce type d'approche remonte aux méthodes de reconstruction algébrique (ART) des années 1970 [11]. Dans ce cas, J était constante et les S_i étaient des intersections d'hyperplans associés aux relations idéales (1). Des fonctions-objectif quadratiques et entropiques ont également été proposées, sous contraintes affines associées aux mesures [3]. Le but de cette communication est de proposer une méthode générale pour résoudre (3)–(4) qui procède par éclatement et approximation des contraintes, et qui offre une grande souplesse quant à sa mise en œuvre sur des architectures à processeurs parallèles. Elle nous permet notamment d'introduire de nouveaux types de contraintes sur l'image originale et son sinogramme.

2 Algorithme

2.1 Projection sous-différentielle

Nous rappelons ici brièvement le concept de projection sous-différentielle (cf. [5] pour de plus amples détails).

Sous nos hypothèses, pour tout $x \in \mathbb{R}^N$, chaque fonction f_i admet au moins un sous-gradient en x, i.e., un point $g_i \in \mathbb{R}^N$ tel que

$$(\forall y \in \mathbb{R}^N) \;\; \langle y - x \mid g_i \rangle + f_i(x) \leq f_i(y). \tag{5}$$

L'ensemble des sous-gradients de f_i en x est le sous-différentiel de f_i en x et se note $\partial f_i(x)$; si f_i est différen-

tiable en x, alors $\partial f_i(x) = \{\nabla f_i(x)\}$. La projection sous-différentielle $G_i x$ de x sur S_i est définie en prenant un sous-gradient quelconque $g_i \in \partial f_i(x)$ et en posant

$$G_i x = \begin{cases} x + \dfrac{\delta_i - f_i(x)}{\|g_i\|^2} g_i, & \text{si } f_i(x) > \delta_i; \\ x, & \text{si } f_i(x) \leq \delta_i. \end{cases} \quad (6)$$

Le calcul de $G_i x$ est beaucoup plus économique que celui de la projection exacte $P_i x$, qui revient à résoudre un problème de minimisation sous contrainte ; on rappelle cependant que si la projection $P_i x$ de x sur S_i est facile à calculer, alors on peut prendre $G_i x = P_i x$ [5]. En général, $G_i x$ est la projection sur une approximation affine de S_i.

2.2 Méthode parallèle itérative par blocs

Pour ne pas surcharger l'exposition, nous décrivons l'algorithme dans le cas correspondant à nos simulations numériques, c'est-à-dire quand

$$J\colon x \mapsto \|x - r\|^2. \quad (7)$$

Il s'agit donc de trouver la meilleure approximation, dans l'ensemble des solutions admissibles S, d'une image de référence r.

Notre algorithme est tiré de [6, Section 6.5] ; on se reportera à [6, Section 6.4] pour le cas général et les détails.

Algorithme 1

1. Fixer $\varepsilon \in]0, 1/m[$. Poser $x_0 = r$ et $n = 0$.
2. Prendre un ensemble d'indices $\varnothing \neq I_n \subset \{1, \ldots, m\}$.
3. Poser $z_n = x_n + \lambda_n(\sum_{i \in I_n} \omega_{i,n} p_{i,n} - x_n)$, où :
   a. $(\forall i \in I_n)\ p_{i,n} = G_i x_n$;
   b. $\{\omega_{i,n}\}_{i \in I_n} \subset [\varepsilon, 1]$ et $\sum_{i \in I_n} \omega_{i,n} = 1$;
   c. $\lambda_n = \begin{cases} \dfrac{\sum_{i \in I_n} \omega_{i,n} \|p_{i,n} - x_n\|^2}{\left\| \sum_{i \in I_n} \omega_{i,n} p_{i,n} - x_n \right\|^2}, & \text{si } x_n \notin \bigcap_{i \in I_n} S_i; \\ 1, & \text{sinon.} \end{cases}$
4. Poser $\pi_n = \langle x_0 - x_n \mid x_n - z_n \rangle$, $\mu_n = \|x_0 - x_n\|^2$, $\nu_n = \|x_n - z_n\|^2$, et $\rho_n = \mu_n \nu_n - \pi_n^2$.
5. Poser
$$x_{n+1} = \begin{cases} z_n, & \text{si } \rho_n = 0 \text{ et } \pi_n \geq 0; \\ x_0 + \left(1 + \dfrac{\pi_n}{\nu_n}\right)(z_n - x_n), & \text{si } \rho_n > 0 \text{ et } \pi_n \nu_n \geq \rho_n; \\ x_n + \dfrac{\nu_n}{\rho_n}\big(\pi_n(x_0 - x_n) + \mu_n(z_n - x_n)\big), & \text{si } \rho_n > 0 \text{ et } \pi_n \nu_n < \rho_n. \end{cases}$$
6. Poser $n = n + 1$ et aller à l'étape 2.

Proposition 2 (cf. [6, Thm. 6.4(i)]) *Supposons que*

$$(\exists\, M \in \mathbb{N}^*)(\forall n \in \mathbb{N}) \quad \bigcup_{k=n}^{n+M-1} I_k = \{1, \ldots, m\}. \quad (8)$$

Alors toute suite $(x_n)_{n \in \mathbb{N}}$ générée par l'Algorithme 1 converge vers l'unique solution de (3) *sous les hypothèses* (4) *et* (7)*.*

Remarque 3

- Dans l'algorithme 1, l'ensemble des solutions admissibles S est «éclaté» et l'algorithme n'agit que sur les ensembles-contrainte $(S_i)_{1 \leq i \leq m}$ individuellement.
- Le noyau d'une itération consiste en le calcul simultané d'un bloc de projections sous-différentielles $(p_{i,n})_{i \in I_n}$ de x_n, qui sont ensuite recombinées pour former le point z_n (étape 3). On calcule alors le nouvel itéré x_{n+1} en fonction de x_0, x_n, et z_n (étape 5).
- La suite $(I_n)_{n \in \mathbb{N}}$ définit les blocs d'ensembles balayés à chaque itération. La règle (8) impose que chaque ensemble S_i soit utilisé au mois une fois au cours de M itérations consécutives.
- L'Algorithme 1 peut efficacement être mis en œuvre en adaptant, par le choix des blocs d'ensembles, la charge de calcul de chaque itération à la puissance des processeurs parallèles disponibles.
- Contrairement aux méthodes de projections (exactes ou sous-différentielles) classiques [1, 5, 15], qui produisent une image admissible *quelconque*, l'Algorithme 1 converge vers la solution admissible la plus proche de l'image de référence r. Plus généralement, la méthode de [6, Section 6.4] converge vers la solution admissible qui minimise un critère strictement convexe J.

3 Contraintes

3.1 Contraintes sur le sinogramme

Nous disposons de q observations bruitées du sinogramme de $\overline{x}$ données par (2). Une contrainte naturelle sur une estimée x de $\overline{x}$ est d'imposer que, pour chaque $i \in \{1, \ldots, q\}$, le résidu $s_i - L_i x$ se comporte comme le bruit w_i.

Historiquement, ce type de contraintes remonte à [10, 11], où il est imposé que l'amplitude de chaque point de $s_i - L_i x$ soit statistiquement compatible avec la distribution du bruit w_i. Ceci conduit à des ensembles convexes de la forme

$$\big\{x \in \mathbb{R}^N \mid \ |\langle s_i - L_i x \mid e_k \rangle| \leq \delta_{i,k}\big\}, \quad (9)$$

où $(e_k)_{1 \leq k \leq M}$ est la base canonique de $\mathbb{R}^M$ et $\delta_{i,k}$ est une borne sur les réalisations de la variable aléatoire $|\langle w_i \mid e_k \rangle|$. Notons qu'en pratique il est très délicat d'ajuster les paramètres $\delta_{i,k}$ sur ces qM ensembles de manière à aboutir à un niveau de confiance global satisfaisant [7] (ils doivent être suffisamment grands pour donner une région de confiance acceptable pour $\overline{x}$, tout en étant suffisamment petits pour que S soit sélectif). Une autre approche est celle de [12], qui consiste à considérer le bruit comme global sur l'ensemble du sinogramme. C'est une approche certes grossière et assez éloignée de la réalité physique, mais qui présente l'avantage de s'affranchir d'une modélisation fine du bruit sur les détecteurs.

La question de la modélisation du bruit en tomographie est difficile. Néanmoins, on sait que le bruit trouve son origine majoritairement au niveau des détecteurs. Il est donc raisonnable de supposer que les bruits sur deux vues différentes s_i et s_j sont indépendants. Suivant l'analyse de [9, 14], si Ψ est une propriété connue du bruit w_i dans (2), l'ensemble associé est

$$S_i = \big\{x \in \mathbb{R}^N \mid s_i - L_i x \ \text{ est compatible avec } \ \Psi\big\}. \quad (10)$$

Par exemple, l'expérience montre que, pour certaines classes de problèmes, on peut estimer avec une bonne précision l'énergie

du bruit sur le sinogramme à un angle donné θ_i, ce qui conduit à l'ensemble

$$S_i = \big\{x \in \mathbb{R}^N \mid ||s_i - L_i x||^2 \leq \delta_i\big\}. \quad (11)$$

3.2 Contraintes sur l'image originale

Diverses contraintes sur l'image originale ont été utilisées en tomographie pour traduire des propriétés connues a priori [3, 4, 13]. Par exemple, les amplitudes minimale et maximale des pixels sont souvent connues. Après normalisation, cette contrainte est représentée par l'ensemble

$$S_1 = [0,1]^N. \quad (12)$$

De même, on connaît souvent une approximation K du support de l'objet à reconstruire. L'ensemble associé est

$$S_2 = \big\{x \in \mathbb{R}^N \mid x \cdot 1_K = x\big\}, \quad (13)$$

où 1_K désigne la fonction caractéristique de K.

Il est également possible d'imposer une contrainte de moment. En effet, la moyenne de l'image $\overline{x}$ est invariante par la transformée de Radon à un angle donné. Donc, au vu de (2),

$$(\forall i \in \{1,\dots,q\}) \quad \int_{\mathbb{R}^2} \overline{x} = \int_{\mathbb{R}} L_i\overline{x} = \int_{\mathbb{R}} s_i - \int_{\mathbb{R}} w_i. \quad (14)$$

Donc en supposant les moyennes des bruits $(w_i)_{1\leq i\leq q}$ connues et en prenant les valeurs minimale et maximale des moyennes observées sur les q vues $(s_i)_{1\leq i\leq q}$, on obtient un encadrement de la moyenne de $\overline{x}$. L'ensemble associé à cette contrainte est de la forme

$$S_3 = \big\{x \in \mathbb{R}^N \mid \mu \leq \langle x \mid \underline{1}\rangle \leq \nu\big\}. \quad (15)$$

Notons que les projections sur les trois ensembles ci-dessus peuvent être calculées explicitement [4].

Comme on l'a observé dans [8], dans de nombreux types de problèmes, la variation totale $\mathrm{tv}(\overline{x})$ de l'image originale, qui est une mesure de la quantité d'oscillations, ne dépasse pas une certaine borne connue ζ (cf. [2] pour une autre utilisation de la variation totale en tomographie). Par exemple, en tomographie médicale, les images ont souvent un nombre fini de niveaux de gris et la variation totale s'identifie à la somme des longueur des contours des objets élémentaires composant l'image. Il s'agit donc d'une quantité qui peut être estimée à partir d'expérimentations ou de banques d'images du même type. L'ensemble associé est

$$S_4 = \big\{x \in \mathbb{R}^N \mid \mathrm{tv}(x) \leq \zeta\big\}, \quad (16)$$

où la variation totale d'une image $x = [x^{i,j}] \in \mathbb{R}^{\sqrt{N}\times\sqrt{N}}$ est

$$\begin{aligned}\mathrm{tv}(x) = &\sum_{i=0}^{\sqrt{N}-2}\sum_{j=0}^{\sqrt{N}-2}\sqrt{|x^{i+1,j}-x^{i,j}|^2+|x^{i,j+1}-x^{i,j}|^2}\\ &+\sum_{i=0}^{\sqrt{N}-2}|x^{i+1,\sqrt{N}-1}-x^{i,\sqrt{N}-1}|\\ &+\sum_{j=0}^{\sqrt{N}-2}|x^{\sqrt{N}-1,j+1}-x^{\sqrt{N}-1,j}|. \quad (17)\end{aligned}$$

L'opérateur de projection sous-différentielle G_4 sur cet ensemble est donné dans [8].

FIG. 1 – Image originale.

4 Simulations numériques

Nos simulations sont basées sur l'image test classique de Shepp-Logan (cf. Fig. 1).

4.1 Simulation du bruit

Le bruit provient essentiellement du fait que des quanta ont été déviés de leur trajectoire initiale. Pour simuler le bruit w_i dans (2), nous ajoutons un bruit gaussien centré w à l'image originale $\overline{x}$. La vue correspondant à l'angle θ_i est donc

$$L_i(\overline{x}+w) = L_i\overline{x} + L_i w. \quad (18)$$

Puisque qu'un détecteur ne peut détecter qu'un nombre positif de quanta, nous tronquons les valeurs négatives et donc $w_i = \max\{L_i w, 0\}$. Le rapport signal-sur-bruit sera toujours défini par rapport au bruit «artificiel» w et exprimé en décibels (dB) ; sa valeur est donc $10\log_{10}\big(\|\overline{x}\|^2/\|w\|^2\big)$.

4.2 Mise en oeuvre

Nous utilisons l'Algorithme 1 avec $r = 0$ dans (7). Le problème (3) est donc de trouver l'image admissible d'énergie minimale. Nous utilisons $q = 17$ vues, ce qui est très peu comparé au nombre de vues habituellement utilisées. Le rapport signal-sur-bruit est de 20 dB. La projection sur $S_1 \cap S_2$ de la reconstruction classique par rétroprojection apparaît dans la Fig. 2.

4.3 Contraintes

Nous utilisons les ensembles (12), (13), (15), et (16). Pour la prise en compte des données, nous utilisons les ensembles (11), à savoir,

$$S_{i+4} = \big\{x \in \mathbb{R}^N \mid ||s_i - L_i x||^2 \leq \delta_i\big\}, \quad 1 \leq i \leq q. \quad (19)$$

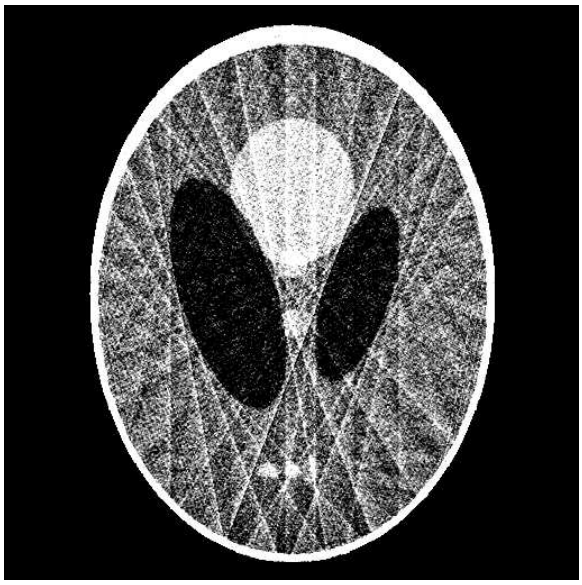


FIG. 2 – Reconstruction par rétroprojection classique sous les contraintes (12) et (13).

FIG. 3 – Reconstruction par la méthode proposée.

On tire de (6) que la projection sous-différentielle d'une image x sur cet ensemble est

$$G_{i+4}x = \begin{cases} x + \dfrac{\delta_i - \|L_i x - s_i\|^2}{2\|L_i^\top (L_i x - s_i)\|^2} L_i^\top (L_i x - s_i), \\ \qquad\qquad \text{si } \|L_i x - s_i\|^2 > \delta_i; \\ x, \qquad\qquad \text{si } \|L_i x - s_i\|^2 \leq \delta_i. \end{cases} \quad (20)$$

On a donc un total de $m = q + 4 = 21$ contraintes. L'Algorithme 1 est porté sur une machine à 8 processeurs parallèles. On choisit donc des blocs de 8 ensembles à chaque itération, activés de manière cyclique, c'est-à-dire

$$I_n = \big\{(8n) \bmod(m) + 1, \ldots, (8n+7) \bmod(m) + 1\big\}. \quad (21)$$

Les poids $(w_{i,n})_{i \in I_n}$ sont pris égaux sur les contraintes actives de $(S_i)_{i \in I_n}$ et nuls sur les contraintes inactives.

4.4 Résultats

La Fig. 3 montre le résultat obtenu en appliquant l'Algorithme 1. On remarque que les ellipses de l'image originale sont toutes présentes et qu'il y a peu d'artefacts.

Références

[1] H. H. Bauschke and J. M. Borwein, "On projection algorithms for solving convex feasibility problems," *SIAM Rev.*, vol. 38, pp. 367–426, 1996.

[2] M. M. Bronstein, A. M. Bronstein, M. Zibulevsky, and H. Azhari, "Reconstruction in diffraction ultrasound tomography using nonuniform FFT," *IEEE Trans. Medical Imaging*, vol. 21, pp. 1395–1401, 2002.

[3] Y. Censor and S. A. Zenios, *Parallel Optimization : Theory, Algorithms and Applications.* New York : Oxford University Press, 1997.

[4] P. L. Combettes, "The convex feasibility problem in image recovery," *Advances in Imaging and Electron Physics*, vol. 95, pp. 155–270. New York : Academic, 1996.

[5] P. L. Combettes, "Convex set theoretic image recovery by extrapolated iterations of parallel subgradient projections," *IEEE Trans. Image Process.*, vol. 6, pp. 493–506, 1997.

[6] P. L. Combettes, "Strong convergence of block-iterative outer approximation methods for convex optimization," *SIAM J. Control Optim.*, vol. 38, pp. 538–565, 2000.

[7] P. L. Combettes and T. J. Chaussalet, "Combining statistical information in set theoretic estimation," *IEEE Signal Process. Lett.*, vol. 3, pp. 61–62, 1996.

[8] P. L. Combettes and J. C. Pesquet, "Image restoration subject to a total variation constraint," *IEEE Trans. Image Process.*, vol. 13, pp. 1213–1222, 2004.

[9] P. L. Combettes and H. J. Trussell, "The use of noise properties in set theoretic estimation," *IEEE Trans. Signal Process.*, vol. 39, pp. 1630–1641, 1991.

[10] G. T. Herman, "A relaxation method for reconstructing objects from noisy X-rays," *Math. Programming*, vol. 8, pp. 1–19, 1975.

[11] G. T. Herman, *Image Reconstruction from Projections, the Fundamentals of Computerized Tomography.* New York : Academic, 1980.

[12] H. Kudo and T. Saito, "Sinogram recovery with the method of convex projections for limited-data reconstruction in computed tomography," *J. Opt. Soc. Amer. A*, vol. 8, pp. 1148–1160, 1991.

[13] H. Stark (Ed.), *Image Recovery : Theory and Application.* San Diego, CA : Academic, 1987.

[14] H. J. Trussell and M. R. Civanlar, "The feasible solution in signal restoration," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, pp. 201–212, 1984.

[15] D. C. Youla and H. Webb, "Image restoration by the method of convex projections : Part 1 – theory," *IEEE Trans. Medical Imaging*, vol. 1, pp. 81–94, 1982.

Restauration myope d'images pour l'inspection des ouvrages d'art par un système de vision embarqué sur drône *

Benoit Vozel[1], Kacem Chehdi[1], Jean Dumoulin[2]

[1]IETR, UMR CNRS 6164, Equipe TSI2M, Université de RENNES I,
BP 80518 - 22305 Lannion Cedex, France

[2]Laboratoire Central des Ponts et Chaussées
Route de Bouaye - BP 41291 - 44341 Bouguenais Cedex, France
vozel@enssat.fr, chehdi@enssat.fr, Jean.Dumoulin@lcpc.fr

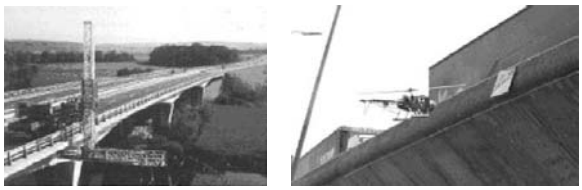
Résumé – Cet article présente les premiers résultats d'une étude de faisabilité du développement d'un système aérien de vision pour l'inspection des structures de génie civil telles que des ponts, tunnels. L'objectif final est d'effectuer l'inspection à l'aide d'un dispositif de vision embarqué sur un drône évoluant à l'extérieur, ou même à l'intérieur des ouvrages d'art. Les images acquises par un tel système de vision embarqué sont inévitablement dégradées. Les dégradations sont dues au comportement dynamique du drône ou aux conditions opérationnelles pendant l'inspection de la structure. Ils se traduisent essentiellement par un flou de mouvement qui peut être combiné avec un flou de défocalisation sur les images acquises. Dans la pratique, ils sont tous les deux inconnus. Nous simulons expérimentalement les dégradations qui peuvent affecter la qualité des images acquises par un tel système de vision embarqué sur drône. Nous considérons ensuite leur restauration numérique. La restauration est réalisée au moyen d'un système de traitement myope. Elle inclut deux modules de traitement, un pour identifier la PSF et un autre pour la restauration de l'image. L'identification et l'évaluation de la PSF est réalisée en utilisant le critère de la validation croisée généralisée et une contrainte de douceur sur l'image. L'étape de restauration utilise une régularisation de type Tikhonov-Miller. Les premiers résultats prouvent que ce système donne des résultats plutôt bons sur les diverses configurations (statique, dynamique) d'acquisition des images. Il est tout à fait efficace sur les images flouées et bruitées, sans connaissance *a priori* des caractéristiques de bruit ou de l'image.

Abstract – This paper presents the first results of a feasibility study focusing on the development of an aerial vision system for the inspection of civil engineering structures such as bridges, tunnels. The final goal is to carry out the inspection by using an onboard vision device on an UAV (Unmanned Aerial Vehicle) evolving outside, or even inside the structures. The images acquired by such an onboard vision system are unavoidably degraded. The degradations are either due to the dynamic behavior of the UAV or the operative conditions during the inspection of the structure. They essentially translate in a motion blur which may be combined with a defocusing blur on the acquired images. In practice, they are both unknown. We first experimentally simulate the degradations that can affect the quality of images acquired by such a vision system onboard an UAV and then consider their numerical restoration. The restoration is achieved by means of a myopic processing system. It includes two processing modules, one for identifying the blur PSF and another for image restoration. The blur PSF identification and estimation is achieved using the Generalized Cross-Validation Criterion and a smoothness constraint on the image. The deblurring step is based on a Tikhonov-Miller regularization technique. First results show that this system yields rather good results on various configuration (static, dynamic) of images acquisition. It is quite effective in restoring blurred and noise corrupted images, without prior knowledge of either the noise or image characteristics. It can also discriminate between the blurred (defocusing and/or motion blurs) and unblurred images.

1 Introduction

La surveillance des ouvrages d'art est une opération nécessaire pour le maintien en état du patrimoine et pour satisfaire à la sécurité des usagers. Actuellement, celle-ci est principalement effectuée visuellement par des inspecteurs utilisant des moyens tels que des passerelles (Fig. 1), des nacelles pour accéder en tout point de l'ouvrage. Cette surveillance porte aussi bien sur le gros oeuvre, que sur la chaussée, les corniches, les gardes corps et les câbles dans le cas des ponts suspendus. Cette inspection détaillée à lieu tous les six ans, avec éventuellement un suivi plus détaillé pour les ouvrages souffrant de désordres préoccupants. Cette méthode d'auscultation des ouvrages d'art entraîne quelques difficultés. En effet pour examiner certaines parties d'ouvrages, il faut pouvoir y accéder sans mettre en danger les intervenants. De plus la mise en place du système pour la surveillance de l'ouvrage impose l'installation de barrages sur les routes et entraîne donc un coût trop élevé. Il est donc envisagé d'automatiser certaines tâches et de réaliser ces contrôles à l'aide de dispositifs de vision embarqués sur un drône (appareil volant sans pilote) évoluant à l'extérieur (Fig. 2) voire à l'intérieur de l'ouvrage d'art [6],[8]. L'efficacité opérationnelle des contrôles réalisés par un drône devrait se traduire par une réduction des risques d'accidents en évitant une in-

*Les auteurs remercient Christian Gourraud[2] pour sa contribution technique dans la mise en place des essais en laboratoire



Une passerelle automotrice Un drône en position d'acquisition

FIG. 1 – Inspection des ouvrages d'art

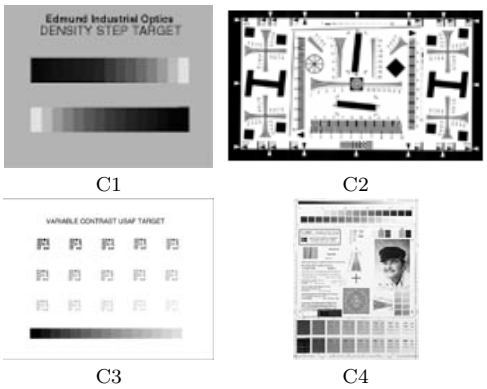
tervention humaine dans une zone à risque. Elle devrait aussi se traduire par une diminution du coût global des mesures en réduisant le nombre d'intervenant et en évitant l'utilisation d'une passerelle automotrice qui nécessite une fermeture partielle ou totale de l'ouvrage à la circulation. L'usage d'une plate-forme d'auscultation de type drône devrait autoriser une meilleure investigation, car il permettra l'inspection des zones impossibles d'accès. A terme, le suivi à long terme de l'ouvrage inspecté devrait en être faciliter et la compréhension de son comportement améliorée. Lors des inspections sur ouvrages d'art, les inspecteurs s'intéressent notamment à des défauts de type fissure dont la largeur est inférieure au millimètre, c'est pourquoi il faudra être en mesure d'obtenir à partir du système de vision embarqué des images dont la qualité soit suffisante pour les détecter. Dans ce contexte, un travail de recherche portant sur la simulation expérimentale et numérique des dégradations pouvant affecter la qualité des images acquises par un système de vision embarqué sur un drone a été engagé. Ces perturbations peuvent être soit liées au comportement dynamique du drone et/ou aux conditions opératoires lors de l'auscultation de l'ouvrage d'art. Ces perturbations sur le dispositif de prise de vue embarqué se traduisent par l'apparition de flous de mouvement combinés ou pas à des flous de défocalisation sur les images acquises. Des expérimentations ont été développées et conduites en laboratoire mais également sur un plan numérique afin de générer deux types de flous sur les images acquises par un dispositif de prises de vues numériques. Ces essais ont fait et continu de faire l'objet d'amélioration en fonction de l'avancement des connaissances acquises au cours des travaux conduits sur la restauration de ces images floues.

2 Les essais en laboratoire

Lors des essais de mise au point d'un protocole expérimental pour l'évaluation d'algorithmes de restauration des images, deux types de flous ont été réalisés lors de l'acquisition. Les images ont été acquises avec une résolution spatiale d'environ 0.3 mm par pixel et une taille d'image 1280 par 1024 pixels (pleine résolution du capteur CMOS [7]).

Les flous de défocalisation ont été réalisés en statique en utilisant différentes bagues allonges (de 0.5, 1, 1.5 et 5 mm).

Dans un premier temps, les prises d'images en configuration dynamique ont été seulement affectées des déplacements générés par le pot vibrant [9] afin de provoquer des flous de mouvement sans déformation du champ de prise de vue (axe optique de la caméra est perpendiculaire à la cible). Dans un second temps, il a été associé à ces flous de mouvement, un flou de défocalisation en ajoutant une bague allonge.



C1 C2

C3 C4

FIG. 2 – Les différentes cibles

3 La restauration numérique myope

Le but de la restauration myope d'image est d'estimer simultanément la PSF et l'image originale avec une connaissance partielle de leurs caractéristiques ou des caractéristiques du bruit. Nous considérons ici uniquement des PSF linéaires, invariantes spatialement, avec un bruit d'observation additif gaussien blanc indépendant de l'image originale. Bien que ceci puisse sembler tout à fait restrictif, les PSF des flous de défocalisation et de mouvement uniforme peuvent être facilement décrites par une fonction paramétrique avec peu de paramètres. Pour résoudre le problème inverse dans un cadre déterministe, on utilise le fait que le contenu des images est généralement basse-fréquence. L'imposition de cette contrainte de régularisation permet de réduire l'espace de recherche pour des solutions [3],[4].

3.1 Le modèle d'observation

Le modèle d'observation retenu est le suivant

$$g = H\ f + n \tag{1}$$

où $f = vec(F)$, $n = vec(N)$, sont respectivement l'image originale et le bruit, g l'image observée, tous les trois de dimension $N^2 \times 1$ et H est la matrice associée à la transformation linéaire construite à partir d'une version discrète de la PSF, de dimension $N^2 \times N^2$. Compte tenu du contexte expérimental décrit précédemment, nous utiliserons les deux modèles H_1 (défocalisation), H_2 (mouvement uniforme) et H_0 (pas de flou), rassemblés ci-dessous, pour la phase de restauration numérique.

3.2 Identification de la PSF

Nous proposons l'utilisation de l'algorithme suivant pour l'identification de la PSF parmi les deux PSF possibles compte tenu de la génération expérimentale des images considérées dans cette étude : la première mono-paramétrique

(flou de défocalisation), la seconde bi-paramétrique (flou de mouvement uniforme). Les modèles retenus sont les suivants :

H_0 : Dirac

$$\begin{cases} h_{\mathbf{\Theta}_\delta}(n,m) = \delta(n,m) \text{ pour } (n,m) \in S_G \\ \mathbf{\Theta}_\delta^T = \{\emptyset\} \end{cases}$$

H_1 : Défocalisation

$$\begin{cases} h_{\mathbf{\Theta}_D}(n,m) = \begin{array}{ll} \frac{4}{\pi\, d^2} & \text{si } \sqrt{n^2+m^2} \leq \frac{d}{2} \\ 0 & \text{autrement} \end{array} \\ \mathbf{\Theta}_D^T = (d) \in S_D \end{cases}$$

H_2 : Mouvement uniforme

$$\begin{cases} h_{\mathbf{\Theta}_M}(n,m) = \begin{array}{ll} \frac{1}{d} & \text{si } \sqrt{n^2+m^2} \leq \frac{d}{2} \\ 0 & \text{autrement} \end{array} \\ \tan(\phi) = \frac{n}{m} \\ \mathbf{\Theta}_M^T = (d,\phi) \in S_M \end{cases}$$

La solution est donnée par [5] :

$$\mathbf{\Theta}^* = [\lambda^*, \mathbf{\Theta}_h^*]^T = \text{argmin}_{\mathbf{\Theta}\in\mathcal{H}}\, GCV(\mathbf{\Theta}) \qquad (2)$$

où $\mathcal{H}$ est l'ensemble des PSF candidates $\{H_0, H_1, H_2\}$ et où S_D, S_M, sont les intervalles de valeurs possibles pour les différents paramètres d, ϕ. L'expression du critère de la validation croisée généralisée est donnée par :

$$GCV(\lambda, \mathbf{\Theta}_h) = \frac{1}{\left(\sum_k \frac{\lambda\, |C_k|^2}{H(\mathbf{\Theta}_h)_k^2 + \lambda\, |C_k|^2}\right)^2} \qquad (3)$$

$$\sum_k \left(\frac{\lambda^2\, |C_k|^4}{(H(\mathbf{\Theta}_h)_k^2 + \lambda\, |C_k|^2)^2}\, |G_k|^2 \right) \qquad (4)$$

où λ est le paramètre de régularisation, C_k, $H_k(\mathbf{\Theta}_h)$ les valeurs propres des matrices C, $H(\mathbf{\Theta}_h)$ obtenues par transformation de Fourier discrète et G_k est la $k$ième composante de la transformée de Fourier discrète de G.
Le vecteur de paramètres $\mathbf{\Theta}_h$ dépend du type de PSF choisi : $\mathbf{\Theta}_D^T = (\, d\,)$ pour le flou de défocalisation, $\mathbf{\Theta}_M^T = (\, d, \phi\,)$ pour le mouvement uniforme.

3.3 Restauration de l'image

La phase suivante de restauration est obtenue par régularisation de Tikhonov-Miller. Cette régularisation a été développée dans un contexte déterministe [1]. Elle consiste en une fonctionnelle stabilisatrice ajoutée aux moindres carrés liés à l'inversion du modèle (1) :

$$\hat{f}_{TM} = \arg\min_f J(f) = \|g - H\, f\|^2 + \lambda \|C\, f\|^2 \qquad (5)$$

où C traduit la connaissance a priori sur f et λ est le paramètre de régularisation. En pratique, C est un filtre passe-haut, par exemple un opérateur de différentiation d'ordre peu élevé (premier ordre).

Cet algorithme peut fournir de bonnes valeurs initiales pour une procédure plus précise à la suite.

4 Premiers résultats expérimentaux

Nous avons tout d'abord généré des images dégradées par simulation numérique afin de pouvoir évaluer proprement et vérifier l'efficacité de la méthode proposée d'identification de la PSF. En effet, dans ce cas précis, le modèle d'observation considéré est exactement celui ayant été utilisé pour la génération des données. Ces premiers résultats ne sont pas décrits ici. Ils sont semblables aux résultats obtenus dans [5]. Puis, nous avons considéré des images dégradées réelles (en configuration statique : addition d'une bague allonge de 0,5 millimètres, puis de 1,5 millimètres ; en configuration dynamique : fréquence de 15Hz), issues de l'acquisition en laboratoire au LCPC, de façon à appréhender la robustesse de la méthode à la non adéquation idéale du modèle d'observation aux données réelles. Les figures suivantes (Fig.3), (Fig.4) et (Fig.5) montrent les images restaurées.

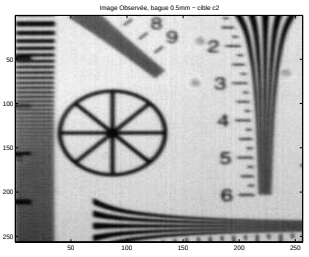


Image observée (256 × 256) - cible c2
bague allonge 0.5 mm

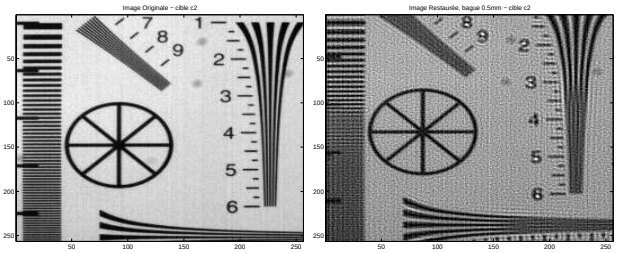


Image originale — Image restaurée (d=5, λ=0.005)

FIG. 3 – Restauration après identification du flou (bague allonge 0.5 mm) - $S_D = [1 : 99]$

Ces résultats montrent que les modèles retenus de PSF dans (2) sont bien adaptés pour les étapes d'identification et de restauration. Les résultats obtenus en configuration dynamique sont plus contrastés, compte-tenu de la non stationnarité du flou sur l'image considérée (Fig.5). Les essais en configuration dynamique sont actuellement en phase de rafinement. Si la PSF réelle n'appartient pas à l'ensemble des PSF candidates, un module de restauration aveugle peut être considéré pour compenser la distance entre les deux PSF. Il utilise comme valeur initiale de PSF, la valeur obtenue lors de l'étape précédente d'identification et réalise l'estimation alternée de la PSF et de l'image originale [2]. Pour réduire les coûts de calcul, il est possible d'effectuer l'étape d'identification (2) sur une image de résolution suffisamment réduite ("binned image" de taille 256×256, voire 128×128) sans altérer la qualité de l'identification, ni diminuer le taux de bonne identification (les profils des deux PSF candidates retenues étant

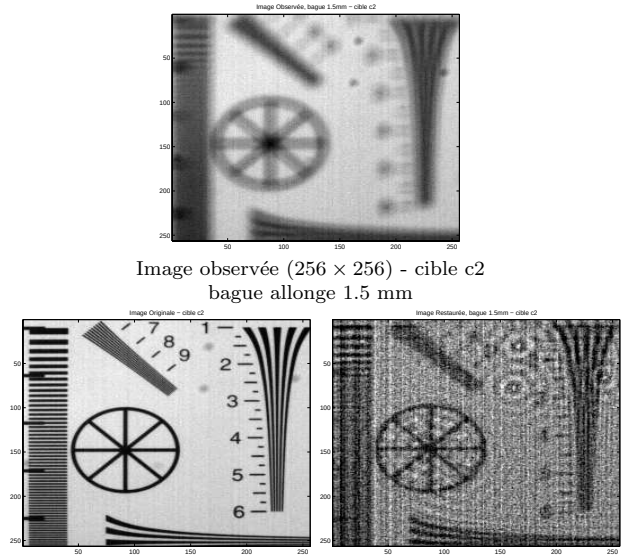


Image observée (256 × 256) - cible c2
bague allonge 1.5 mm

Image originale — Image restaurée (d=15, λ=0.0012)

FIG. 4 – Restauration après identification du flou (bague allonge 1.5mm) - $S_D = ([1 : 99])$

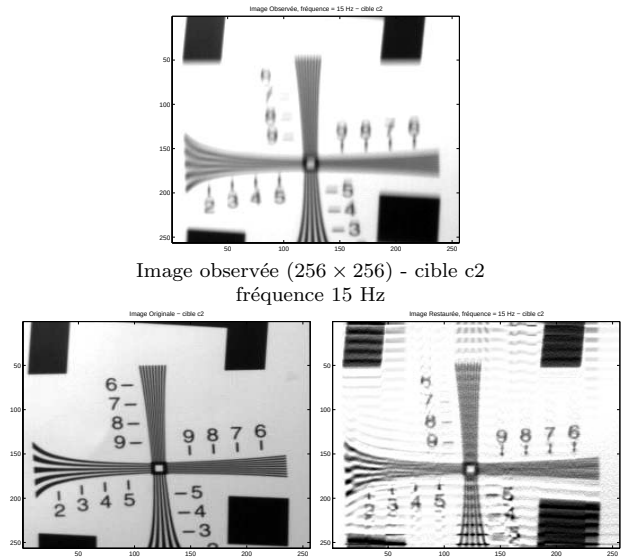


Image observée (256 × 256) - cible c2
fréquence 15 Hz

Image originale — Image restaurée (d=9, $\phi = \frac{\pi}{2}$, λ=0.005)

FIG. 5 – Restauration après identification du flou (fréquence 15 Hz) - $S_M = ([1 : 99], [0.8\frac{\pi}{2} : 1.2\frac{\pi}{2}])$

suffisamment distincts).

D'autres méthodes de restauration [10], [11], [12], peuvent également être utilisées, notamment afin de palier les inconvénients bien connus des méthodes d'inversion fréquentielle (oscillations observées (Fig.4)). Elles induisent néanmoins un coût de calcul plus important. Le choix définitif dépend surtout des traitements en aval (caractérisation et reconnaissance des fissures).

La méthode proposée et testée ici permet donc d'obtenir une première estimation de la PSF (sa nature : défocalisation, mouvement, et son support : valeur de d, ϕ). Cette estimation est ensuite utilisée pour définir la résolution d'une fenêtre glissante de restauration. Après réécriture de la PSF, le traitement des fenêtres successives de l'image dégradée de pleine résolution (1024 × 1280) donne finalement l'image restaurée.

5 Conclusion

La restauration numérique des images expérimentales obtenues en configuration statique et dynamique avec différents réglages des paramètres d'acquisition, montrent le potentiel du système global. Les travaux futurs incluent la détermination du taux d'identification sur la population complète d'images expérimentales ainsi que la restauration d'images expérimentales dégradées par une combinaison de flous de type défocalisation et mouvement uniforme.

Références

[1] Andrews H.C., Hunt B.R. " Digital Image Restoration ". Prentice-Hall, 1977.

[2] Y.-L. You, M. Kaveh. "A Regularization Approach to Joint Blur Identification and Image Restauration". IEEE Transactions on Image Processing, 5(3) :416-428, March 1996.

[3] R.L Lagendijk, J. Biemond, D.E. Boekee. "Regularized iterative image restauration with ringing reduction" IEEE Tr. ASSP, 36(12), 1874-1888, December 1988.

[4] Galatsanos N.P., Katsaggelos A.K. " Methods for Choosing the Regularization Parameter and Estimating the Noise Variance in Image Restoration and Their Relation " IEEE Tr. IP, July 1992, Vol. 1, No. 3, pp. 322-336.

[5] B. Vozel, K. Chehdi, M.-P. Vandecandelaere. "A Blind Restoration System of Blurred and Noisy Numerical Images". ICIP 2000, Vancouver, Canada, 10-13 September 2000.

[6] F Derkx, J. Dumoulin, J-L Sorin et V. Legeay, "Inspection of the Civil Engineering Structures : Interest of the UAV". Revue Essais Industriels,pp 23-27, mars 2003.

[7] Vitana Corporation. " Pixelink : Megapixel Firewire camera developer's manual". release 3.0, 121 pages, 2002.

[8] J-M Moliard. "New User's Guide of the PICTURE Image Processing Software ". août 2004.

[9] Dactron Inc. " Premier shaker control manual V5.0 ". 509 pages, janvier 2003 ;

[10] T.F. Chan and C.K. Wong, "Total variation blind deconvolution", IEEE Tr. IP, vol. 7(3), pp 370-375,mar 1998

[11] S. Teboul and L. Blanc-Féraud and G. Aubert and M. Barlaud "Variational approach for edge preserving regularization using coupled PDE's", IEEE Tr. IP, vol. 7(3), pp 387-397,mar 1998

[12] G. Aubert and P. Kornpobst, "Mathematical problems in image processing (Partial differential equations and the calculus of variations)", Springer, 2002

Scintigraphic Image Denoising using Fisz Transformation and Redundant Wavelet Packets

Khlifa Nawrès[1], Hamrouni Kamel[1], Ellouze Noureddine[1]

[1]Labortaoire des Systèmes et de Traitement du Signal : Ecole Nationale d'Ingénieurs de Tunis

Khalifa_nawres@yahoo.com, Kamel.Hamrouni@enit.rnu.tn, Noureddine.Ellouze@enit.rnu.tn

Résumé *— Le débruitage d'images par seuillage des coefficients d'ondelettes est une approche en pleine évolution vu le nombre croissant des publications qui s'intéressent à cette méthode de filtrage. Dans ce papier, nous proposons une nouvelle méthode de débruitage des images scintigraphiques basée sur la décomposition en ondelettes redondantes (utilisant une version modifiée du seuil de Bayes) et la transformation de Fisz pour changer la nature de la distribution du bruit d'une poissonienne à une gaussienne. En collaboration avec des spécialistes en médicine nucléaire, les résultats sont jugés importants.*

Abstract *— Image denoising by wavelet coefficients thresholding is an approach in full evolution considering the growing number of publications, which are interested in this methodology of filtering. In this paper, we propose a new method of denoising scintigraphic images based on redundant wavelet packet decomposition (using a modified version of Bayes's threshold) and Fisz transformation to transform the poissonnian noise in a Gaussian one. In collaboration with physicians of nuclear medicine, results are considered important.*

1. Introduction:

In medical imaging, noise suppression is a difficult task. A compromise between the preservation of useful diagnostic information and noise suppression must be respected. Generally, we rely on the intervention of an expert to control the quality of processed images. For example, in certain cases, such as the ultrasonic images, the noise can contain information which is useful for the physician. Medical images is thus, very variable, it is necessary to operate case by case.

In this paper, will present a new method to denoising scintigraphic images, using the Fisz transformation and the redundant wavelet packet. The originality of this method is his ability to suppress poison noise while preserving the useful details.

2. Denoising by wavelet coefficients thresholding:

Actually, denoising using wavelets proves his ability to satisfy the compromise between smoothing and conserving important features.

The first attempt of medical images denoising using wavelet was that of Weaver et al [5]. Results were encouraging except that the method eliminates small structures which were confused with the noise.

In 1994, Johnston and Donoho [6][7][8][9] formalized the principle of denoising by thresholding wavelet coefficients and since several work were published. [10,11,12,13,14,15,16]

At the beginning, the noise was supposed as Gaussien noise, white and additive. More recently, much of other work were interested in other models of noise. Thus the multiresolution denoising became, par excellence, a robust method of filtering.

In this paper, we propose a new method to denoise scintigraphic images, using the Fisz transformation and the redundant wavelet packet.

3. Denoisng scintigraphic images using Packet wavelet decomposition:

Wavelet packet decomposition gives better exploration of image details, so denoising using wavelet packet decomposition gives better results than classic wavelet decomposition. In the case of scintigraphic image, this decomposition will be associated with the Fisz Transformation.

In fact, in scintigraphic image, the noise distribution is Poissonnian [1]. We then propose the use of a transformation to change the nature of this noise into Gaussian.

In literature, several transformations were proposed such as that of Kolaczyk [2] and Anscombe.

In 2002, P.Fryzlewicz proposed a transformation known as the Fisz one for the monodimensional case [3]. More recently, in December 2003, an extension to 2D was proposed by Fadili which showed the considerable contribution of this transformation compared to other transformations [4].

Our denoising method is composed of the following steps:

- Apply the Fisz transformation.
- Apply our denoising method by thresholding the stationary wavelet packet coefficients:
- Apply the inverse Fisz transformation.

4. The Fisz transformation:[4]

If M indicates the original image of size N^2, m indicates the level of decomposition and I and J indicate the indices of the pixels in each image, the Fisz transformation is resumed in the following steps:

- Apply a level of decomposition by using the Haar wavelet with non normalised filters.

- we note D^m_{ij} the coefficients of details on each level of decomposition and A^m_{ij} the coefficients of approximation. The coefficients of details will then be modified as follows:

$$Z^m_{ij} = 0 \quad if \quad A^m_{ij} = 0$$

$$\frac{D^m_{ij}}{\sqrt{A^m_{ij}}} \quad else$$

- Repeat the first two stages on each level of decomposition (until $j=log_2(N)$) by keeping intact coefficients of approximation.

It results from this that the coefficients of details in all the scales and orientations, thus modified, are a Gaussien version of original coefficients.

- Reconstruct the new image N (by using the Haar wavelet with non normalised filters)

We can then write: $N = F(M)$ where F is the operator of the Fisz transformation.

5. Wavelet packet transform and best bases selection:

In the wavelet transform, one step calculates a low pass result and a high pass result. The low pass result is a smoother version of the original signal. The low pass result recursively becomes the input to the next wavelet step, which calculates another low and high pass result. For more details on the wavelet transform see .

As indicated below, the wavelet transform applies the wavelet transform step to the low pass result. The wavelet packet transform applies the transform step to both the low pass and the high pass result.

The wavelet packet transform can be viewed as a tree. The root of the tree is the original image. The next level of the tree is the result of one step of the wavelet transform. Subsequent levels in the tree are constructed by recursively applying the wavelet transform step to the low and high pass filter results of the previous wavelet transform step.

The best basis algorithm finds a set of wavelet bases that provide the most desirable representation of the data relative to a particular cost function. One cost function that has been proposed in the literature is the Shannon entropy function.

The best basis set can be obtained by traversing the tree, top down, from left to right. The traversal on a branch stops when a best basis node is encountered and this node is added to the best basis list

6. Modified version of Bayes's threshold:

It is an adaptive threshold which depends on the data in each sub-band in each level. Indeed, a threshold T is applied to each sub-band:

$T=\frac{\beta\hat{\sigma}^2}{\sigma_y}$ Where β is a parameter calculated for each level of decomposition:

$$\beta = \sqrt{\frac{L_k}{J}}$$

with L K: the length of sub-band in K 2nd scale (equal to the length of the original image), and J the number of the level. $\hat{\sigma}^2$: noise variance estimated by the formula of Donoho and Johstone. σ_y: Standard deviation of sub-band.

7. Proposed method of thresholding:

In literature, soft threshloding is applied to just the last level of wavelet packet decomposition. So the noise variation is estimated just one time. And it is applied to all sub-bands (except that of approximation).

As shown in figure 2, in our method, we apply a one level redundant wavelet packet decomposition, the threshold is calculated, a soft thresholding is applied to different sub-bands (except that of approximation).

The noise variation is then re-estimated for each image to calculate the new value of the threshold witch will be applied to the next redundant wavelet packet decomposition.

8. Results:

The described method was applied to a variety of cardiac scintigrahpic images; results show the suppression of noise and the preservation of ventricles edges (figure 1). In collaboration with physician of nuclear medicine, results are considered important. We mention that wavelets were used in another work to desoining scintigraphic images [17] but the method doesn't respect the particularity of this type of images (bad spatial resolution, poissonnian noise…), it's use the universal thresholding witch known of its incapability to preserve small details.

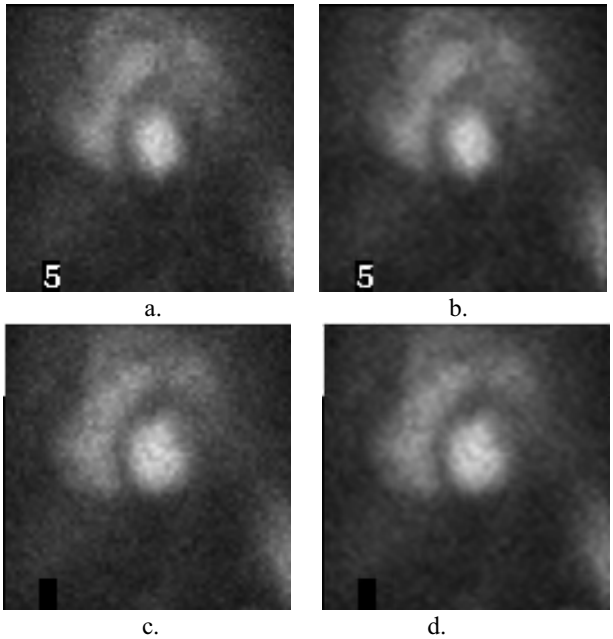


Figure 1: Scintigraphic image denoising : a,c: original images. b,d: denoising images

References:

[1]. Ridha Ben Salah, «Plethysmographie électrique thoracique localisée. Application à la détermination des paramètres cardiovasculaires au diagnostic des cardiopathies», Thèse à la Faculté des Sciences de Tunis. Juin 1988.

[2]. Kolaczyk. Wavelet shrinkage estimation of certain poisson intensity signals using corrected thresholds. Statistica Sinica, 9, pp. 119-135, 1999.

[3]. P. Fryzlewicz and G. P. Nason. A wavelet-fisz algorithm for poisson intensity estimation. Tech. Rep., Department of Mathematics, University of Bristol, 2002.

[4]. Jalal FADILI1, Jerôme MATHIEU1, Michel DESVIGNES2, La transformation de Fisz pour l'estimation d'images d'intensité Poissonnienne dans le domaine des ondelettes. GRETSI'03.vol. I, page 317

[5]. Weaver JB, Xu YS, Healy DM Jr, Cromwell LD. Filtering noise from images with wavelet transforms. Magn Reson Med. 21(2):288-95. Oct, 1991.

[6]. D.L. Donoho, I.M. Johnston. "Threshold selection for wavelet shrinkage of noisy data". Proc. 16th Annual Conf. of the IEEE Engineering in Medicine and Biology Society .1994.

[7]. D.L. Donoho. « Denoising by soft thresholding ». IEEE Transactions on Information Theory, Vol. 41, pp. 613617, Mar.1995.

[8]. D.L. Donoho, I.M. Johnston. "Ideal spatial adaptation via wavelet shrinkage", Biometrika, vol. 81, pp. 425-455, 1994.

[9]. D.L. Donoho, I.M. Johnston. "Adapting to unknown via wavelet shrinkage". Journal of the American Statistical Association, vol. 90, no. 432, pp. 1200-1224, December 1995.

[10]. Bruce and H. Y. Gao. "WaveShrink: Shrinkage Functions and Thresholds", Proc. SPIE Vol. 2569, pp. 270-281,1995.

[11]. Junhui Qian. "Denoising by wavelet transform". Tech. Rep. Rice University. Department of Electrical Engineering. 2001.

[12]. Bruce and H. Gao. "Applied wavelet Analysis with S-PLUS". Springer Verlag, 1996.

[13]. S. G. Chang, B. Yu, M. Vetterli. "Adaptive wavelet thresholding for image denoising and compression". IEEE Transaction on Image Processing, vol. 9, no. 9, pp. 1532-1546. Sep. 2000

[14]. Fodor, I.K. and C. Kamath. "Denoising through Wavelet Shrinkage: An Empirical Study". Journal of Electronic Imaging, vol. 12, pp.151-160, January 2003.

[15]. Khaliha Nawres, Hamrouni Kamel et Ellouze Noureddine. Etude comparative des méthodes de réduction de bruit par seuillage des coefficients d'ondelettes. Journées Scientifiques Francophones JSF'03. Douz-Tunisie. Décembre 2002.

[16]. Martin Vetteli et Grace Chang, Bin Yu. Adaptive wavelet thresholding for image denoising and compression. IEEE Transactions on Image Processing, 9: 1532-1546, Sep 2000.

[17]. N. S. Arikidis, J. Kalatzis, S.Tsantis, V. Prassopoulos, D. Cavouras. Suppression of high frequency noise from scintigraphic images using the discrete wavelet transform. Third European Symposium in Biomedical Engineering and Medical Physics, p:36, 30/08 – 1/09, 2002, Patras, Greece.

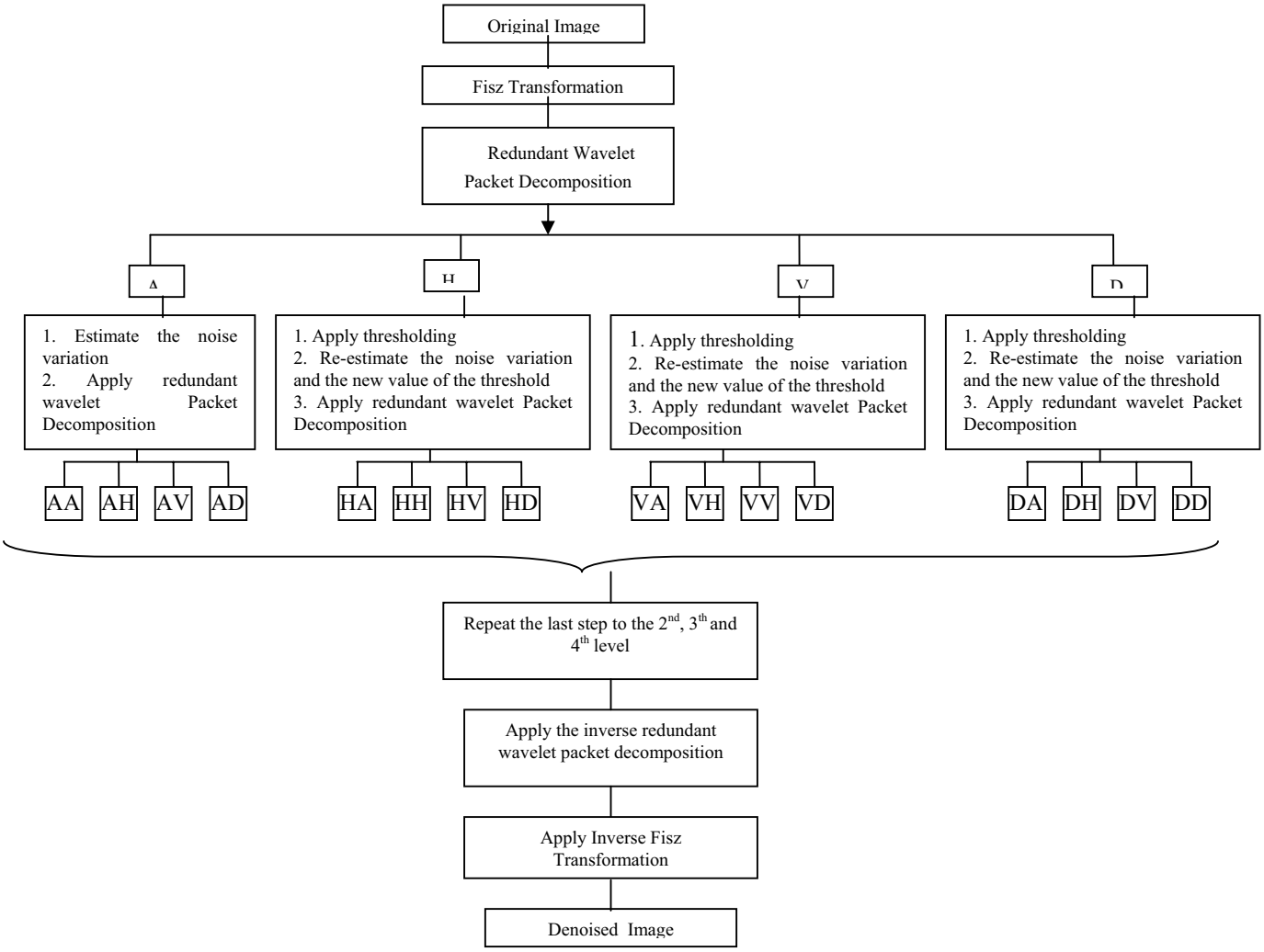


Figure 2: Architecture of the proposed method

Restauration des images floues et bruitées multiplicativement*

Luc KLAINE[1], Benoit VOZEL[2], Kacem CHEHDI[3]

[1, 2, 3]Équipe TSI2M de l'IETR de l'Université de Rennes I
6 rue de Kerampont, BP 80518, 22305 Lannion Cedex
luc.klaine@enssat.fr, benoit.vozel@enssat.fr, kacem.chehdi@enssat.fr

Résumé – Dans cet article, nous décrivons une méthode de restauration des images floues et bruitées multiplicativement inspirée du filtre de Kuan. Un modèle variationnel de forme intégrale ou différentielle est obtenu à partir de l'erreur quadratique moyenne globale. La forme intégrale introduit un noyau d'estimation locale équivalent à une réponse impulsionnelle localement adapté à l'image et ouvre des perspectives d'extension à d'autres types de noyau. La forme différentielle conduit à un nouveau modèle où l'*a priori* d'homogénéité dépend du gradient et de l'image.

Abstract – In this paper, we describe a method of restoration of the images perturbed by a blur and a multiplicative noise inspired by the Kuan filter. A variational model in an integral or a differential form is obtained from the total mean square error. The integral form introduces a kernel of local estimates, that is equivalent to a point spread function locally adapted to the image and opens perspectives for extension to other kind of kernel. The differential form leads to a new model in which the *a priori* homogeneity depends on the gradient and the image.

1 Introduction

Une nouvelle méthode de restauration par minimisation de l'erreur quadratique moyenne globale est décrite. L'image observée est supposée dégradée par un flou et un bruit multiplicatif de moyenne 1 décorrélé de l'image originale. L'idée est d'établir un lien entre le principe du filtre de Kuan et les méthodes de restauration variationnelles.

Trois formulations alternatives de l'erreur quadratique moyenne globale sont présentées dans la **section 2**. La première est stochastique et coïncide avec l'erreur quadratique moyenne de l'image globale. Elle justifie le couplage de cette énergie de régularisation avec un terme d'attache aux données. La seconde est intégrale et consiste à incorporer des statistiques locales. Elle autorise l'utilisation d'un noyau d'estimation locale adapté à la configuration locale de l'image. La troisième est différentielle et correspond au cas limite où le pas d'échantillonnage est proche de zéro. Le traitement devient alors indépendant de la résolution de l'image et permet d'établir un lien avec les méthodes variationnelles où l'*a priori* d'homogénéité est caractérisé par le gradient.

Enfin, les résultats expérimentaux de la nouvelle méthode de restauration comparant les formes intégrale et différentielle sont présentés dans la **section 3**.

2 Formulations alternatives de l'erreur quadratique moyenne globale

2.1 Forme stochastique

Supposons d'abord que l'image est simplement bruitée.

Pour rendre globale l'erreur quadratique moyenne locale (dépend de chaque pixel), il suffit de l'intégrer sur le support de l'image (*déf.* 2.1). Les propriétés du bruit conduisent à une simplification de l'erreur quadratique moyenne globale stochastique (*prop.* 2.1).

Définition 2.1 *Soient $u(\cdot)$ l'image originale, $n(\cdot)$ le bruit et $K(\cdot)$ le gain alors on définit l'*erreur quadratique moyenne globale *ou* énergie potentielle EQMG *stochastique $J(\cdot)$ par :*

$$J(u,K) \triangleq \frac{1}{2}\int_{\Omega}\mathbb{E}\Big[\Big(u-\big[\mathbb{E}[u.n]+K\big((u+n)-\mathbb{E}[u+n]\big)\big]\Big)^2\Big] \tag{1}$$

représentant la moyenne de l'erreur quadratique moyenne entre l'image originale $u(\cdot)$ et la sortie du filtre de Kuan [9, 8].

Proposition 2.1 *Soit σ_n^2 la variance du bruit alors l'énergie potentielle EQMG s'exprime également :*

$$J(u,K) = \frac{1}{2}\int_{\Omega}\Big[(1-K)^2\mathbb{V}[u] + K^2\sigma_n^2\mathbb{E}[u^2]\Big]. \tag{2}$$

PREUVE : Conformément au modèle de pseudo-observation introduit par Kuan, on pose $b = u.(n-1)$. D'après les hypothèses sur le bruit multiplicatif, on a $\mathbb{E}[b]=1$, $\mathbb{V}[b]=\mathbb{E}[u^2]\sigma_n^2$ et $\mathbb{E}\Big[\big(u-\mathbb{E}[u]\big)\big(b-\mathbb{E}[b]\big)\Big]=0$ et on obtient :

$$\begin{aligned}
&J(u,K)\\
&= \frac{1}{2}\int_{\Omega}\mathbb{E}\Big[\big(u-\mathbb{E}[u]\big)^2 + K^2\big(u-\mathbb{E}[u]\big)^2 + K^2\big((b)-\mathbb{E}[b]\big)^2\\
&\qquad -K^2\big(u-\mathbb{E}[u]\big)\big(b-\mathbb{E}[b]\big) - 2K\big(u-\mathbb{E}[u]\big)^2\\
&\qquad -2K\big(u-\mathbb{E}[u]\big)\big(b-\mathbb{E}[b]\big)\Big]\\
&= \frac{1}{2}\int_{\Omega}\Big(\mathbb{V}[u]+K^2\mathbb{V}[u]+K^2\mathbb{V}[b]-2K\mathbb{V}[u]\Big)\\
&= \frac{1}{2}\int_{\Omega}\Big[(1-K)^2\mathbb{V}[u]+K^2\mathbb{E}[u^2]\sigma_n^2\Big].
\end{aligned}$$

*Ces travaux sont financés par le Conseil Regional de Bretagne et le FEDER au travers du projet Interreg 3B PIMHAI.

■

Supposons maintenant que l'image est également dégradée par un flou. On décide arbitrairement de ne pas introduire le flou dans l'énergie potentielle EQMG. On impose cependant au minimum de rester suffi samment proche de l'image observée conformément au modèle de dégradation. L'énergie potentielle EQMG perturbée par un terme d'attache aux données est appelée énergie totale EQMG :

$$J(\lambda;u,K) \triangleq J(u,K) + \lambda \int_\Omega \Big| \frac{u_0}{H(u)} - 1 \Big|^2. \qquad (3)$$

Le terme d'attache aux données caractérisant le bruit multiplicatif est issu des travaux de Rudin-Osher [11]. Le procédé est largement utilisé pour les méthodes variationnelles [2]. L'énergie potentielle EQMG est perçue comme une énergie de régularisation. Elle permet d'introduire un *a priori* d'homogénéité pour reconstruire proprement l'image. L'énergie potentielle ou totale EQMG dépendent de l'image mais aussi du gain. La première raison est que la sortie du fi ltre de Kuan ne donne pas la meilleure image au sens de l'erreur quadratique moyenne. La seconde raison est que l'énergie potentielle EQMG est quadratique donc convexe par rapport à l'image ou au gain séparément. L'algorithme naturel qui en découle consiste à minimiser alternativement [3, 4] l'énergie totale EQMG bien qu'il n'y ait aucune garantie quant à l'existence d'un couple gain et image optimal.

2.2 Forme intégrale

Jusqu'à présent l'énergie potentielle EQMG était stochastique, on désire la rendre déterministe. L'espérance et la variance sont remplacées par des statistiques locales. On obtient l'énergie potentielle EQMG intégrale (*déf.* 2.2)

Définition 2.2 *Soient $\Phi(\cdot) \in L^\infty(\Omega^2)$ un noyau d'estimation locale, $u(\cdot) \in L^2(\Omega)$ l'image originale et $K(\cdot) \in L^\infty(\Omega)$ le gain alors on définit l'*énergie potentielle EQMG intégrale *par :*

$$J(u,K) \triangleq \frac{1}{2} \int_\Omega \Big[(1-K)^2 \mathrm{var}_\Phi[u] + K^2 \sigma_n^2 \mathrm{moy}_\Phi[u^2] \Big]. \quad (4)$$

On rappelle que les statistiques locales peuvent être défi nies à partir d'un noyau d'estimation locale $\Phi(\cdot) \in L^\infty(\Omega^2)$ par :

$$\mathrm{moy}_\Phi[u](x) \triangleq \|\Phi(x)\|^{-1}_{L^1(\Omega)} \int_\Omega \Phi(x,y) u(y) \mathrm{d}y,$$
$$\mathrm{var}_\Phi[u](x) \triangleq \|\Phi(x)\|^{-1}_{L^1(\Omega)} \int_\Omega \Phi(x,y) \big| u(y) - \mathrm{moy}_\Phi[u](x) \big|^2 \mathrm{d}y$$

où la densité d'estimation locale vaut $\Phi(x)(y) \triangleq \Phi(x,y)$. Afi n de minimiser alternativement l'énergie totale EQMG, on détermine l'image optimale en fonction du gain et le gain optimal en fonction de l'image par dérivation au sens de Gâteaux (*prop.* 2.2 et 2.3).

Proposition 2.2 *Le gain optimal $K(\cdot)$ lorsque l'image $u(\cdot)$ est fixée vaut :*

$$K(x) = \frac{\mathrm{var}_\Phi[u](x)}{\mathrm{var}_\Phi[u](x) + \sigma_n^2 \mathrm{moy}_\Phi[u^2](x)}. \qquad (5)$$

PREUVE : La dérivée de $J(u,\cdot)$ par rapport à $K(\cdot)$ vaut :

$$\langle J_K(u,K), K_h \rangle = \int_\Omega \Big[K\big(\mathrm{var}_\Phi[u] + \sigma_n^2 \mathrm{moy}_\Phi[u^2]\big) - \mathrm{var}_\Phi[u] \Big] K_h.$$

et s'annule pour le gain $K(\cdot)$ donné par la proposition. ■

Proposition 2.3 *L'image optimale $u(\cdot)$ lorsque le gain $K(\cdot)$ est fixé vérifie l'équation :*

$$\int_\Omega \|\Phi(y)\|^{-1}_{L^1(\Omega)} \Phi(y,x) \Big[(1-K)^2(y) \big[u(x) - \mathrm{moy}_\Phi[u](y) \big] + K^2(y) \mathrm{moy}_\Phi[u](y) \Big] \mathrm{d}y$$
$$- \lambda H^* \left[\frac{u_0}{H(u)^2} \right] \left[\frac{u_0}{H(u)} - 1 \right] (x) = 0 \text{ sur } \Omega \qquad (6)$$

où $H^*(\cdot)$ est l'adjoint de $H(\cdot)$.

PREUVE : Les dérivées de $\mathrm{moy}_\Phi[u^2](x)$ et $\mathrm{moy}_\Phi[u]^2(x)$ par rapport à $u(\cdot)$ valent respectivement :

$$\lim_{\varepsilon \to 0+} \varepsilon^{-1} \Big[\mathrm{moy}_\Phi[(u+\varepsilon u_h)^2] - \mathrm{moy}_\Phi[u^2] \Big](x)$$
$$= \int_\Omega 2 \|\Phi(x)\|^{-1}_{L^1(\Omega)} \Phi(x,y) u(y) u_h(y) \mathrm{d}y$$

$$\lim_{\varepsilon \to 0+} \varepsilon^{-1} \Big[\mathrm{moy}_\Phi[u+\varepsilon u_h]^2 - \mathrm{moy}_\Phi[u]^2 \Big](x)$$
$$= \int_\Omega 2 \|\Phi(x)\|^{-1}_{L^1(\Omega)} \Phi(x,y) \mathrm{moy}_\Phi[u](x) u_h(y) \mathrm{d}y.$$

La dérivée de Gâteaux de $J(\cdot,K)$ par rapport à $u(\cdot)$ vaut :

$$\langle J_u(u,K), u_h \rangle = \int_\Omega u_h(x) \int_\Omega \|\Phi(y)\|^{-1}_{L^1(\Omega)} \Phi(y,x)$$
$$\Big[(1-K)^2(y) \big[u(x) - \mathrm{moy}_\Phi[u](y) \big] + K^2(y) \mathrm{moy}_\Phi[u](y) \Big] \mathrm{d}y \mathrm{d}x$$

La dérivée du terme d'attache aux données s'identifi ant à :

$$-H^* \left[\frac{u_0}{H(u)^2} \right] \left[\frac{u_0}{H(u)} - 1 \right],$$

celle de l'énergie totale s'annule pour l'image $u(\cdot)$ donnée par la proposition. ■

En toute rigueur, les propriétés ne sont valables que pour un noyau d'estimation locale indépendant de l'image. Toutefois, si le noyau d'estimation est construit à l'aide d'une fonction de seuillage, la dérivée est nulle presque partout sur la sous-fenêtre. En d'autre terme, la contribution de la réponse impulsionnelle par rapport à l'image reste négligeable. Par conséquent, il est possible d'utiliser un noyau d'estimation locale adaptatif dans l'algorithme de restauration.

2.3 Forme différentielle

Intéressons-nous au cas limite où le pas d'échantillonnage spatial est très faible ou tend vers zéro. De manière équivalente, la taille de la sous-fenêtre tend également vers zéro. En conséquence, les opérateurs intégraux deviennent des opérateurs différentiels. Afi n de bien le comprendre, on étudie trois noyaux d'estimation locale signifi catifs.

- SUR LE DISQUE DE RAYON ε : la moyenne locale est consistante avec le Laplacien à l'ordre 2, la variance locale avec le carré du gradient à l'ordre 2 (*lem.* 2.1). On retrouve la propriété d'isotropie du Laplacien puisque le disque est invariant par rotation ;

Lemme 2.1 *Soit $\Phi(x,y) = \psi_\varepsilon(y-x)$ où $\psi_\varepsilon(\cdot)$ est la fonction indicatrice sur le disque de rayon ε alors pour toute image $u(\cdot) \in C^3(\Omega)$:*

$$\begin{aligned} \mathrm{moy}_\Phi[u](x) &= u(x) + \frac{\varepsilon^2}{8}\Delta u(x) + o\big(\varepsilon^2\big), \\ \mathrm{var}_\Phi[u](x) &= \frac{\varepsilon^2}{4}|\nabla u(x)|^2 + o\big(\varepsilon^2\big). \end{aligned} \tag{7}$$

- SUR LE MASQUE DE LEE DE RAYON ε : la moyenne locale est consistante avec le gradient à l'ordre 1, la variance locale avec le carré du gradient à l'ordre 2 (*lem.* 2.2). La moyenne locale n'est finalement pas très bien adaptée puisque elle se ramène simplement à un opérateur à l'ordre 1 ;

Lemme 2.2 *Soit $\Phi(x,y) = H\big(-\langle\nabla u(x), y-x\rangle\big)\psi_\varepsilon(y-x)$ où $\psi_\varepsilon(\cdot)$ est la fonction indicatrice sur le disque de rayon ε et $H(\cdot)$ la fonction de Heavyside alors pour toute image $u(\cdot) \in C^3(\Omega)$:*

$$\begin{aligned} \mathrm{moy}_\Phi[u](x) &= u(x) - \frac{4\varepsilon}{3\pi}|\nabla u(x)| + \frac{\varepsilon^2}{8}\Delta u(x) + o\big(\varepsilon^2\big), \\ \mathrm{var}_\Phi[u](x) &= \varepsilon^2\Big(\frac{1}{4} - \frac{16}{9\pi^2}\Big)|\nabla u(x)|^2 + o\big(\varepsilon^2\big). \end{aligned} \tag{8}$$

- SUR LE SEGMENT PARALLÈLE AU CONTOUR DE LONGUEUR 2ε : la moyenne locale est consistante avec la dérivée seconde tangentielle à l'ordre 2, la variance locale est négligeable à l'ordre 2 (*lem.* 2.3). La variance locale ne caractérise pas le désordre au voisinage du point.

Lemme 2.3 *Soit $\Phi(x,y) = \delta\big(\langle\nabla u(x), y-x\rangle\big)\psi_\varepsilon(y-x)$ où $\psi_\varepsilon(\cdot)$ est la fonction indicatrice sur le disque de rayon ε et $\delta(\cdot)$ la masse de Dirac alors pour toute image $u(\cdot) \in C^3(\Omega)$:*

$$\begin{aligned} \mathrm{moy}_\Phi[u](x) &= u(x) + \frac{\varepsilon^2}{12}u_{TT}(x) + o\big(\varepsilon^2\big), \\ \mathrm{var}_\Phi[u](x) &= o\big(\varepsilon^2\big). \end{aligned} \tag{9}$$

où T désigne le vecteur tangent ou vecteur unitaire parallèle au contour.

Les résultats de consistance entre la variance et le carré du gradient sont introduits sous le signe intégral. On obtient l'erreur quadratique moyenne globale différentielle (*déf.* 2.3).

Définition 2.3 *Soient $u(\cdot) \in H^1(\Omega)$ l'image originale, σ_n^2 la variance du bruit et $K(\cdot) \in L^\infty(\Omega)$ le gain alors on définit l'*énergie potentielle EQMG différentielle *$J(\cdot)$ par :*

$$J(u,K) \triangleq \frac{1}{2}\int_\Omega \Big[(1-K)^2|\nabla u|^2 + K^2\sigma_n^2u^2\Big]. \tag{10}$$

On obtient de nouvelles formules pour caractériser l'image optimale lorsque le gain est fixé et le gain optimal lorsque l'image est fixée (*prop.* 2.4 et *prop.* 2.5).

Proposition 2.4 *Le gain optimal $K(\cdot)$ lorsque l'image $u(\cdot)$ est fixée vaut :*

$$K = \frac{|\nabla u|^2}{|\nabla u|^2 + \sigma_n^2u^2}. \tag{11}$$

PREUVE : La dérivée de Gâteaux de $J(u,\cdot)$ par rapport à $K(\cdot)$ vaut :

$$\langle J_K(u,K), K_h\rangle = \int_\Omega \Big[K\big(|\nabla u|^2 + \sigma_n^2u^2\big) - |\nabla u|^2\Big]K_h.$$

et s'annule pour le gain $K(\cdot)$ donné par la proposition. ■

Proposition 2.5 *L'image optimale $u(\cdot)$ lorsque le gain $K(\cdot)$ est fixé vérifie :*

$$\begin{cases} -\mathrm{div}\Big((1-K)^2\nabla u\Big) + K^2\sigma_n^2u \\ \qquad -\lambda H^*\left[\dfrac{u_0}{H(u)^2}\right]\left[\dfrac{u_0}{H(u)} - 1\right] = 0 \text{ sur } \Omega \\ (1-K)^2u_N = 0 \qquad\qquad \text{sur } \partial\Omega \end{cases} \tag{12}$$

où N désigne le vecteur normal unitaire intérieur au bord.

PREUVE : La dérivée de Gâteaux de $J(\cdot,K)$ par rapport à l'image $u(\cdot)$ vaut d'après la formule de Green :

$$\begin{aligned} \langle J_u(u,K), u_h\rangle = &\int_{\partial\Omega}\Big[(1-K)^2u_N\Big]u_h \\ &- \int_\Omega\Big[\mathrm{div}\Big((1-K)^2\nabla u\Big) - K^2\sigma_n^2u\Big]u_h. \end{aligned}$$

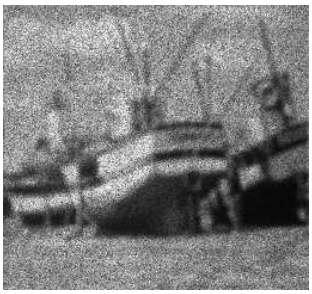
D'après la formule de la dérivée du terme d'attache aux données, la dérivée de l'énergie totale s'annule précisément lorsque l'image $u(\cdot)$ vérifie les équations précédentes. ■

Tous les résultats obtenus via la dérivée de Gâteaux sont purement formels. Le gain optimal lorsque l'image est fixée existe, il est borné et compris entre 0 et 1. Comme le gain est borné, les résultats classiques sur les équations aux dérivées partielles garantissent l'existence d'une solution faible de l'équation caractéristique de l'image optimale dans l'espace de Sobolev $H^1(\Omega)$. La minimisation alternée assure l'existence de l'image optimale et du gain optimal à chaque itération. En aucun cas, elle n'assure la converge du procédé vers un couple image/gain optimal.

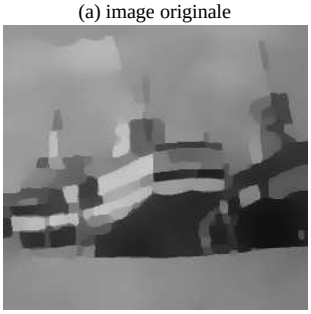
3 Etude expérimentale

Les tests ont été menés sur les images **[bateau]** et **[voiture]** dégradées par un flou gaussien ou de défocalisation et un bruit d'écart-type $\sigma_n = 0.2$ (*fig.* 1 et 2). Les deux méthodes de restauration testées sont basées sur la minimisation alternée de l'énergie EQMG en fonction du gain et de l'image. La première méthode forme différentielle est directement issue de la minimisation de l'énergie totale EQMG différentielle. La seconde méthode forme intégrale repose sur un noyau d'estimation locale adaptatif [6, 7] conformément au développement sur l'énergie totale EQMG intégrale. Le noyau d'estimation locale est construit de sorte à tenir compte de la nature des pixels selon qu'ils se trouvent dans une région homogène ou près d'un contour. Il coïncide exactement avec la réponse impulsionnelle du filtre moyenneur de Pomalaza-Raes [5] dans les zones homogènes. Il coïncide précisément avec la réponse impulsionnelle du filtre moyenneur de Asano-Yokoga [1] près des contours. Les résultats objectifs des simulations (*tab.* 1) regroupent les critères habituels, à savoir l'erreur moyenne absolue, l'erreur moyenne quadratique, l'erreur maximale et le

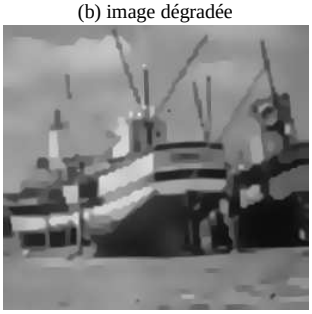
(a) image originale



(b) image dégradée



(c) image restaurée forme différentielle

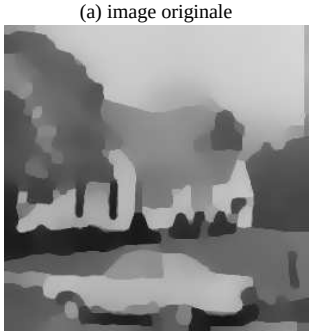


(d) image restaurée forme intégrale

FIG. 1 – Restauration de l'image **[savoise]** pour un flou gaussien et un bruit multiplicatif d'écart-type $\sigma_n = 0.2$

(a) image originale

(b) image dégradée



(c) image restaurée forme différentielle

(d) image restaurée forme intégrale

FIG. 2 – Restauration de l'image **[voiture]** pour un flou de défocalisation et un bruit multiplicatif d'écart-type $\sigma_n = 0.2$

rapport signal/bruit crête. On constate visuellement que l'utilisation du noyau d'estimation locale adaptatif affine la restauration. En effet, les bords semblent plus francs et le flou moins présent dans l'image ainsi restaurée. Toutefois, on observe un effet d'escalier sur les différents bords obliques. Cela montre que la configuration locale de l'image n'est pas totalement respectée. Par conséquent, le noyau d'estimation locale mérite d'être encore amélioré. Globalement, les résultats objectifs confirment l'impression visuelle. En effet, la méthode de restauration basée sur la réponse impulsionnelle adaptative minimise quasiment toujours les différents critères d'erreur.

4 Conclusion

L'énergie potentielle EQMG correspond à une énergie de régularisation et est adaptée à la restauration des images floues et bruitées. Cependant, l'existence et l'unicité du minimum ne sont pas assurées conformément au modèle de Geman-McClure où la fonction régularisante n'est pas convexe vu comme un cas particulier du modèle de Perona-Malik [10] mal posé. Par conséquent, la convergence de l'algorithme n'est pas assurée et peut générer une solution *a priori* instable. Toutefois, il semble au moins aussi légitime d'introduire un terme d'attache aux données dans l'énergie totale EQMG que dans le modèle de Geman-McClure. Or bien que ne s'appuyant pas à l'heure actuelle sur une théorie bien formalisée, les énergies de régularisations non-convexes sont largement utilisées d'autant que les résultats sont globalement prometteurs. Enfin, signalons que l'optimisation du noyau d'estimation locale adaptatif devrait également améliorer sensiblement les résultats.

TAB. 1 – Tableau comparatif des méthodes de restauration appliquées aux images **[bateau]** et **[voiture]** pour différents flous et un bruit multiplicatif d'écart-type $\sigma_n = 0.2$

		Critères de comparaison			
		mae	**mse**	**me**	**psnr**
[bateau]	dégradée	23.205	30.458	168	18.457
	forme différentielle	14.259	21.170	155	21.65
	forme intégrale	**12.248**	**16.545**	126	**23.791**
[voiture]	dégradée	25.038	32.450	157	17.906
	forme différentielle	18.124	25.101	151	20.171
	forme intégrale	**11.894**	**18.569**	**105**	**22.780**

Références

[1] T. Asano and N. Yokoga. Image segmentation scheme for low level computer vision. *Pattern Recognition*, 1981.

[2] G. Aubert and P. Kornpobst. *Mathematical problems in image processing (Partial differential equations and the calculus of variations)*. Springer, 2002.

[3] P. Charbonnier. *Reconstruction d'image : Régularisation avec prise en compte des discontinuités*. PhD thesis, Université Sophia-Antipolis de Nice, September 1994.

[4] P. Charbonnier, K. Blanc-Féraud, G. Aubert, and M. Barlaud. Deterministic edge-preserving regularization in computed imaging. *IEEE Transactions on Image Processing*, 7(3) :370–375, March 1997.

[5] Y.S. Fong, C.A. Pomalaza-Raez, and X.H. Wang. Comparison strudy of non-linear studies in image processing applications. *Optical Engineering*, 28(7) :749–760, 1989.

[6] L. Klaine, B. Vozel, and K. Chehdi. Adaptive filtering of multiplicative noise by a new differential method. In *Proceedings of ISSPA' 2003*, July 2003.

[7] L. Klaine, B. Vozel, and K. Chehdi. An integro-differential method for adaptive filtering of additive or multiplicative noise. In *Proceedings of ICASSP' 2005*, March 2005.

[8] D.T. Kuan. Adaptative noise smoothing filter for images with signal dependant noise. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, PAMI-7(2) :165–177, 1980.

[9] J.S. Lee. Digital image enhancement and noise filtering by use of local statistics. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, PAMI-2(2) :165–168, 1980.

[10] P. Perona and J. Malik. Scale space and edge detection using anisotropic diffusion. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 12(7) :629–639, 1990.

[11] L.I. Rudin and S. Osher. Total variation based image restoration with free local constraints. In *IEEE International Conference on Image Processing*, volume 1, pages 31–35, November 1994.

Une nouvelle transformation pour la localisation des yeux dans une image de visage monochrome

M. MILGRAM L. PREVOST R. BELAROUSSI

[1] Université Pierre & Marie Curie, Laboratoire des Instruments & Systèmes d'Ile de France Groupe PARC

4 Place Jussieu, 75252 Paris Cedex 5, BC252

Résumé – Nous présentons une méthode de localisation des yeux dans des images de visages et les résultats sur environ 3000 visages de la base ECU. Cette méthode fonctionne sur des images monochromes à l'aide de la Transformation Chinoise (TC). Celle ci est un outil original capable de détecter certaines formes par accumulations de votes. Après la TC, un modèle géométrique Gaussien est utilisé ainsi qu'une technique de projection des niveaux de gris afin de raffiner la localisation. Une notion de disimilarité entre les images des 2 yeux fournit un critère de validation. De nombreux résultats sont présentés et commentés.

Abstract *– A method for the localization of the eyes in a facial image and the results on nearly 3000 faces from the ECU database are presented. This method works on grayscale images, applying the Chinese Transformation (CT) on edge pixels. This transformation is an original tool able to detect patterns by accumulating votes. After the CT , we apply a Gaussian model to find eyes location and we refine the model with a grayscale profile approach. The dissimilarity between the 2 eyes is then introduced as a validation criterion. Experimental results on a large database is presented and commented.*

1. Introduction

Nous présentons dans cet article un nouvel algorithme de traitement d'images permettant d'estimer la position des yeux dans une image de visage. Cette image peut-être indifféremment monochrome ou couleur car nous n'utilisons pas l'information colorimétrique comme attribut pertinent ; l'appartenance ethnique ne modifie donc pas les performances du système. Les artefacts comme le port de lunette ne sont pas un obstacle majeur pour notre méthode. Cette approche est basée sur une technique inspirée de la Transformée de Hough (la « Transformée Chinoise ») qui utilise la disposition des pixels de contour ainsi que l'orientation du gradient en ces pixels. Elle est associée à des modèles probabilistes et à deux experts de raffinement et de validation. Les résultats expérimentaux, estimés sur une base de 2500 images de visages, sont très encourageants avec un taux de détection supérieur à 98% et une erreur de localisation relative supérieure à 92%.

La detection/localisation de visages est devenue un enjeu très important pour des applications comme la vidéo-conférence, la biométrie, la détection de présence, l'indexation d'images, etc. Cette détection est reliée à plusieurs sous-problèmes :

- détection d'éléments caractéristiques (bouche, yeux)
- détection et suivi du regard
- estimation de la pose
- analyse de l'expression faciale

Nous présentons ici une solution au sous-problème de la détection/ localisation des yeux. Pour cela, on trouve dans la littérature deux approches génériques, parfois combinées :

- l'approche par décisions binaires qui analyse un ensemble de sous-images et prend une décision pour chacune d'elles (exemple : *pattern matching*)
- l'approche globale qui traite directement le visage en entier (exemple : techniques de votes dans le style Transformée de Hough)

Dans les 2 cas, un codage préalable est parfois utilisé comme dans Huang et Weshler [10] où ce sont les ondelettes associées à un réseau RBF. Daugman [8] utilise aussi les ondelettes de Gabor pour localiser l'iris. D'autres comme Braathen, Bartlett et Littleworth-Ford [7] s'attaquent aux expressions faciales avec ce même type de codage. Les points caractéristiques (contours, coins) sont combinés à des traitements de morphologie mathématique par Gu, Su et Du [9] dans leur approche complètement ascendante du problème. La dimension fractale a été mise à contribution par Lin, Lam et Siu [11] qui l'emploie, combinée avec des contraintes géométriques. D'autres chercheurs proposent de localiser la partie blanche de l'œil, la sclera, et non l'iris à l'aide d'un modèle multi-gaussien [6].

Notre contribution comprendra une description des pré traitements suivie par la présentation de la TC. Les résultats de cette TC seront ensuite renforcés par notre modèle probabiliste et raffinés par une technique de profil. Enfin, nous présenterons et analyserons nos résultats sur la base ECU.

2. Pré-traitement et normalisation

L'étude a été mené sur la base ECU [5] qui comporte 2997 images RGB de visages. Parmi ces images, seules 2640 contiennent un visage où les deux yeux sont visibles. L'étiquetage de ces images (position des yeux) a été réalisé manuellement. Nous avons supposé que les visages étaient cadrés de manière relativement uniforme tout en nous efforçant d'accepter certaines variations dans le cadrage. Les visages ayant une orientation (angle entre l'horizontal et la droite joignant les yeux) supérieure à 30° ont été écartés automatiquement. Finalement, 2450 images ont passé cette étape de filtrage. C'est sur ces données qu'ont été estimés les moyennes empiriques mx_1, mx_2, my_1 et my_2 des abscisses (x_1 et x_2)et ordonnées (y_1 et y_2) des deux yeux et la matrice de variance-covariance géométrique Σ (supposée diagonale).

Les images RGB sont ramenées en niveau de gris par simple moyenne des 3 plans couleurs, sans prise en compte de connaissance colorimétrique *a priori* comme la teinte chair [11]. On détermine ensuite l'image *ORIENX* des orientations de gradient en évaluant les deux composantes *Gx* et *Gy* du gradient par un opérateur de Roberts, et sa direction $\alpha=arctan(Gy/Gx)$ discrétisée sur N valeurs (N=8). Le module du gradient permet de sélectionner les points de contour par seuillage absolu (le seuil a été déterminé empiriquement et validé *a posteriori* en fonction des résultats du système complet de localisation).

3. Transformée chinoise

La Transformation Chinoise (TC) est une transformation globale qui travaille sur les orientations de gradient des points de contour d'une image et fournit en sortie deux tableaux de vote. Le principal tableau de vote (*VOTE*), dans notre application, donne une information sur la position du centre des yeux. L'idée de base est que les yeux occupent dans le visage, en incluant les sourcils, une zone qui est plus sombre que son environnement immédiat et possède une forme elliptique. Du fait de la quasi symétrie centrale de l'ensemble, le centre de l'œil se trouve au milieu de nombreux segments joignant 2 points ayant des orientations de gradient opposés (figure 1). La TC peut rappeler la THG (qui utilise le gradient) mais elle en diffère fortement car elle peut détecter à la fois des ellipses de toute taille et excentricité. Elle repose sur les propriétés de symétrie centrale de l'ellipse.

Le principe général de la TC consiste d'abord à déterminer, pour chaque paire de pixels de contour *(M',M")* d'orientations respectives *(O',O")* *(O =0,1,..,N-1)* s'ils font ou non partie de l'ensemble des pixels votants. Le critère d'appartenance est basé sur la position relative de *M'* et *M"* et sur la condition *O'*=2 et *O"*=6 (soit Nord et Sud). Les points *M(x',y')* et *M"(x",y")* doivent vérifier les contraintes géométriques suivantes :

(1) $y'-y'' < \delta_{VERT}$

(2) $45° < \beta < 135°$ avec $\beta = arctan[(y'-y'') / (x'-x'')]$

(3) $| M'M'' | < \delta_{DIST}$

Des ajustements ont conduit à prendre δ_{VERT}=0.08xHauteur_image et δ_{DIST} =0.25xHauteur_image.

Deux tableaux de vote *VOTE* et *DIAM* (figure 2), de même taille que l'image *ORIENX* des orientations de gradient sont crées. Pour chaque paire *(M',M")* de l'ensemble des votants, on va incrémenter la valeur de *VOTE (M)* où *M* est le milieu du segment *[M'M"]*. (d'où le nom de Transformée Chinoise, en référence à la Chine, « Empire du Milieu »). Pour conserver une information sur la distance entre deux points votants pour un point *M*, on stocke dans *DIAM(M)* (comme diamètre) la moyenne des distances *d(M',M")*. Cette information permettra de déterminer un rectangle englobant chaque œil.

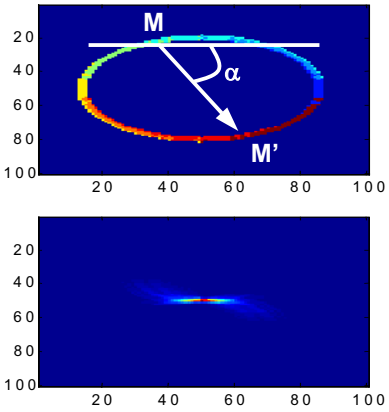


FIG. 1 : Principe de la Transformée Chinoise appliquée à une ellipse et tableau VOTE.

Une analyse *a posteriori* montre que la position des yeux correspond de manière assez fiable aux pixels ayant reçus le plus de votes. Malheureusement, les 2 premiers maxima (maximum absolu et second maximum) ne sont pas toujours les bons.

4. Modèle probabiliste

Pour renforcer le précédent détecteur, nous avons mis en œuvre un modèle probabiliste sur les positions des yeux dans le visage. Nous devons intégrer 2 types d'informations :

- les résultat du tableau de *VOTE* (information radiométrique)
- les positions relatives des paires candidates dans le visage (information géométrique)

Afin de normaliser les données issues du tableau de *VOTE* et pour ne pas être trop dépendant des variations inter-images, nous avons décidé d'utiliser pour quantifier la première information une loi géométrique (inspirée des « *neural gas* » [3]). La log-vraisemblance du $n^{ième}$ candidat est donc donnée par :

$LR(n) = log(Prob(n)) = - rang(n)$

où $rang(n)$ désigne la position du $n^{ième}$ candidat dans la liste des candidats ordonnées par score (résultat du tableau *VOTE*) décroissant. $LPR(n)$ ne prend donc en compte que le score lié à la radiométrie.

Le second terme LA (A pour abscisse) prend en compte la distribution statistique conjointe des abscisses des deux yeux (x_1, x_2) (normalisées entre 0 et 1 en divisant par la largeur de l'image) :

$$LA(x_1, x_2) = ½\ X^T \Sigma^{-1} X \text{ où } X=[x_1-mx_1\ x_2-mx_2]^T$$

Finalement, nous supposons classiquement l'indépendance des 2 types d'évènements pour mesurer la log-vraisemblance du couple (x_1, x_2):

$$LC(n, m) = LA(x1(n,m) , x2(n,m)) + LR(n) + LR(m)$$

où $x_1(n,m)$ désigne l'abscisse du candidat n ou m selon leur position et $x_2(n,m)$ l'abscisse du candidat restant.

Dans un premier temps, on sélectionne le couple (n^*,m^*) qui maximise la log-vraisemblance $LC(n^*,m^*)$. Si les 2 positions ainsi déterminées correspondent effectivement à 2 yeux, la vraisemblance $LC(n^*,m^*)$ doit être élevée (en fait supérieure à un seuil δ_{LPC} déterminé empiriquement à la valeur –100). Sinon , on va rejeter ce couple et :

- chercher la détection la plus probable (et l'œil correspondant) selon un modèle gaussien pour chaque œil
- prédire la position du second œil en utilisant un prédicteur linéaire d'ordre 1 :

$x_2 = mx_2 + (x_1 - mx_1).C_{12}$ et $y_2 = y_1$ (C_{12} coefficient de corrélation empirique centré entre x_1 et x_2)

Si nous n'avons qu'un seul candidat (maximum local dans le tableau de VOTE), ce qui précède nous permet aussi de prédire la position du second œil.

5. Amélioration de la précision

A ce stade, l'erreur selon l'axe vertical est nettement dominante car la TC a tendance à générer des confusions avec le sourcil dont l'apparence est parfois très proche de celle de l'œil. Nous avons donc développé un expert de raffinement et un expert de validation.

L'expert de raffinement permet d'augmenter de manière substantielle la précision de la détection selon l'axe vertical en analysant les profils verticaux des deux yeux (valeur moyenne des niveaux de gris pour chaque ligne de la vignette contenant l'œil). La position du centre de l'œil correspond à un minimum local du profil vertical, mais plusieurs minima locaux sont parfois détectés dus aux sourcils ou à l'ombre présente sous l'œil (figure 2). Seuls les deux premiers minima locaux sont conservés, le second pouvant être utilisé en cas de rejet du premier par le critère de dissimilarité présenté dans la suite.

L'expert de validation permet de filtrer une partie importante des fausses détections en comparant l'image de l'œil droit avec l'image miroir de l'œil gauche obtenue par symétrie verticale. On mesure la dissimilarité des candidats en évaluant la distance euclidienne entre les deux vignettes normalisées. Si cette dissimilarité est trop faible, on considère la détection comme peu fiable et on choisit de la rejeter. Ceci provoque une augmentation du taux de rejet et une diminution de l'erreur de localisation. On notera que pour les visages inclinés ou ceux dont les pupilles ne sont pas centrées dans les yeux, cette dissimilarité peut être mise en défaut. Ceci n'est pas trop gênant dans la mesure où ce test ne nous sert qu'à rejeter des détections douteuses.

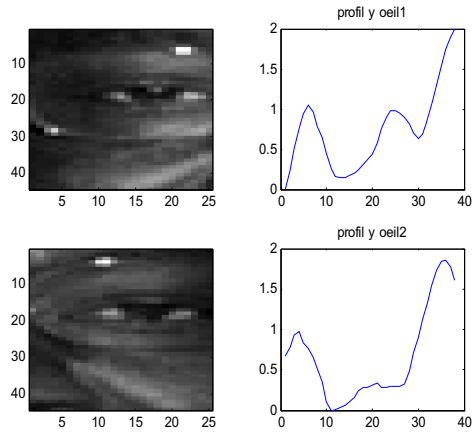


FIG 2 : profils verticaux des deux yeux et résultats de localisation.

6. Résultats expérimentaux

Nos mesures de performance tiennent compte du fait que nous demandons au système de nous fournir deux détections exactement. Dans les cas ou aucun maximum local n'est trouvé dans le tableau de *VOTE* (soit environ 2% des images de la base ECU), nous avons deux détections manquées. Dans les autres cas, le critère d'évaluation est la distance entre la détection (pour chaque œil) et la vérité terrain (normalisée, c'est à dire divisée par la distance entre les deux yeux).

L'utilisation du rejet avec un seuil de 0.1 sur les 2450 visages testés donne moins de 13% de rejet et une diminution significative de l'erreur de localisation, qui passe de 8% à 7.2% pour l'œil 1 et de 7.7% à 7.0% pour l'œil 2. Cette diminution peut sembler faible mais elle provient de l'élimination d'erreurs importantes (*outliers*). Ceci apparaît quand on observe les médianes au lieu des moyennes (tableau 1). Le gain est alors faible (de 6.0% à 5.7% pour l'œil 1 par exemple) ce qui est du au fait que l'on supprime surtout de grandes valeurs ayant une faible incidence sur cette médiane.

On a étudié en détail les cas de mauvaises détections sur les 300 premières images (cas où on obtient, pour au moins un des deux yeux, une erreur relative de position supérieure à 10%). Ceci produit 70 images, ce qui est cohérent avec la performance trouvée de 11% qui concerne la moyenne des distances relatives, d'où la proportion double quand on sélectionne les cas où une seule détection est distante. Nous avons classé ces erreurs par taux d'occurrence décroissant : lunettes (14%), yeux fermés (10% : confusion avec les sourcils), peau très sombre (7%), résolution trop faible (7% : moins de 15 pixels de large pour la vignette contenant un œil), cheveux sur les yeux (6%), bandeau (3%) … divers

(53%). La figure 3 rappelle les différentes étapes du processus de localisation.

TAB 1 : Résultats expérimentaux(avec seuil de dissimilarité : 0.1)

	Moyenne	Moyenne + rejet	Médiane	Médiane + rejet	Taux de rejet
Œil 1	0.080	0.072	0.060	0.058	0.13
Œil 2	0.077	0.070	0.059	0.056	0.13

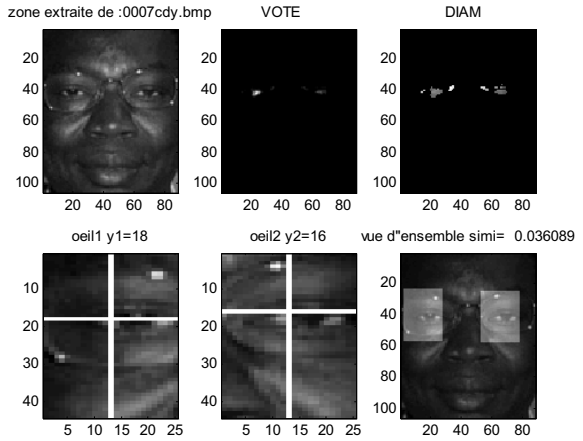


FIG 3 : image initiale, tableaux *VOTE* et *DIAM*, profil verticaux des deux yeux et résultats de localisation.

7. Conclusions

Une méthode globale de localisation des yeux dans une image de visage a été présentée et les résultats sur près de 2500 visages sont concluants. Elle repose sur une nouvelle transformation, dite « chinoise », qui peut servir dans d'autres contextes comme la détection de symétries axiales ou la localisation de patterns [4]. L'utilisation de cette technique pour la détermination de pose ou pour la confirmation d'une hypothèse de pose est la prochaine étape que nous nous sommes assignés. Par ailleurs, cette méthode pourra aussi être utilisée pour assister la détection du visage proprement dite. En effet, dans ce domaine, nous avons obtenu des performances intéressantes [1] mais avec un taux de fausse alarme élevé, donc perfectible. Précisons enfin que l'extension de cette méthode aux images couleurs de visages d'orientation non contrainte est en cours.

Références

[1] Belaroussi R., Prevost L. & Milgram M., Combining model-based classifiers for face localization, MVA'2005. Tsukuba, Japon, May 2005

[2] Hsu R.L. ,Abdel-Mottaleb. M. & Jain A. K., Face detection in color images, IEEE Trans. PAMI, 24(5):696-706, 2002.

[3] Martinetz T. & Schulten K., Topology representing networks, Neural Networks, 7(2), 1994.

[4] Negri P., Clady X. & Milgram M., Visual perception for human grasping gestures, ACIVS Conference, University of Antwerp, Antwerp, Belgium , September 2005

[5] Phung S. L., Bouzerdoum A. & Chai D., Skin segmentation using color pixel classification: Analysis and comparison, IEEE Trans. PAMI, to be published in 2005.

[6] Betke M., Mullally W. & Magee J., "Active Detection of Eye Sclera", *HMAS'2000 (Real Time Proceedings of the IEEE CVPR Workshop on Human Modeling, Analysis and Synthesis), 2000.*

[7] Braathen B. and Bartlett M. S. Littlewort-Ford G. & Movellan J. R., "3-D head pose estimation from video by nonlinear stochastic particle filtering", *UCSD MPLab TR 2001.05, 2001.*

[8] Daugman J.G., "Complete Discrete 2-D Gabor Transform by Neural Networks for Image analysis and Compression", *IEEE Trans. ASSP, 36(7):1169-1179, 1988.*

[9] Gu H, Su G and Du C., "Feature points extraction from face", *IVCNZ'2003 (Image and Vision Computing New Zealand), 2003.*

[10] Huang, J. & Wechsler H., "Eye location using genetic algorithm", *AVBPA'1999 (Audio and Video-Based Person Authentification), 1999.*

[11] Lin K.-H., Lam K.-M. & Siu, W.-C. "Locating the eye in human face images using fractal dimensions", *IEE Proc.Vis.Image Signal Process., 148(6) , 2001.*

Extraction d'un modèle 3D de visage en temps-réel et de manière robuste

Marc CHAUMONT

LIRMM, UMR CNRS 5506, Université de Montpellier II,
161 rue ADA, 34392 Montpellier Cedex 05, France.
marc.chaumont@lirmm.fr

Résumé – Cet article traite de l'extraction d'un modèle 3D de visage et de son positionnement 3D à partir d'un faible nombre de couples de points 2D-3D en correspondance. Cette extraction est envisagée comme une phase d'initialisation pour une application de suivi de visage temps-réel basé modèle 3D. L'inconvénient majeur des solutions actuelles pour l'extraction de modèle 3D sont soit la complexité calculatoire, soit la modélisation simplifiée. Notre solution, quant à elle, est rapide, robuste et suffisamment descriptive, ce qui la rend exploitable comme initialisation pour un suivi de visage temps-réel basé modèle 3D. Cette solution se décompose en deux étapes : une approximation suivie d'un raffinement d'un modèle 3D générique. Les résultats obtenus montre des performances en terme de robustesse, de rapidité et de réalisme.

Abstract – This article deals with 3D-face model and 3D-pose extraction from a small set of couples of 2D-3D corresponding-points. This extraction is considered as an initialization step of a real-time face tracking application based on a 3D model. Major drawbacks of current 3D model extraction solutions are either the computationally complexity or the over-simplified modeling. Our solution is rapid, robust and descriptive enough, which make it useful for a face tracking, in real-time, based on a 3D model. This solution is based on a two step approach: an approximation followed by a refinement of a generic 3D model. The results obtained show rapid, robust and realistic performances.

1 Introduction

Il existe deux grandes classes de techniques d'extraction de modèle 3D : la première classe nécessite un minimum de deux images pour faire un calcul de triangulation et retrouver les positions 3D des points utilisés lors de la triangulation [1] ; la seconde classe nécessite la connaissance d'un modèle 3D qu'il faut positionner [2]. Notons qu'une contrainte supplémentaire s'ajoute lors de l'extraction d'un modèle de visage, puisque les modèles 3D sont génériques et doivent être déformés pour être adaptés à la morphologie et aux émotions d'un visage particulier.

Ces deux approches sont coûteuses en temps de calcul et nécessitent pour être robustes de disposer d'un grand nombre de points 2D. Or, dans le cadre du suivi de visage par modèle 3D, lorsque les séquences vidéo sont acquises par une caméra non calibrée, la phase d'initialisation (c'est-à-dire d'extraction et de positionnement du modèle 3D) doit être à la fois robuste et temps-réel. La solution que nous proposons permet d'extraire en temps-réel un modèle 3D à partir d'un faible nombre de points 2D caractérisant le visage.

Le schéma général que nous proposons se base sur l'utilisation d'un modèle 3D générique et se compose de deux étapes. La première étape détaillée en partie 3 consiste à extraire une approximation du modèle 3D de visage. La seconde étape détaillée en partie 4 vise à améliorer le modèle 3D et à extraire sa position 3D. Pour introduire les deux parties relatives à notre contribution, nous formalisons au préalable le problème dans la partie 2.

2 Formulation énergétique générale

Avec une caméra classique de type sténopé, la projection T d'un point 3D $M'_i = (X'_i, Y'_i, Z'_i)^t$ (dans le repère objet) donne un point 2D $m'_i = (u'_i, v'_i)^t$ (dans le repère image) qui peut être exprimé en coordonnées homogènes [3] par l'équation 1. f est la longueur focale de la caméra ; k_u et k_v sont les facteurs d'échelle horizontaux et verticaux (mesurés en pixels/m) ; u_0 et v_0 sont les coordonnées du point principal ; $(t_x, t_y, t_z)^t$ est le vecteur de translation et $r_{ij:i,j\in[1,3]}$ sont les coeffi cients de la matrice de rotation. La Figure 1 illustre les différents systèmes de coordonnées ainsi que la projection d'un point 3D M'_i en un point 2D m'_i.

Formellement, l'extraction et le positionnement du modèle 3D de visage consiste à minimiser la distance E (voir l'équation 2) entre l'ensemble observé des points 2D de l'image $\{(u_i, v_i)^t\}$ et l'ensemble des points projetés $\{(u'_i, v'_i)^t\}$. L'ensemble des points projetés $\{(u'_i, v'_i)^t\}$ est obtenu en projetant les sommets du modèle 3D de visage, correspondant aux points 2D observés, en utilisant la projection T de l'équation 1.

$$E = \sum_i (u_i - u'_i)^2 + (v_i - v'_i)^2. \qquad (2)$$

Notons que les sommets du modèle 3D de visage utilisés dans la minimisation de l'équation 2 appartiennent à un modèle 3D de visage "*mis-en-forme*". Par "*mis-en-forme*", nous voulons dire que la morphologie et les émotions courantes du visage traité sont décrites par le modèle. Pour obtenir ce modèle 3D "*mis-en-forme*" nous déplaçons les sommets d'un modèle 3D générique. Le modèle 3D que nous avons retenu se nomme CANDIDE-3 [4] ; il décrit un visage moyen en position neutre par un maillage composé de 113 sommets et 168 facettes.

L'ensemble des points projetés $\{(u'_i, v'_i)^t\}$ (voir l'équation

$$\begin{pmatrix} s.u'_i \\ s.v'_i \\ s \end{pmatrix} = \underbrace{\begin{pmatrix} \alpha_u & 0 & u_0 & 0 \\ 0 & \alpha_v & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}}_{paramètres\ intrinsèques} . \underbrace{\begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix}}_{paramètres\ extrinsèques} . \begin{pmatrix} X'_i \\ Y'_i \\ Z'_i \\ 1 \end{pmatrix} = T_{3\times4}.M'_i,\ avec \begin{cases} \alpha_u = -k_u.f \\ \alpha_v = k_v.f \end{cases} . \quad (1)$$

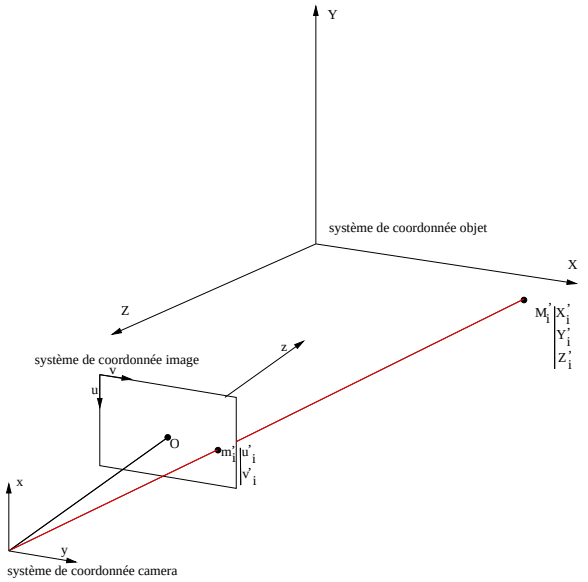


FIG. 1 – Les différents systèmes de coordonnée

2) est obtenu par l'application de trois opérations consécutives sur les sommets du modèle CANDIDE-3. Ces trois opérations

$$\begin{pmatrix} s.u'_i \\ s.v'_i \\ s \end{pmatrix} = T.\underbrace{[M_i + S_i\sigma + A_i.\alpha]}_{M'_i}. \quad (3)$$

sont : **1.** un *déplacement de forme* $S_i.\sigma$; **2.** un *déplacement d'animation* $A_i.\alpha$; **3.** la projection T. Ceci est exprimé par l'équation 3. S_i et A_i sont respectivement la *matrice d'unité de forme* et la *matrice d'unité d'animation*, exprimant le déplacement possible d'un sommet i du modèle CANDIDE-3. L'intensité du déplacement est exprimée par les vecteurs de pondération σ et α. Pour plus de détails, le lecteur pourra consulter le rapport d'Ahlberg [4]. La minimisation de l'équation 2 permet d'obtenir les paramètres T, σ et α.

Si elle est traitée directement, la minimisation de l'équation 2 est diffi cile, la solution n'est pas obtenue en temps-réel, et le résultat est peu robuste. Pour parvenir à une solution acceptable en peu de temps, nous proposons une solution qui se décompose en deux étapes. La première étape (partie 3) consiste à approximer la forme du modèle 3D de visage (T, σ et α sont grossièrement calculés). La seconde étape (partie 4) permet d'améliorer la forme du modèle 3D et d'extraire sa position 3D (T et la forme du modèle 3D sont raffi nés).

3 Approximation du modèle 3D de visage

3.1 Approximation de la pose

Le calcul de la projection T (obtenue en résolvant l'équation 2) n'est pas une tâche facile. L'expression immédiate de la solution mène en effet à un système linéaire homogène. Pour résoudre l'équation, il est classique d'introduire des contraintes sur la solution comme dans [5]. Pour cela, la connaissance d'un grand nombre de couples de points 2D-3D en correspondances est nécessaire. Or, dans le cas d'un visage Humain, le nombre de points caractéristiques est faible. Le nombre de couples de points 2D-3D en correspondance n'est donc pas suffi sant pour utiliser ce type d'approche. Nous proposons donc de simplifi er la projection T.

Cette simplifi cation consiste à supposer que tous les sommets 3D sont dans un même plan. Cette hypothèse est réaliste lorsque la distance entre la caméra et le visage est grande en comparaison des différences de profondeurs entre les points 3D du visage. La projection T est donc simplifi ée en une projection perspective faible[1] dont la matrice associée est de dimension 2×4 tel que :

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} = \begin{pmatrix} (\alpha_u.r_{1j} + u_0.r_{3j})/t_z & \alpha_u.t_x/t_z + u_0 \\ (\alpha_v.r_{2j} + v_0.r_{3j})/t_z & \alpha_v.t_y/t_z + v_0 \end{pmatrix}_{j\in[1,3]} . \begin{pmatrix} M_i \\ 1 \end{pmatrix} = \underbrace{\begin{pmatrix} a_0 & b_0 & c_0 & d_0 \\ a_1 & b_1 & c_1 & d_1 \end{pmatrix}}_{T_{2\times4}} . \begin{pmatrix} M_i \\ 1 \end{pmatrix}.$$

En annulant, à partir de l'équation 2, chacune des dérivées partielles de E fonction des paramètres de T, nous obtenons deux systèmes linéaires (σ et α sont positionnés à zéro). Le premier système linéaire est donné Équation 4 (le deuxième système linéaire est obtenu en remplaçant u_i par v_i et a_0 par a_1, b_0 par b_1 et ainsi de suite). Les deux systèmes sont résolus en utilisant des outils classiques d'algèbre linéaire.

$$\begin{pmatrix} \sum_i X_i^2 & \sum_i X_i.Y_i & \sum_i X_i.Z_i & \sum_i X_i \\ \sum_i X_i.Y_i & \sum_i Y_i^2 & \sum_i Y_i.Z_i & \sum_i Y_i \\ \sum_i X_i.Z_i & \sum_i Y_i.Z_i & \sum_i Z_i^2 & \sum_i Z_i \\ \sum_i X_i & \sum_i Y_i & \sum_i Z_i & \sum_i 1 \end{pmatrix} . \begin{pmatrix} a_0 \\ b_0 \\ c_0 \\ d_0 \end{pmatrix} = \begin{pmatrix} \sum_i X_i.u_i \\ \sum_i Y_i.u_i \\ \sum_i Z_i.u_i \\ \sum_i u_i \end{pmatrix}. \quad (4)$$

3.2 Approximation de la forme

Une fois que la projection $T_{2\times4}$ est calculée, nous pouvons extraire le modèle 3D du visage. Pour cela, nous allons "*mettre-en-forme*" le modèle 3D moyen *CANDIDE-3*. La minimisation de l'équation 2 est effectuée en fi xant la projection $T_{2\times4}$. L'équation 2 est réécrite telle que :

[1] La projection perspective faible est également appelée projection orthographique, projection orthographique graduée ou projection affi ne.

$$E = \sum_i [U_i - N.S_i.\sigma]^t.[U_i - N.S_i.\sigma], \quad (5)$$
$$avec\, U_i = \begin{pmatrix} u'_i \\ v'_i \end{pmatrix} - T.\begin{pmatrix} M_i \\ 1 \end{pmatrix},$$
$$et \quad N = T_{2\times 3}.$$

On obtient alors un système linéaire (équation 6) en annulant les dérivées partielles $\frac{\partial E}{\partial \sigma}$. La solution est telle que $\sigma = (A^tA)^{-1}A^t.B$.

$$\underbrace{(\sum_i S_i^t.N^t.N.S_i)}_{A}.\sigma = \underbrace{\sum_i S_i^t.N^t.U_i}_{B}. \quad (6)$$

Du fait du faible nombre de couples de points 2D-3D en correspondance, la matrice A est souvent non inversible. On la re-conditionne donc de manière à ce que les coeffi cients diagonaux à l'intersection d'une ligne de zéros et d'une colonne de zéros soient non nuls. Notons également que la solution du système linéaire ne doit pas mener à des retournements de maille 3D ni à des déplacements de grande amplitude par rapport à la taille du modèle 3D. Pour cela, on force les coeffi cients de σ à appartenir à leur intervalle de défi nition $[-1, 1]$. Enfi n, pour plus de robustesse, seuls les points peu animés ou non animés sont considérés pour la construction de la matrice A et du vecteur B. Le même raisonnement peut être tenu pour effectuer le calcul de α.

4 Raffinement du modèle 3D et de la pose 3D

Dans la partie précédente, nous avons expliqué comment obtenir rapidement une pose 3D approximative (la matrice $T_{2\times 4}$) et un modèle 3D approximatif. Notre objectif est maintenant d'extraire une information de position 3D plus descriptive (en particulier l'information de profondeur : t_z) et d'avoir un modèle 3D plus précis (c'est-à-dire avoir une meilleure mise en correspondance entre points 2D et points 3D projetés).

4.1 Extraction de la position 3D et raffinement du modèle 3D

Pour extraire les paramètres extrinsèques (la rotation et la translation), nous utilisons l'algorithme *POSIT*[2] de DeMenthon [6] en fi xant les paramètres intrinsèques[3]. Ce choix est guidé par le fait que les paramètres intrinsèques peuvent être approximés sans grande erreur de reconstruction [7, 8] et parceque l'algorithme *POSIT* est simple et rapide.

Des erreurs de mise en correspondance sont introduites : **1.** lors de l'approximation de T (sous-partie 3.1), **2.** lors de la déformation du modèle 3D par l'utilisation d'unités de forme et d'animation (sous-partie 3.2), **3.** lors du positionnement des paramètres intrinsèques. Pour obtenir une mise en correspondance précise nous déplaçons séparément chaque sommet 3D du modèle et pour plus de robustesse, nous déplaçons uniquement les sommets 3D dont l'erreur de mise en correspondance est faible. Le déplacement de chaque sommet M_i est obtenu en résolvant l'équation linéaire 7 (X_i et Y_i sont les inconnus) pour les couples de points valides. Remarquons que cette dernière étape doit être mené uniquement si l'erreur moyenne de mise en correspondance est faible.

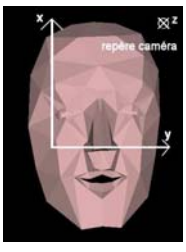
$$\begin{pmatrix} ((u_i-u_0).r_{31}-\alpha_u.r_{11}) & ((u_i-u_0).r_{32}-\alpha_u.r_{12}) \\ ((v_i-v_0).r_{31}-\alpha_v.r_{21}) & ((v_i-v_0).r_{32}-\alpha_v.r_{22}) \end{pmatrix}.\begin{pmatrix} X_i \\ Y_i \end{pmatrix}$$
$$= \begin{pmatrix} (\alpha_u.r_{13}-(u_i-u_0).r_{33}).Z_i + \alpha_u.t_x - (u_i-u_0).t_z \\ (\alpha_v.r_{23}-(v_i-v_0).r_{33}).Z_i + \alpha_v.t_y - (v_i-v_0).t_z \end{pmatrix}. \quad (7)$$

5 Résultats

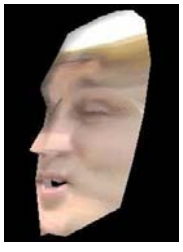
La Figure 2 illustre sur l'image *Foreman* quelques étapes de notre technique d'extraction et de positionnement d'un modèle 3D. La fi gure 2(a) donne le résultat de la "*mise-en-forme*" du modèle 3D moyen *CANDIDE-3* ($T_{2\times 4}$ et σ sont calculés). La fi gure 2(b) montre le maillage fi nal du modèle 3D de visage projeté. On peut constater que le modèle représente très bien la morphologie et l'animation du visage. Les fi gures 2(c) et 2(d) montrent que l'utilisation d'un modèle 3D pour le suivi de visage est envisageable puisqu'il est possible de synthétiser de manière réaliste le visage sous un angle de vue différent, ceci avec un faible nombre de points en entrée, un modèle 3D possédant peu de facettes, et un temps de calcul très faible.

(a) Approximation de la forme.
[Points 2D (en gris)]
[Points 3D projetés (en noir)]

(b) Maillage du modèle 3D



(c) Modèle 3D vu par la caméra



(d) Modèle 3D avec texture

FIG. 2 – Illustration de certaines étapes de l'extraction et du positionnement du modèle 3D sur l'image *Foreman*

Le tableau 1 donne les erreurs moyennes de mise en correspondance entre les points 2D placés manuellement et les points 3D projetés. L'erreur est donnée pour le modèle projectif perspective faible : $T_{2\times 4}$ et pour le modèle projectif sténopé : $T_{3\times 4}$ (paramètres intrinsèques fi xés ; paramètres extrinsèques calculés par l'algorithme *POSIT*). On peut constater qu'en règle gé-

[2] L'algorithme *POSIT* consiste à réduire de manière itérative l'erreur de projection due à l'approximation de la projection de l'équation 1 par une projection perspective faible, ceci dans l'objectif d'extraire les paramètres extrinsèques.

[3] f est positionnée à 0.05, k_u et k_v sont positionnés à 5000, le point $(u_0, v_0)^t$ est positionné au centre de l'image

nérale le modèle projectif faible est plus précis que le modèle sténopé.

TAB. 1 – Erreurs moyennes de mise en correspondance avec la projection perspective faible et la projection sténopée

Image	Nb. points	Type de projection	Erreur
Foreman	27	$T_{2\times4}$	3.09
		$T_{3\times4}$	2.88
Lena	23	$T_{2\times4}$	2.78
		$T_{3\times4}$	3.15
Rotation (base M2VTS)	15	$T_{2\times4}$	1.62
		$T_{3\times4}$	5.22

La figure 3 illustre le cas où l'erreur de mise en correspondance est réduite lors du passage au modèle sténopé T_{3x4}. La figure 4 illustre le cas inverse avec une erreur de mise en correspondance fortement augmentée lors du passage au modèle sténopé.

L'erreur obtenue par le modèle sténopé peut s'expliquer par l'imprécision sur les paramètres intrinsèques et extrinsèques. Il pourrait donc être envisagé d'ajouter à l'approche le calcul des paramètres intrinsèques : soit en utilisant l'auto-calibration basée analyse-synthèse [9], soit en utilisant des approches provenant du domaine de la vision [10]. Le calcul des paramètres intrinsèques par de telles approches reste cependant sensible à la phase d'initialisation ainsi qu'aux bruits sur les points 2D. Le calcul des paramètres extrinsèque peut également être amélioré. L'algorithme *POSIT* fournit en effet une approximation des paramètres extrinsèques et une approche linéaire robuste [11] pourrait s'y substituer.

(c) Projection $T_{2\times4}$

(d) Projection $T_{3\times4}$

FIG. 3 – Illustration de la projection du maillage 3D, sur l'image *Foreman*, avec un modèle projectif à perspective faible $T_{2\times4}$ et avec un modèle projectif sténopé $T_{3\times4}$

Le temps d'exécution nécessaire à l'extraction du modèle 3D et de sa pose, à partir d'une trentaine de points 2D, sur un processeur Intel Pentium cadencé à 2.4Ghz est de 1.6 ms (l'implémentation est en C++ et lors de la mesure, une dizaine d'autres applications fonctionnaient). Ce temps de calcul peut par ailleurs être réduit par optimisation du code et utilisation d'une librairie de calcul matriciel plus performante.

6 Conclusion

Dans ce papier nous proposons une solution robuste et rapide pour positionner et extraire un modèle de 3D de visage. Dans ce but, nous utilisons un modèle 3D moyen de visage et quelques couples de points 2D-3D en correspondance. Une succession de calculs robustes et rapides mènent à un modèle 3D reflétant bien la structure du visage traité, et à un positionnement du modèle 3D très complet puisque l'on détermine les paramètres du modèle projectif de caméra sténopé (paramètres de caméra, de rotation et de translation).

(a) Projection $T_{2\times4}$

(b) Projection $T_{3\times4}$

FIG. 4 – Illustration de la projection du maillage 3D, sur une image d'une séquence de la base M2VTS, avec un modèle projectif à perspective faible $T_{2\times4}$ et avec un modèle projectif sténopé $T_{3\times4}$

Références

[1] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.

[2] F. Dornaika and J. Ahlberg, "Fast and reliable active appearance model search for 3D face tracking," *Transactions on Systems, Man, and Cybernetics–Part B : Cybernetics*, vol. 34, no. 4, pp. 1838–1853, 2004.

[3] R. Horaud et O. Monga, *Vision par ordinateur : Outils fondamentaux*, chapter 5, Editions Hermès, Eyrolles, 1995.

[4] J. Ahlberg, "CANDIDE-3 - un updated parameterised face," Tech. Rep., Department of Electrical Engineering, Linköping University, Jan. 2001.

[5] O. D. Faugeras and G. Toscani, "The calibration problem for stereo," in *Computer Vision and Pattern Recognition, CVPR*, 1986 June, pp. 15–20.

[6] D. DeMenthon and L.S. Davis, "Model-based object pose in 25 lines of code," *International Journal of Computer Vision, IJCV*, vol. 15, no. 1, pp. 123–141, June 1995.

[7] S. Bougnoux, "From projective to euclidean space under any practical situation, a criticism of self-calibration," in *International Conference on Computer Vision, ICCV*, Jan. 1998, pp. 790–798.

[8] L.F. Cheong and C-H. Peh, "Characterizing depth distortion due to calibration uncertainty," in *European Conference on Computer Vision, ECCV*. June 2000, vol. 1842, pp. 664–677, Springer-Verlag.

[9] P. Eisert, "Model-based camera calibration using analysis by synthesis techniques," in *Proc. 7th International Workshop Vision, Modeling, and Visualization*, Nov. 2002, pp. 307–314.

[10] A. Fusiello, "Uncalibrated euclidean reconstruction : A review," *Image and Vision Computing*, vol. 18, no. 6-7, pp. 555–563, May 2000.

[11] L Quan and Z Lan, "Linear n-point camera pose determination," *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI*, vol. 21, no. 8, pp. 774–780, Aug. 1999.

Perception visuelle du geste de préhension.

Pablo Augusto Negri, Xavier Clady, Maurice Milgram
LISIF - PARC, UMPC (Paris 6)
3 rue Galilee 94200 Ivry-sur-Seine, FRANCE
pablo.negri@lisif.jussieu.fr,
clady@ccr.jussieu.fr,
maum@ccr.jussieu.fr

Résumé – Nous présentons dans cet article, un nouvel algorithme de traitement d'images, la "Transformée Chinoise", permettant d'estimer la localisation des doigts d'une main. Cette approche utilise une technique inspirée de la Transformée de Hough qui prend en compte la disposition des pixels de contour ainsi que l'orientation du gradient en ces pixels. Elle a été intégrée dans un système d'acquisition visuelle monoculaire des gestes humains de préhension.

Abstract – This article describe a new algorithm, the Chinese Transform, for the localization of the fingers. This approach is inspired in the Hough Transform utilizing the position and the orientation of the gradient from the image edge's pixels. A vision system using this technique is proposed, requiring only one camera and a computer.

1 Introduction

Dans la taxonomie des gestes [9], ceux de préhension se classent dans la catégorie des gestes techniques. Ils font l'objet de nombreuses études dans les communautés cognitivistes et médicales, dont [4, 5] font office de travaux fondateurs. Ces études sont souvent réalisées dans l'objectif de déterminer les influences de maladies motrices ou psychomotrices (Parkinson [1], lésions cérébrales [3], etc.) sur la coordination du geste de préhension. Pour la plupart, les expérimentations consistent à disposer sur un plan plusieurs objets, généralement cylindriques, de différentes tailles et à différentes positions. Les cobayes doivent prendre les objets suivant un protocole défini ; tout en gardant la main parallèle au plan. Des émetteurs infrarouges sont installés sur le pouce et l'index, ainsi que sur la paume. Un système de vision dédié, par exemple l'Optotrack, traque ces marqueurs.

Dans ce papier, nous allons proposer un système non intrusif et peu onéreux, ne nécessitant qu'une seule caméra numérique grand public. D'autres applications seraient possibles telles qu'une Interface Homme-Machine naturelle [14] pour la téléopération de robots manipulateurs [13] ou des jeux en Réalité Virtuelle (par exemple, un jeu d'échecs où les pièces seraient virtuelles).

Nous disposons d'une mire, dite Plateforme d'Evolution (PE), composée d'amers circulaires de couleurs différentes et de géométrie connue. La main, formant une pince avec le pouce et l'index, se déplace dans un plan horizontal (parallèle à la plateforme) sans variation de la hauteur (composante Z, estimée à 5cm). Une caméra fixe est placée en contre-plongée, à une distance relativement grande par rapport à la PE, de sorte que la variabilité de profondeur des points de la main puisse être considérée comme négligeable. La position 3D des doigts est calculée à partir de ces hypothèses et en utilisant aussi une méthode itérative d'estimation de la pose de la caméra (cf. fig. 1).

Dans la prochaine section, nous décrivons les procédures

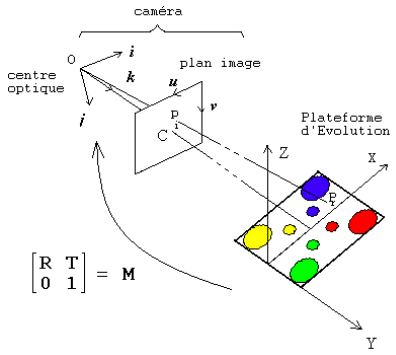


FIG. 1 – Relation "scène-caméra".

pour extraire les positions des doigts dans l'image. Tout d'abord, un algorithme de soustraction de fond nous permet de déterminer une zone d'intérêt dans l'image. Dans cette zone, une image de distance à la couleur de peau est calculée pour l'extraction des bords orientés de la main. Ces bords sont employés dans un algorithme original, la Transformée Chinoise (TC), pour la segmentation et la localisation des doigts. Cette approche est inspirée de la Transformée de Hough. Cet algorithme permet l'extraction des segments des doigts. La section 3 concerne le suivi du geste. Nous utilisons un filtre de Kalman adapté au suivi des segments. Un modèle simplifié permet la reconstruction de la main à partir de ces segments. La section 4 présente quelques résultats obtenus avec notre système. Nous concluons cet article avec quelques perspectives.

2 Détection des doigts

2.1 Localisation de la scène et prétraitements

Les centres des amers colorés nous permettent d'obtenir la pose de la caméra dans la matrice homogène de

transformation M [8]. Cette matrice sert à reconstruire les positions de 3D de points de la main dans les repères liés à la PE, à partir de leurs positions 2D dans l'image. Il existe des solutions plus élégantes [12], permettant de résoudre notre problème sans calcul explicite de la pose ; cependant cette connaissance pourrait s'avérer utile dans des travaux futurs. Dans une évolution future de notre système, nous pourrions ajouter une seconde caméra, liée mécaniquement à la première, qui observerait le visage de l'opérateur, afin de déterminer son champ de vision de la scène. D'où l'importance de localiser la plateforme par rapport au bâti instrumenté des deux caméras.

Pour restreindre la recherche à une fenêtre dans l'image, avec pour conséquence un gain de calcul, nous utilisons une méthode de suppression du fond. La méthode dite de Stauffer et Grimson [11, 2] utilise des mixtures de Gaussiennes pour modéliser la couleur des pixels. Dans notre approche, nous n'utilisons pas l'espace RGB, mais uniquement les chrominances de l'espace couleur YC_bC_r. Cela nous permet de minimiser l'influence des variations de la composante Y (essentiellement dues aux ombres générées par la main ou l'environnement).

De la fenêtre obtenue, nous extrayons une image en niveaux de gris I_{tc}, où les pixels qui ont une chrominance proche à celle de la teinte chair seront nettement distingués. La méthode développée consiste à calculer une image de distance normalisée à la teinte chair dans l'espace des chrominances. L'image I transformé de l'espace RGB à l'espace YC_bC_r est appelée I_{ybr}. Ensuite, nous calculons la soustraction sur chaque composante, rouge (r), et bleue (b), du I_{ybr} avec les valeurs moyennes expérimentales de la teinte chair, b_{tc} and r_{tc} respectivement.

$$\begin{aligned} \overline{I_b} &= |I_{ybr}(b) - b_{tc}| \\ \overline{I_r} &= |I_{ybr}(r) - r_{tc}| \\ \overline{I_{br}} &= \sqrt{\overline{I_b} + \overline{I_r}} \end{aligned}$$

Enfin, nous obtenons l'image de distance à la teinte chair (cf. fig. 4) définie par :

$$I_{tc} = 1 - \frac{\overline{I_{br}}}{max(\overline{I_{br}})}$$

2.2 Transformation Chinoise

La Transformation Chinoise (TC) tient son nom du mot Shongguo, *l'Empire du Milieu* : c'est ainsi que les Chinois nomment leur pays. Il s'agit d'une méthode de votes : chaque vote est attribué au centre ou milieu de deux points ayant des directions de gradients opposées. Cette méthode a le même principe que le travail de Reisfeld [10].

Dans l'exemple de la fig. 2, deux points du contour e_1 et e_2, obtenus de l'image d'une ellipse I_e, sont définis à partir des paramètres suivants : le vecteur normal, n_i, et sa position dans l'image, $p_i(x, y)$, avec $i = \{1, 2\}$. Chaque vecteur normal représente l'orientation du gradient dans ce point. p_{12} est le segment que lie les deux points avec p_v son point milieu.

En superposant e_1 sur e_2, nous pouvons comparer ces orientations. Nous affirmons que e_1 et e_2 ont des orientations opposées si l'angle α_{12} formé entre n_1 et $-n_2$ accomplit la condition :

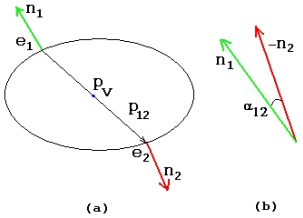


FIG. 2 – (a) montre deux points de l'image et ses vecteurs normals, (b) montre les deux vecteurs normals superposés qui forment un angle α.

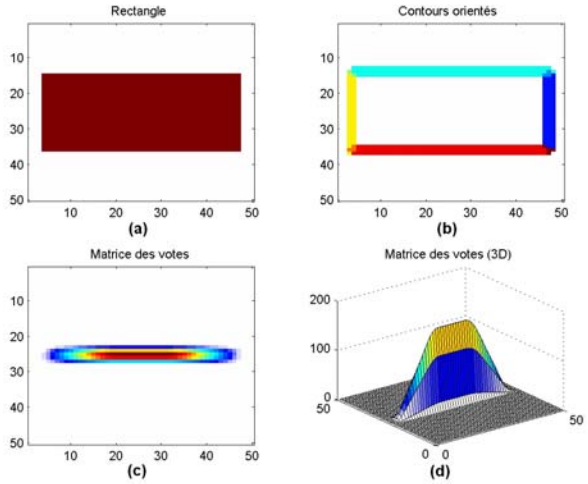


FIG. 3 – Exemple de la TC pour un rectangle. (a) image originale, (b) contours orientés, (c) et (d) matrices des votes en 2D et 3D.

$$\alpha_{12} < \alpha_{threshold} \quad (1)$$

Ensuite, la TC vote pour p_v, le point milieu de p_{12}, si :

$$|p_{12}| < d \quad (2)$$

Nous créons et incrémentons un accumulateur (matrice de votes) avec tous les couples qui accomplissent les conditions (1) et (2).

La fig. 3 est un exemple de l'algorithme de la TC appliqué à un rectangle. La fig. 3.a représente l'image originale et la fig. 3.b montre ces contours orientés en différentes couleurs. Dans la pratique, nous échantillons les orientations du gradient en $N = 8$ directions ; cette opération fixe une valeur pratique pour $\alpha_{threshold}$. La matrice de votes (voir la fig. 3.c et 3.d) est le résultat de l'application de la TC pour tous les points de contour de la fig. 3.a avec $d =$ 35 pixels.

Dans notre application, nous profitons de la forme tubulaire de l'index et du pouce (les deux doigts formant la pince). Leurs bords parallèles satisfont les conditions de direction, de distance et de gradient. Les zones d'accumulation trouvées permettent de définir les régions intérieures des doigts (cf. fig. 4.d). L'application d'une Transformée de Hough sur la matrice de votes permet de déterminer les segments correspondant aux doigts.

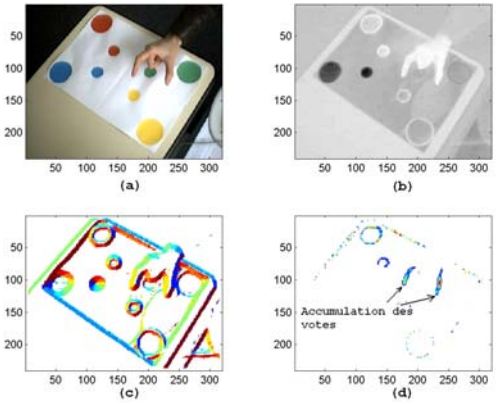


FIG. 4 – Illustration de la TC. (a) Image originale, (b) Distance à la couleur chair I_{tc}, (c) Orientation du gradient aux contour de l'image : chaque orientation est répresentée par une couleur différente, (d) Matrice d'accumulation de votes.

Les résultats de la TC peuvent se comparer à un algorithme de squelettisation morphologique. Cependant, la squelettisation utilise les régions comme primitive et est soumise à ses défauts caractéristiques : les trous, les lacunes et les irrégularités des bords.

3 Suivi du geste et répresentations

Un filtre de Kalman [6] adapté aux segments nous permet de filtrer les pistes, pour ne retenir que celles liées aux doigts formant la "pince" (constituée du pouce et de l'index), et d'éliminer les fauses alarmes.

3.1 Suivi des segments

L'objectif est le suivi des segments appartenant aux doigts dans une séquence d'images. Ceux-ci ont été obtenus à partir de la matrice de votes de la TC et de l'application de la Transformée de Hough. Les paramètres identifiant chaque segment sont (voir fig. 5) : $P_m(x_m, y_m)$, coordonnées du point milieu, l et θ, respectivement la longueur et l'angle du segment.

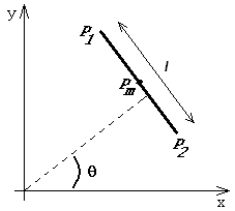


FIG. 5 – Modèle du segment.

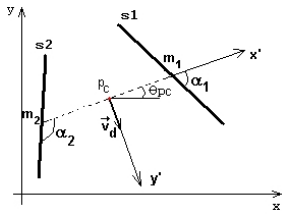


FIG. 6 – Modèle de la pince.

Notre système est composé de trois filtres de Kalman indépendants. Deux filtres scalaires pour la longueur et pour l'orientation, et un filtre vectoriel pour la position. Si nous considérons la vitesse constante, les vecteurs d'état sont :

$$X^{P_m} = \begin{pmatrix} x_m \\ \dot{x_m} \\ y_m \\ \dot{y_m} \end{pmatrix} \qquad X^{l} = \begin{pmatrix} l \\ \dot{l} \end{pmatrix} \qquad X^{\theta} = \begin{pmatrix} \theta \\ \dot{\theta} \end{pmatrix}$$

Nous suivons tous les segments de chaque image de la séquence. Il subsiste des fausses alarmes, qui vont disparaître dans les images successives.

3.2 Suivi de la pince

A partir de deux segments, nous modélisons la pince (cf. fig. 6) avec les paramètres suivants :
- p_c, point milieu du segment m_1m_2,
- Θ_{pc}, angle qui définie l'inclination de la pince par rapport à l'axe x,
- v_d, vecteur unitaire directeur qui définit l'orientation de la pince,
- l_{12}, longueur du segment m_1m_2,

Nous ajoutons deux autres paramètres : les orientations α_1 et α_2 des segments s_1 et s_2, calculées après un changement de repères, de (x, y) au (p_c, x', y'), l'axe lié à la pince, (voir fig. 6).

Les nouveaux vecteurs d'état pour le suivi de la pince sont : X^{p_c}, $X^{\Theta_{pc}}$, X^{α_1}, X^{α_2} et $X^{l_{12}}$.

Tous les couples de segments ne forment pas forcément une pince. Des contraintes sur la longueur l_{12} et les angles α_1 et α_2 nous permettent de filtrer les "fausses pinces". Cette représentation a été inspirée des études réalisées sur le geste de préhension [4, 5, 3, 1]. Elle permet d'observer facilement les principales grandeurs utilisées dans ces études : distance inter-doigts (ouverture de la pince), position et orientation de la main par rapport à la scène (et aux objets). Par ailleurs, elle nous a permis de définir un modèle articulatoire de la main.

4 Résultats

Nous appliquons la TC dans une séquence vidéo composée de trois étapes. La première étape montre une main allant saisir un objet imaginaire dans un coin de la PE (cf. la fig. 7). Dans la prochaine étape, le sujet "pose" l'objet imaginaire dans le coin opposé de la PE. L'étape finale montre la main retournant à la position initiale. Nous présentons les résultats obtenus pour l'étape 1.

Dans la fig. 8.a, nous pouvons voir tous les segments enregistrés durant la première étape. La fig. 8.b montre la trajectoire du point p_c. Les courbes d'évolution de l'ouverture des doigts et de la vitesse de la main pour l'étape 1 sont présentées respectivement dans la fig. 8.c et 8.d. D'après Jeannerod [4, 5], l'acte de préhension d'un sujet normal est divisé en 2 phases : une phase à grande vitesse correspondante au 75% du mouvement d'approche vers l'objet et une phase finale à plus faible vitesse. Dans la fig. 8.d, nous remarquons que la première phase jusqu'au frame 15 est caractérisée par une grande accélération et une augmentation de la distance inter-doigts (fig. 8.c). Ensuite, il y a une désaccélération de la main en s'approchant vers l'objet pour atteindre la distance inter-doigts finale.

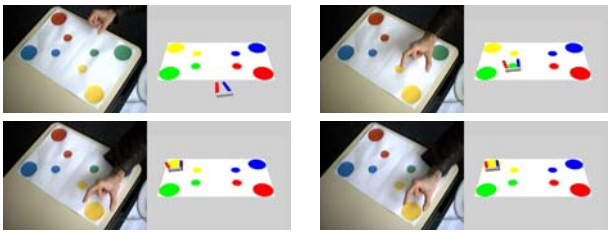


FIG. 7 – Cette figure montre des images de l'étape 1. Sur l'image de gauche, nous pouvons voir le point de vue de la caméra et sur la droite, l'environnement OpenGL avec une pince virtuelle reproduisant simultanément le geste.

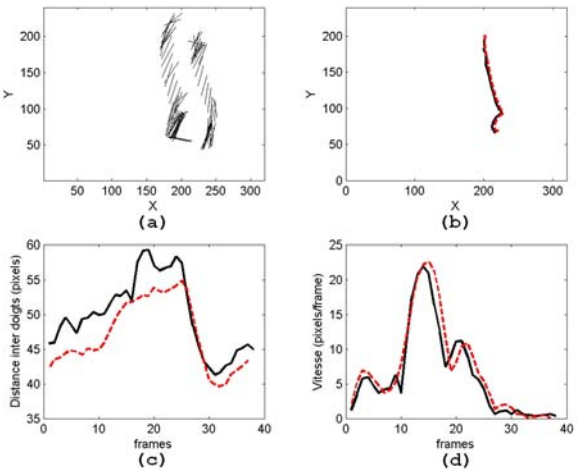


FIG. 8 – Résultats pour l'étape 1 : (a) tous les segments. (b) trajectoire du point p_c (c) distance entre les doigts (d) vitesse de la main. Dans (b), (c) et (d), les courbes en pointillés montrent la vérité terrain.

Ces informations peuvent être employées par les spécialistes afin de mesurer l'habilité du sujet et peuvent également être utiles pour un diagnostic médical.

5 Conclusion and perspectives

Cet article a présenté un système pour l'acquisition du geste humain de préhension. Il utilise une nouvelle méthode pour la détection et localisation des doigts, appelée Transformation Chinoise. Cette technique est une méthode de votes inspirée par la Transformée de Hough. Des filtres de Kalman sont utilisés pour le suivi du geste. Les résultats obtenus sont conformes aux observations des études médicales [4, 5] et peuvent être utilisés pour la détection des maladies psychomotrices.

Les prochaines étapes de nos travaux de recherche seront orientées dans la détermination, voire l'anticipation des points de prise d'un objet [7]. Pour ceci, nous devrons analyser le geste en fonction de ses propres caractéristiques, mais aussi des caractéristiques intrinsèques (forme, taille, etc.) et extrinsèques (position, orientation) de l'objet.

Références

[1] U. Castiello, K. Bennet, C. Bonfiglioli, S. Lim, and R.F. Peppard. The reach-to-grap movement in parkinson's disease : response to a simultaneous perturbation of object position and object size. *Computer Exp. Brain Res*, (125) :453–462, 1999.

[2] E. Hayman and J. Eklundh. Statistical background subtraction for a mobile observer. In *In the Proceedings of the 9th International Conference on Computer Vision*, pages 67–74, Nice, France, 2003.

[3] J. Hermdörfer, and Marquardt C. Ulrich, S., and and Mai N. Goldenberg, G. Prehension with the ipsilesional hand after unilateral brain damage. *Cortex*, pages 35 :139–161, 1999.

[4] M. Jeannerod. Intersegmental coordination during reaching at natural visual objects. *Attention and performance (Long J, Baddeley A, eds)*, pages 153–168, 1981.

[5] M. Jeannerod. The timing of natural prehension movements. *Journal of Motor Behavior*, pages 16 :235–254, 1984.

[6] R.E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, (82) :35–45, 1960.

[7] C. Michel, V. Perdereau, and M. Drouin. An approach to grasp planning in constrained environments. In *Proceedings of the IEEE International Conference on Mechatronics and Robotics*, pages 761–767, Aachen, Germany, Septembre 2004.

[8] P. Negri, X. Clady, and Milgram M. Perception visuelle du geste de préhension : Application à la robotique manipulatrice. In *18ème Journée des JJCR*, Douai, France, Septembre 2004.

[9] V. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction : A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19 :677–695, July 1997.

[10] D. Reisfeld. *Generalized Symmetry Transforms : Attentional Mechanisms and Face Recognition.* PhD thesis, Tel Aviv University, Janvier 1994.

[11] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition*, volume 2, pages 22–46, Fort Collins, Colorado, Juin 1999.

[12] Peter Sturm. Algorithms for plane-based pose estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition, Hilton Head Island, South Carolina, USA*, pages 1010–1017, June 2000.

[13] J. Triesh and C. von der Malsburg. Classification of hand postures agains complex backgrounds using elastic graph matching. *Image and Vision Computing*, 20(13-14) :937–943, Décembre 2002.

[14] M. Turk and M. Kolsch. *Emerging Topics in Computer Vision*, chapter Perceptual Interfaces. Prentice Hall PTR, 2005.

Détection de piétons par stéréovision et noyaux de graphes

Frédéric SUARD, Alain RAKOTOMAMONJY, Abdelaziz BENSRHAIR

Laboratoire PSI (Perception, Systèmes d'Information), CNRS FRE 2645
INSA de Rouen, FRANCE
Tél : 02 32 95 98 84 Fax : 02 32 95 97 08
frederic.suard@insa-rouen.fr

Résumé – Cet article présente une méthode concernant la reconnaissance de piétons à l'aide de graphes et de méthodes à noyaux. La détection du piéton est limitée à cause de la grande variabilité de la forme du piéton : taille, posture. Pour surmonter ce problème, nous avons choisi de le représenter à l'aide d'un graphe. Le but de la méthode est d'extraire le graphe de chaque objet (piétons ou non-piétons) présent dans une base d'images et de calculer un noyau à partir de ces graphes afin d'effectuer un apprentissage supervisé basé sur les SVMs (Séparateurs à Vaste Marge). L'application sur une base d'images réelles nous permet de démontrer l'efficacité de cette méthode, au niveau des invariances en échelle, avec un bon taux de reconnaissance.

Abstract – This article presents a novel method concerning pedestrian detection, thanks to graph kernels. Nowadays, the pedestrian detection is a hard task, due to the variability of its shape : size and posture. To address this problem, we choose to transform a pedestrian into a graph. The aim of this method consists of extracting a graph from each object (pedestrian or non-pedestrian), contained in a database. We compute the kernel with the inner product between graphs in order to apply a supervised classifier, here the SVMs (Support Vector Machine). We applied this method on a real images database in order to test its efficiency, particularly for scale invariance, and we obtained a good classification rate.

1 INTRODUCTION

Depuis quelques années le problème de la détection de piétons a inspiré de nombreux travaux dans le domaine du traitement d'images et de la reconnaissance de formes. Malgré diverses approches pour résoudre ce problème, aucune n'a réellement apporté de solutions définitives, car, d'après Shashua [9], la détection est confrontée à la très grande variabilité dans la forme du piéton.

Aujourd'hui, la tendance serait d'associer des méthodes de vision artificielle et des méthodes d'apprentissage telles que les réseaux de neurones [12], l'Adaboost [11], les moindres carrés régularisés [9] et les Séparateurs à Vaste Marge(SVM) [6].

La méthode présentée suit cette association : stéréovision et méthode à noyaux ([3]). Dans un premier temps, nous utilisons un système de stéréovision afin d'extraire les images des objets appartenant à la scène filmée. Il est ainsi possible de traiter le problème de l'occlusion à l'aide de la 3D. Nous représentons ensuite chaque objet à l'aide d'un graphe, apportant ainsi une réponse au problème de variabilité et d'invariance en échelle. La classification est basée sur les SVMs, dont le noyau est un noyau de graphes, basé sur le produit scalaire de graphes, obtenu en parcourant aléatoirement chaque graphe.

Dans cet article, nous reviendrons sur la construction des graphes à partir d'une paire d'images stéréo, puis nous décrirons rapidement le classifieur utilisé. Enfin, nous présenterons quelques résultats de cette méthode appliquée à des images réelles.

2 DESCRIPTION DE LA METHODE

Nous allons maintenant décrire rapidement la méthodologie que nous avons suivie. Dans un premier temps, il faut extraire les objets de la scène à partir des images, puis appliquer le processus de reconnaissance de formes.

2.1 Extraction des objets

L'idée originale de la méthode est de décrire un piéton à l'aide d'un graphe. En effet, un piéton est composé de plusieurs parties : une tête, un ou deux bras, un buste, une ou deux jambes. Le graphe correspond donc grossièrement au squelette, avec les arcs désignant les membres et les noeuds les articulations entre eux. La transformation d'un piéton en graphe rend ainsi la détection moins sensible aux différentes postures du piéton. Dans notre cas, un piéton de profil et un piéton de face auront quasiment le même squelette, ce qui nécessite par la suite un apprentissage moins volumineux mais produisant des résultats similaires par rapport à d'autres méthodes qui sont plus sensibles à la forme du piéton observé. Les graphes possèdent quelques propriétés que nous nous proposons d'exploiter. Un graphe permet en effet de décrire la forme générale de l'objet en conservant ses caractéristiques géométriques. Un graphe est plus facilement manipulable qu'une image et prend moins de place en mémoire, ce qui est intéressant dans le cas de grandes bases d'apprentissage. Il est également possible, au travers des étiquettes du graphe, de choisir les meilleures informations à retenir pour caractériser l'objet. Enfin, la comparaison de deux graphes est assez rapide pour un faible nombre de nœuds.

Avant de transformer le piéton en graphe, il convient donc d'extraire l'image du piéton de l'image globale. Pour cela, une segmentation est effectuée selon deux critères : une segmentation région et une segmentation 3D. Cette double segmentation est effectuée dans le but d'améliorer la qualité des images des objets extraits. La figure 1 montre un exemple de piéton po-

sant des problèmes pour la reconnaissance. Si le piéton est mal segmenté, le graphe devient alors de mauvaise qualité et la classification devient difficile.

(a) (b) (c)

FIG. 1: *Exemple de piéton mal segmenté.*

Une des deux images stéréo est tout d'abord segmentée par régions selon une méthode simple : nous regroupons les pixels voisins dans une même région dont le niveau de gris est proche. Cependant, cette segmentation ne permet pas de récupérer l'intégralité des objets. Dans le but d'étudier également le problème d'occlusion, nous utilisons la deuxième image stéréo afin de calculer les disparités de chaque région obtenues dans la première image. Ce calcul permet ainsi de regrouper les régions voisines dont la disparité est proche, c'est à dire qui appartiennent au même plan de l'image, et de différencier les régions voisines n'appartenant pas au même plan. La disparité de chaque région est calculée selon un algorithme de mise en correspondance classique : on recherche dans l'autre image la région correspondant le mieux à la région considérée [7]. Cette mise en correspondance est efficace car elle compare des régions et non des points, mais elle se révèle au final beaucoup plus coûteuse en temps de calcul. Lorsque la disparité de toutes les régions a été calculée, nous procédons à une fusion de régions. Les régions voisines dont la disparité est proche sont ainsi regroupées afin de reconstituer l'intégralité des objets. La figure 2 montre un exemple de traitement à partir d'un paire d'images stéréoscopiques (a) et (b), le résultat de la segmentation (c) région sur l'image droite, et l'image obtenue après le calcul de disparité (d). Les régions les plus claires sont les plus proches de la caméra.

Chaque région dont la disparité est non nulle est ensuite extraite et transformée en graphe.

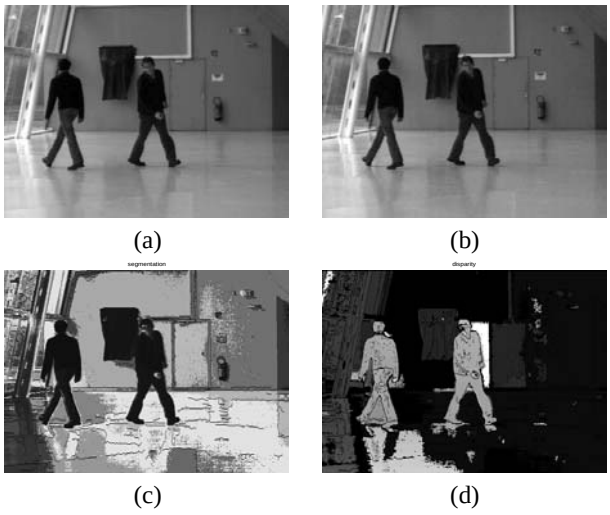
Il faut tout d'abord binariser l'image du piéton avant d'extraire le squelette de cette forme. Le squelette possède les propriétés suivantes

- son épaisseur est de 1 pixel sur tout le squelette, sauf aux intersections,
- la géométrie et la topologie de l'objet sont conservées,
- le squelette est situé au milieu de l'objet.

Il est obtenu de la façon suivante ([8]): Le squelette d'un ensemble X selon une famille d'éléments structurant δ_n est le lieu géométrique des centre de tous les éléments maximaux.

Le squelette contient donc l'information de la forme générale de l'objet, exploitable pour la reconnaissance. Il faut ensuite transformer le squelette en graphe. Un graphe $G = (N, A)$ est constitué d'un ensemble de nœuds N, reliés entre eux par un ensemble d'arcs A. Un nœud est défini de la façon suivante : c'est un pixel du squelette placé à l'extrémité d'un arc. C'est donc soit l'extrémité d'une branche du squelette, soit un pixel placé à une intersection de plusieurs branches.

Lorsque tous les nœuds ont été trouvés, il faut ensuite les relier entre eux afin de créer les arcs. Un arc est présent entre 2 nœuds si il existe un chemin de pixels appartenant au squelettre entre les 2 nœuds, sans que ce chemin ne contienne d'autres nœuds. La figure 4 illustre la transformation d'une image en graphe.



(a) (b) (c) (d)

FIG. 2: *Image originale gauche (a) et droite (b). Segmentation région de l'image droite (c). Disparité (d) obtenue à partir des images (a) et (b).*

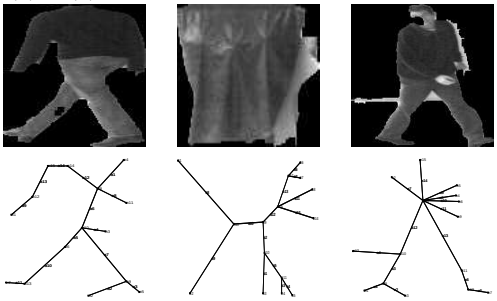


FIG. 3: *Exemples d'images d'objets extraits de la scène de la figure 2 et leur graphe correspondant.*

(a) (b) (c)

FIG. 4: *Image originale (a), son squelette (b) et le graphe résultat (c)*

L'étape suivante consiste à étiqueter le graphe, c'est à dire à ajouter de l'information relative à l'objet au niveau des arcs et des nœuds. Pour les arcs, nous avons choisi les caractéristiques suivantes :

- la longueur de l'arc,
- l'orientation de l'arc par rapport à l'orientation de l'axe principal de l'ellipse contenant l'intégralité de l'objet,
- l'aire du voisinage de l'arc, c'est à dire l'ensemble des pixels décrits par la branche du squelette représentée par

l'arc.

Pour les nœuds, nous avons retenu quelques caractéristiques :

- les coordonnées du nœud dans l'image,
- la taille du plus grand élément structurant contenu dans l'image binaire à l'emplacement du nœud.

2.2 Reconnaissance de formes

La reconnaissance de formes est basée sur une méthode d'apprentissage supervisée. Dans notre cas, les SVMs basés sur un noyau de graphes. Les SVMs étant des classifieurs binaires [2, 1, 10], ils correspondent parfaitement au problème posé en indiquant l'appartenance à la classe 'piéton' ou non. La fonction de décision est calculée selon un ensemble d'apprentissage contenant des graphes de piétons et des graphes de non-piétons.

Le noyau de graphe est défini comme étant le produit scalaire entre graphes : $K(G_i, G_j) = \langle G_i, G_j \rangle$. Le produit scalaire de deux graphes est obtenu en comparant les chemins présents dans chaque graphe en tenant compte des étiquettes d'arcs et de nœuds présents sur le chemin. Pour calculer des produits scalaires entre ces représentations graphiques, nous nous sommes basés sur l'article de H. Kashima [4].

Les graphes sont représentés par plusieurs matrices, chacune définissant les étiquettes des nœuds et des arcs. Un graphe comporte autant de matrices que de catégorie d'étiquettes. Le calcul du noyau aboutit à la résolution d'un système linéaire de taille $(|G^1||G^2|)^2$, avec $|G^i|$ le nombre de nœuds du graphe i. Cette complexité étant assez important, elle nous a forcé à réduire au maximum le nombre de nœuds dans chacun des graphes. Etant donné qu'un graphe est censé représenter les articulations d'un piéton, d'après David Moore[5], 15 nœuds sont suffisants pour représenter un piéton. Nous avons donc développé une fonction permettant de réduire le nombre de nœuds d'un graphe en conservant les informations contenues dans les nœuds et les arcs afin de travailler avec des graphes comportant au maximum 15 nœuds.

3 RESULTATS

Nous allons maintenant détailler les résultats obtenus. Nous avons réalisé plusieurs acquisitions vidéos stéréoscopiques, en intérieur, avec un éclairage constant, et une dizaine de piétons différents. Chaque image acquise comporte au maximum 4 piétons.

Nous avons ainsi obtenu environ 300 images de piétons et 500 de non-piétons (mur, poteau, escalier, etc.)

Chaque image a été transformée en graphe, et nous avons sélectionné plusieurs type d'étiquettes différentes :

- les coordonnés du nœud,
- la longueur des arcs,
- les paramètres des droites support des arcs.

Voici le mode opératoire pour tester la méthode. Nous prenons successivement un certain nombre de données d'apprentissage : 1, 5, 10, 25, 50 et 100, piétons et non-piétons tirés aléatoirement dans la base d'exemple. Pour chaque apprentissage, nous testons plusieurs paramètres pour le noyau et le classifieur SVM. Chaque procédé est itéré 10 fois afin d'obtenir des résultats robustes vis-à-vis du tirage aléatoire.

Lors du calcul du noyau, nous avons le produit $K_n(N_1, N_2)$, résultat de la comparaison des nœuds selon un noyau gaussien de largeur de bande σ_i :

$K_n(N_1, N_2) = e^{-\Sigma_{i=1}^{m} \frac{N_1^i - N_2^i}{2\sigma_i^2}}$, avec N_j^i l'étiquette j du nœud i, m le nombre de caractéristiques de chaque nœud. Nous testons donc plusieurs valeurs pour σ_i afin de retenir le meilleur paramètre.

Au niveau du classifieur, un paramètre C est utilisé pour pondérer l'importance des graphes mal classés. Nous testons également différentes valeurs pour ce paramètre de 1 à 1000.

Le résultat est affiché sur la figure 5. Comme nous pouvons le voir, lorsque C est trop faible, les graphes mal classés n'ont pas beaucoup de poids, et la frontière de décision est mal choisie d'où un taux de classification peu élevé. A partir d'une certaine valeur de C, le taux de classification se stabilise, l'augmentation de C ne modifie plus le taux de classification.

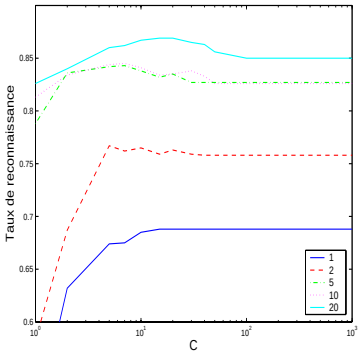


FIG. 5: *Taux de classification en fonction du paramètre C obtenus avec 3 étiquettes (a) : coordonnées(x,y) du nœud, longueur de l'arc. Chaque courbe est le résultat obtenu pour un nombre de graphes en apprentissage différent : 1,2,5,10,20.*

Nous avons également souhaité connaître l'influence des étiquettes sur les résultats. Dans un premier temps, le test est effectué sans prendre en compte les coordonnées de la droite pour les étiquettes d'arcs. Puis le test est refait en tenant compte de toutes les étiquettes citées ci-dessus. La figure 6 montre les résultats obtenus pour 3 étiquettes et 5 étiquettes, en fonction du nombre de graphes par classe utilisées lors de l'apprentissage. Nous affichons également la courbe AUC : aire sous la courbe ROC. La courbe ROC permet de connaître le lien entre le taux de vrais positifs et le taux de faux positifs. Plus l'aire est importante, plus le rapport est correct, c'est à dire que nous obtenons très peu de faux positifs pour un certain nombre de vrais positifs. Les résultats sont légèrement meilleurs en rajoutant des étiquettes, mais le gain n'est pas significatif. Les étiquettes ajoutées sont donc redondantes et ne permettent pas de renforcer les étiquettes initiales, nous allons donc continuer à tester d'autres étiquettes. Sur la figure 6, nous constatons ainsi que la courbe (a) dont les graphes possèdent 3 types d'étiquettes donne des résultats légèrement inférieurs à ceux de la courbe (b), dont les graphes possèdent 5 types d'étiquettes. Nous constatons également qu'avec peu de graphes utilisés lors de l'apprentissage, le taux de classification est satisfaisant. Nous en déduisons donc que quelques graphes de piétons permettent de caractériser correctement un ensemble important de piétons.

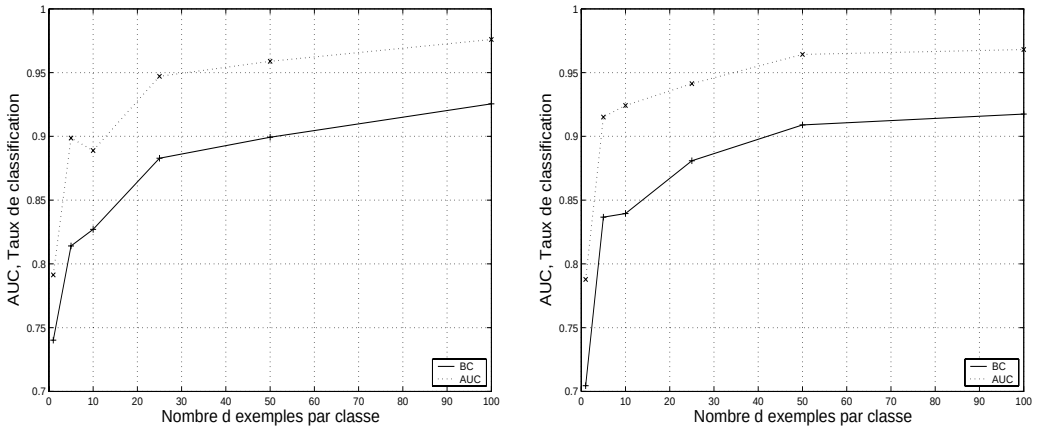


FIG. 6: *Taux de classification (BC) et Aire Sous la Courbe ROC (AUC) en fonction du nombre de graphes par classe pour l'apprentissage obtenus avec 3 étiquettes (a) : coordonnées(x,y) du nœud, longueur de l'arc. Taux et AUC pour 5 étiquettes (b) : coordonnées(x,y) du nœud, longueur de l'arc et paramètres de la droite support de l'arc.*

4 CONCLUSION

Nous avons présenté une nouvelle méthode de détection de piétons basée sur la stéréovision et les noyaux de graphes (Random Walk Kernel). A partir d'une base d'images stéréoscopiques, nous avons extrait plusieurs images de piétons et non-piétons afin de constituer un ensemble d'apprentissage et de test. Le noyau utilisé pour le classifieur SVM est un noyau de graphes, calculé à partir du produit scalaire entre les graphes obtenus à partir des formes des piétons/non-piétons présents dans la base. Les résultats préliminaires sont très encourageant vis-à-vis de la viabilité de la méthode, notamment pour le taux de reconnaissance, élevé pour un faible ensemble d'apprentissage. La méthode apporte également une réponse au problème de variabilité du piéton. Nous prévoyons d'étendre l'expérience à d'autres séquences vidéos présentant des situations et des environnements variés. Si la méthode donne toujours de bons résultats, nous envisagerons alors une application mutli-classes et pourrons ainsi envisager la détection d'autres obstacles : vélos, poussettes, animaux, etc. Nous prévoyons également d'utiliser les graphes dans le cadre du suivi temporel de piétons. Les graphes étant a priori pratiques à utiliser pour cette application.

Remerciements Ce travail est financé en partie par le Programme IST de la Communauté Européenne, avec le réseau d'excellence PASCAL, IST-2002-506778. Cette publication reflète uniquement le point de vue des auteurs.

Références

[1] O. Chapelle, P. Haffner, and V. Vapnik. SVMs for histogram-based image classification. *IEEE Transaction on Neural Networks*, 10(5):1055–1064, 1999.

[2] N. Cristianini and J. Shawe-Taylor. *Introduction to Support Vector Machines*. Cambridge Univeristy Press, 2000.

[3] T. Gartner. Survey of kernels for structured data. In *SIGKDD Explorations*, 2003.

[4] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *Proceedings of the Twentieh International Conference on Machine Learning*, 2003.

[5] David Moore. *A real-world system for human motion detection and tracking*. PhD thesis, California Institute of Technology, 2003.

[6] C. Papageorgiou and T. Poggio. Trainable pedestrian detection. In *Proceedings of the 1999 International Conference on Image Processing*, pages 35–39, 1999.

[7] C. Tomasi S. Birchfield. Dethp discontinuities by pixel-to-pixel stereo. In *IEEE Int. Conf on Computer Vision*, 1998.

[8] J. Serra. Morphologie mathématique. *Traité d'Informatique Géologique*, 6(6):194–238, 1972.

[9] A. Shashua, Y. Gdalyahu, and G. Hayon. Pedestrian detection for driving assistance systems: Single-frame classification and system level performance. In *Proceedings of IEEE Intelligent Vehicles Symposium*, 2004.

[10] V. Vapnik. *Statistical Learning Theory*. Wiley, 1998.

[11] P. Viola, M. Jones, and D. Snow. Pedetrian using patterns of motions and appearance. In *IEEE Int. Conf on Computer Vision*, pages 734–741, 2003.

[12] L. Zhao and C. Thorpe. Stereo and neural network based pedestrian detection. *IEEE Trans. on Intelligent Vehicles Transportation systems*, 1(3):148–154, 2000.

Apprentissages discriminants en reconnaissance de mots cursifs en-ligne

E. CAILLAULT[1], C. VIARD-GAUDIN[1]

[1] IRCCyN – UMR CNRS 6597, Image et Video Communications

Ecole polytechnique de l'université de Nantes, rue Christian Pauc, BP 50609, FR-44306 NANTES cedex 3

{emilie.caillault,christian.viard-gaudin}@univ-nantes.fr

Résumé – Ce papier décrit différents modes d'apprentissage de systèmes hybrides basés sur un schéma neuro-markovien (TDNN multi-états – HMM) appliqués à la reconnaissance de mots cursifs saisis en-ligne. Nous avons considéré différentes fonctions de coût, incluant à la fois des critères d'information mutuelle (MMI) avec un apprentissage discriminant et une estimation du maximum de vraisemblance, pour entraîner le système globalement au niveau mot. Nous avons analysé l'impact de la modélisation markovienne en variant de un à trois le nombre d'états d'un modèle de Markov caché lettre HMM lettre). Plusieurs expérimentations sur ces critères et modélisations ont été menées sur la base IRONOFF dans un contexte de reconnaissance mots non contraints et omni scripteurs et sont retranscrits dans ce papier

Abstract *– This article analyses the behavior of various hybrid architectures based on a neuro-markovian scheme (MS-TDNN HMM) applied to online handwriting word recognition systems. We have considered different cost functions, including maximal mutual information criteria with discriminant training and maximum likelihood estimation, to train the systems globally at the word level and also we varied the number of states from one up to three to model the basic hidden Markov models at the letter level. We report on experimental results for non constrained, writer independent, word recognition obtained on the IRONOFF database.*

1. Introduction

Nos travaux s'intègrent dans le contexte de la reconnaissance de l'écriture en ligne destiné aux systèmes mobiles communicants (assistant numérique personnel, ardoise électronique, smart-phone). Dans ce domaine, les systèmes de reconnaissance mot les plus contraints imposent une écriture de type script [1, 2] facilitant ainsi le problème de segmentation. Les systèmes supportant une écriture cursive sans contrainte conduisent à une architecture nettement plus complexe, c'est précisément le cas du système présenté ici. Nous avons opté pour un système hybride neuromarkovien avec une architecture de type TDNN (Time Delay Neural Network) pour le réseau de neurones. Si de tels systèmes ont déjà été abordés dans la littérature [3, 4, 5], leurs apprentissages posent encore de nombreux problèmes. Nous proposons ici une méthode d'apprentissage global au niveau mot basée sur une fonction de coût générique qui permet de mixer apprentissage de type discriminant et apprentissage de type modèle générateur. La solution proposée représente une simplification intéressante par rapport aux approches nécessitant plusieurs étapes, quelques fois alternées, pour mener à bien l'apprentissage du niveau graphème ou lettre au niveau mot.

Pour illustrer ces travaux, nous décrirons dans un premier temps le processus de reconnaissance du système hybride neuro-markovien (section 2). En section 3, nous détaillons les différents modes d'apprentissage obtenus par variation des paramètres de la fonction de coût présenté dans ce papier avec ces résultats en quatrième partie. Pour améliorer notre système de reconnaissance mot nous présentons en section 5 une première solution, soit une extension multi états de la modélisation markovienne au niveau lettre.

2. Architecture du système

La figure 1 illustre le système de reconnaissance global [6]. Il est basé sur une approche analytique avec segmentation implicite et un apprentissage global au niveau mot.

De plus, il permet de manipuler un lexique de taille variable et ne requiert aucun apprentissage additionnel pour ajouter de nouvelles entrées au dictionnaire mot.

La première étape consiste à normaliser le signal d'écriture, principalement en taille, orientation des lignes de base et s'affranchir de la vitesse d'écriture. Une fois le signal normalisé, sept caractéristiques sont extraites en chaque point, elles seront présentées en entrée du système TDNN+HMM. Le TDNN balaie régulièrement la trajectoire du signal d'écriture avec un pas de segmentation implicite et une fenêtre d'une largeur donnée, elle-même balayée par une fenêtre de convolution de taille fixée à la hauteur du corps du mot. Pour chaque position du TDNN, celui-ci calcule un vecteur O_t correspondant aux probabilités a posteriori des classes associées à ses différentes sorties (66 classes : 2*26 lettres + 14 extensions), elles représentent les probabilités d'observations du HMM. Le HMM [7] effectue la tâche de reconnaissance au niveau mot en trouvant le meilleur

alignement temporel (programmation dynamique) sur les différents modèles mots du lexique.

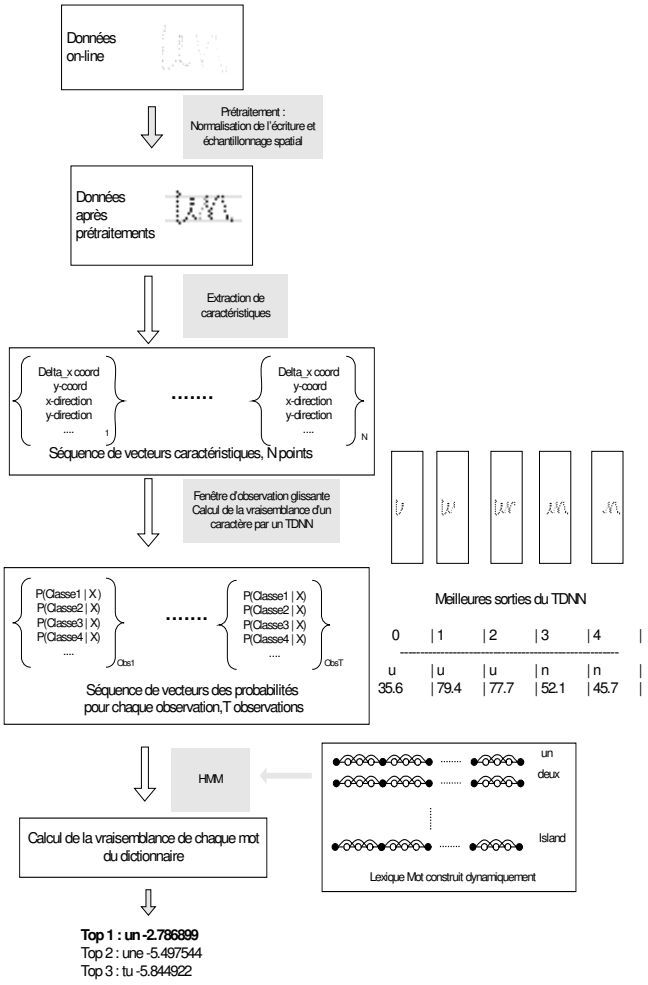


FIG 1 : Processus de reconnaissance du système TDNN+HMM (cas du mot « un »)

Avec une telle approche, une segmentation du mot en lettres est obtenue à l'issue de la reconnaissance, elle correspond aux changements de modèles-lettres sur le meilleur chemin du treillis de reconnaissance du modèle-mot considéré (Viterbi). Le mécanisme d'apprentissage associé à ce système tire profit de l'étiquetage des observations issues du TDNN résultant de cette segmentation implicite. Celui-ci est décrit dans la section suivante.

3. Apprentissages discriminants au niveau mot

L'objectif ici est de rétro-propager directement l'erreur au niveau mot dans le réseau de neurones pour mettre à jour ses paramètres en se basant sur l'étiquetage implicite. Pour cela, nous avons construit une fonction de coût générique L_G combinant à la fois un critère MLE (Maximum Likehood) et un critère MMI (Maximum Mutual Information) permettant de discriminer le modèle mot HMM reconnu à partir d'un lexique et/ou les sorties du TDNN mal classées.

$$L_G = (1+\varepsilon)\ \log P(O|\lambda_{HMMvrai}) - \beta.\begin{bmatrix}(1-\alpha)\log P(O|\lambda_{HMMreconnu}) \\ +\alpha\ \log P(O|\lambda_{TDNNreconnu})\end{bmatrix} \quad (1)$$
$$avec\ HMMreconnu = \arg\max_i P(O|\lambda_i)$$

Les paramètres $\alpha, \beta,$ et ε sont compris entre 0 et 1, le tableau 1 illustre les variantes possibles du critère. Avec $\varepsilon = \beta = 0$, nous obtenons le critère classique du maximum de vraisemblance (MLE). Avec $\beta = 1$, on introduit un apprentissage discriminant (MMI simplifié) qui prend en compte uniquement le mot reconnu en 1ère position par l'algorithme de Viterbi ($\lambda_{HMMreconnu}$) si $\alpha=0$, et dans le cas où $\alpha=1$ seulement les classes reconnues en 1ère position en sortie du TDNN ($\lambda_{TDNNreconnu}$). La valeur α permet de pondérer ces deux derniers types de discrimination.

TAB. 1 : Paramètres en fonction des critères

Paramètres/Critères	ε	β	α
(1) MLE_HMM	0	0	0
(2) MMI_HMM	0	1	0
(3) MLE + MMI_HMM	1	1	0
(4) MLE + MMI_TDNN	1	1	1
(5) Mixte	1	1	0,5

Par souci de place, nous ne détaillerons pas les calculs permettant d'aboutir au gradient de l'erreur L_G donné ci-dessous pour la couche de sortie du TDNN, les couches cachées suivant l'algorithme de rétro propagation classique [7] :

$$\frac{\partial L_G}{\partial w_{ji}} = \sum_t \delta_{j,t} \cdot x_i(O_t)$$
$$avec\ \delta_{j,t} = Grad_{j,t} - x_{j,t}\sum_k Grad_{k,t}$$
$$Grad_{j,t} = \begin{pmatrix}(1+\varepsilon)\dfrac{P(O, q_t = j|\lambda_{HMMvrai})}{P(O,|\lambda_{HMMvrai})} \\ -\beta.\begin{bmatrix}(1-\alpha)\dfrac{P(O, q_t = j|\lambda_{HMMreconnu})}{P(O,|\lambda_{HMMreconnu})} \\ +\alpha\dfrac{P(O, q_t = j|\lambda_{TDNNreconnu})}{P(O,|\lambda_{TDNNreconnu})}\end{bmatrix}\end{pmatrix} \quad (2)$$

0ù j est la position du neurone considéré, i celle du neurone associé de la couche inférieure, t l'indice temporel de l'observation et $x_j(O_t)$ *la sortie du neurone j* de la couche de sortie avec $x_j(O_t) = b_j(O_t)$ selon la notation courante des HMMs : $\lambda(A,B,\pi)$ [8]. $P(O,q_t=j|\lambda)$ est calculé par programmation dynamique [9]. Ainsi pour chaque observation O_t, un gradient positif est rétropropagé pour le vrai modèle HMM (dont une fraction ε correspond à un pur critère ML) et un gradient négatif pour le modèle HMM reconnu et /ou pour les classes reconnues par le TDNN.

TAB. 2 : Calcul du gradient selon les critères

x(j,t) = HMM Vrai (j,t)	x(j,t) = HMM reconnu (j,t)	x(j,t) = TDNN reconnu (j,t)	$Grad_{j,t}$ MLE (1) ($\varepsilon=0$, $\beta=0$)	$Grad_{j,t}$ MMI (2) ($\varepsilon=0$, $\beta=1$)	$Grad_{j,t\,t}$ Generic (5)
F(Faux)	F	F	*0*	*0*	*0*
F	F	V(vrai)	*0*	$-\alpha$	$-\beta\alpha$
F	V	F	*0*	$-(1-\alpha)$	$-\beta(1-\alpha)$
F	V	V	*0*	*–1*	$-\beta$
V	F	F	*1*	*1*	$1+\varepsilon$
V	F	V	*1*	$1-\alpha$	$1+\varepsilon-\beta\alpha$
V	V	V	*1*	$1-(1-\alpha)$	$1+\varepsilon$ $-\beta(1-\alpha)$
V	V	V	*1*	*0*	$1+\varepsilon-\beta$

Le tableau 2 illustre pour différents critères les valeurs prises par la variable $Grad_{j,t}$ en fonction de la sortie du réseau $x(j,t)$ et de son appartenance aux différents chemins : faux (F) signifie que le chemin ne passe par la sortie j du réseau en cette observation t et vrai (V) le contraire. Le HMM vrai correspond au chemin Viterbi du modèle vrai connu, le HMM reconnu au chemin du mot du lexique avec la plus grande vraisemblance et le TDNN reconnu aux sorties de plus forte probabilité. Une normalisation du gradient entre -1 et 1 est ensuite réalisée pour assurer la convergence du réseau.

4. Expérimentations des critères d'apprentissage

Nous avons testé différentes combinaisons du critère précédent sur la base IRONOFF [10]. Un sous-ensemble de 20 898 mots, représentant 197 labels différents, est utilisé pour l'apprentissage et un ensemble séparé de 10 448 mots est testé en généralisation. Le tableau 3 donne les taux de reconnaissance en première position de la base IRONOFF pour chaque critère donné au tableau 1.

TAB. 3 : Taux de reconnaissance sur la base de généralisation (non apprise) IRONOFF

Critères	(1)	(2)	(3)	(4)	(5)
Taux	77,43	83.82	86,34	82,26	87.09

Ce tableau montre qu'un apprentissage uniquement avec un critère de type MLE (1) donne des résultats assez médiocres (77.4 %). Cela doit être mis en relation avec la taille de la base d'apprentissage qui ne permet pas une modélisation suffisamment précise de chaque classe. Dans ces conditions, faire intervenir une compétition entre les classes, avec un critère de discrimination simple (2) où le gradient est nul si le modèle vrai correspond au modèle reconnu permet d'améliorer sensiblement le taux de reconnaissance (83.8 %). La combinaison des deux précédents critères (3) apporte encore un bénéfice notable sur le taux de reconnaissance (86.3 %).

Enfin, un point important apparaît : une discrimination au niveau caractère ($\alpha=0.5$) en plus d'une discrimination au niveau mot (5) accroît les performances du reconnaisseur mot. Il est intéressant de souligner la convergence constatée expérimentalement de ce système d'apprentissage en dehors de toute initialisation spécifique du TDNN et de toute labellisation explicite au niveau caractère. Toutefois, mais c'est le cas assez systématiquement avec les approches neuronales, la valeur du paramètre de pas de gradient influence de manière importante la qualité de la convergence.

5. Modélisation markovienne

Les expériences précédentes étaient fondées sur une modélisation simplifiée : une lettre était représentée par un HMM à un état avec des probabilités de transitions équiprobables et un mot résulte de la concaténation de modèles HMM lettres. Nous avons cherché à assouplir ce modèle, en utilisant des modèles lettres multi-états avec des transitions équiprobables, la figure 2 illustre les modèles lettres à 1 et 3 états. Cette modification a bien sûr une incidence sur la topologie de la couche de sortie du TDNN, celle-ci comporte autant de sorties que d'états distincts. Pour un modèle 3-états, cela fait donc 3 × 66 = 198 sorties.

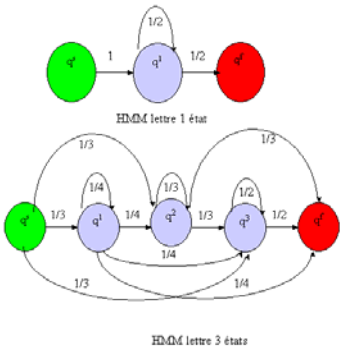


FIG 2 : Modélisation des HMM lettres 1 et 3 états

Le tableau 4 montre l'intérêt d'étendre le nombre d'états du modèle lettre, et cela quelque soit le critère choisi. Par exemple, dans le cas d'un critère MLE-MMI, le modèle 3 états permet de réduire l'erreur de reconnaissance de 41 pourcents. Deux raisons claires sont la conséquence de ces

résultats. La première réside dans les prétraitements du signal d'écriture et la topologie du réseau. La fenêtre d'observation est égale à 4/3 de la longueur moyenne d'un caractère, elle est décalée d'une observation à l'autre de 1/3 de cette longueur. Ainsi un même caractère participe à plusieurs observations consécutives. Avec un seul état il est plus difficile de rendre compte des bruits générés par les lettres qui l'entourent et par les ligatures. Avec plusieurs états il est plus facile d'absorber les différentes situations du caractère dans la fenêtre (début, milieu et fin de la fenêtre). La seconde raison est la plus grande souplesse du modèle HMM pour intégrer les variabilités de longueur intra et inter-lettres. Contenu du nombre variable de graphèmes dans une lettre, une lettre courte à un seul graphème (comme le « i », le « c ») pourra ainsi traverser le modèle par un seul état tandis qu'une lettre longue pourra s'étaler sur deux ou trois états (cas du « f », du « m », du « w »).

TAB. 4 : Taux de reconnaissance du système multi-états

Critères	(1)	(2)	(3)	(5)
1 état	77,43	83,82	86,34	87,09
2 états	80,87	87.46	88.22	ND
3 états	84,69	90,57	92,01	92,78

(ND : non disponible actuellement).

Les approches neuro-markoviennes permettent d'introduire un apprentissage discriminant local et apportent une première réponse intéressante au problème de la discrimination (MMI, combinaison MMI-MLE). L'introduction d'un critère plus local telle une discrimination des sorties du TDNN par rapport au modèle vrai donné par Viterbi permet de renforcer l'apprentissage tant dans sa phase d'initialisation qu'en terme de reconnaissance.

6. Conclusion et perspectives

Dans ce papier nous avons étudié différents modes d'apprentissage global d'un reconnaisseur mot en-ligne basé sur une architecture TDNN-HMM. Nous avons établi une fonction générique qui permet en fonction de ses paramètres de jouer avec des critères plutôt discriminants MMI ou des critères classiques MLE. La combinaison de ces critères est très intéressante en terme de taux de reconnaissance. Il est important de souligner qu'aucune phase d'apprentissage sur des sous-entités segmentées à la main n'a été réalisée, contrairement à ce qui est souvent le cas dans les schémas actuels dans la littérature.

Plusieurs améliorations peuvent être apportées à ce système. Il pourrait être envisagé d'optimiser le nombre d'états lettre par lettre, de travailler sur les modèles de durée et cela de façon dynamique au cours de l'apprentissage afin d'introduire d'abord un *a-priori* assez fort, et également de faire évoluer la fonction de coût générique, en jouant au fur et à mesure des itérations sur les paramètres de contrôle.

Références

[1] L. Oudot, L. Prevost L. & M. Milgram. *Un modèle d'activation vérification pour la lecture de textes manuscrits dynamique*. Colloque International Francophone sur l'écrit et le Document (*CIFED'04)*, La Rochelle, France, 2004.

[2] N. Ragot, E. Anquetil. *A generic hybrid classifier based on hierarchical fuzzy modeling experiments on on-line handwritten character recognition*. Seventh International Conference on Document Analysis and Recognition *(ICDAR'03),* UK, vol. 2, pages 963-967, 2003.

[3] M. Schenkel, I. Guyon, D. Henderson. *On-line cursive script recognition using Time Delay Neural Networks and Hidden Markov Models*. Machine Vision and Applications, special issue on Cursive Script Recognition, (8):215--223, 1995.

[4] S. Jaeger, S. Manke, J. Reichert, A. Waibel. *On-Line Handwriting Recognition: The NPen++ Recognizer*. International Journal on Document Analysis and Recognition *(IJDAR'00)*, volume 3, pages 169-180, 2000.

[5] Z. Wimmer, S. Garcia-Salicetti, A. Lifchitz, B. Dorizzi, P. Gallinari, T. Artières. « REMUS », sur la toile http://www-connex.lip6.fr/~lifchitz/Remus/ .

[6] E. Poisson, C. Viard-Gaudin, P.-M. Lallican. *Système TDNN/HMM de reconnaissance de mots cursifs en ligne à apprentissage simplifié*. Colloque International Francophone sur l'écrit et le Document (*CIFED'04),* La Rochelle, Juin 2004.

[7] Y. LeCun, L. Bottou., Y. Bengio , P. Haffner. *Gradient-Based Learning Applied to Document Recognition*. Intelligent Signal Processing, pages 306-351, 2001.

[8] L.R. Rabiner. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*. Proceedings of IEEE, volume 77, pages 257-285, 1989

[9] R. S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998, A Bradford Book.

[10] C. Viard-Gaudin, P.M. Lallican, S. Knerr, P. Binter. *The IRONOFF Dual Handwriting Database*. International Conference on Document Analysis and Recognition *(ICDAR'99)*, pages 455-458. Bangalore, Sept. 1999.

Analyse des Orientations pour la Caractérisation d'Images de Documents de la Renaissance

N. Journet[1] J.Y Ramel[2] R. Mullot[1] V. Eglin [3]

[1]L3I, Pôle Sciences et Technologie 17042 La Rochelle Cedex 1 - FRANCE njournet@univ-lr.fr

[2]LI, 64 Avenue Jean Portalis 37200 TOURS - FRANCE Jean-yves.ramel@univ-tours.fr

[3]LIRIS INSA de Lyon, Bâtiment Jules Verne Campus de la doua 69622 Villeurbanne cedex - FRANCE eglin@rfv.insa-lyon.fr

Résumé – Cet article présente une nouvelle méthode de caractérisation d'images de documents imprimés datant de la Renaissance. Notre approche se base sur une extraction des différentes orientations présentes sur la totalité de la surface de la page et qui sont caractéristiques de la présence de différentes entités textuelles, ou graphiques (incluant les enluminures, les ornements et bandeaux, les lettrines, ainsi que diverses illustrations). Cette caractérisation s'appuie sur le calcul et l'exploitation de la fonction d'autocorrélation qui a la particularité, lorsqu'elle est estimée sur une zone de texte ou de dessin, de générer une signature unique facilement identifiable. Ce choix nous permet de séparer le texte des dessins, tout en minimisant la quantité d'a priori relatif aux images traitées.

Abstract – This paper presents a method of a page layout characterization adapted to ancient printed documents. Our approach is based on the extraction and the quantification of the various orientations that are present in a document image. The documents are analysed using a normalized square windows in which we analyze significant orientations with a directional rose. Each kind of information (textual or graphical) is typically identified and marked by its orientation distribution. This choice of characterization allows us to separate textual regions from graphics by minimizing the a priori knowledge.

1. Introduction

Ces dernières années ont été marquées par la mise en place de nombreuses campagnes de numérisation. Au fil des années les historiens ont ainsi accumulé une grande quantité d'images numériques de documents anciens. Ces ouvrages proviennent de différents pays (France, Allemagne, Italie…) et recouvrent plus de trois siècles de notre histoire (du milieu du 15ème à la fin du 17ème) pour le patrimoine imprimé. Pendant de longues décennies le monde de l'imprimerie s'est naturellement inspiré des traditions héritées de l'écriture [1]. Les choix relatifs à la mise en page ou la typographie s'en trouvaient donc souvent liés aux seuls choix de l'imprimeur. Ainsi, d'un pays à un autre (ou d'un siècle à l'autre) les ouvrages produits sont caractérisés par une forte variabilité de mises en page et de règles d'édition. Du fait des contraintes techniques et historiques, on observe des structures visuellement très complexes basées sur le multi-colonnage, des corps de texte contenus dans des zones non rectangulaires, des localisations imprévisibles des dessins, ou encore des espacements entre zones graphiques ou textuelles non normalisés, voir figure 1.

FIG. 1 : Exemples de pages de documents anciens actuellement archivés au CESR.

Ce constat nous a poussés à aborder le problème de l'analyse d'images de documents imprimés anciens sous un autre angle. En effet, cette forte hétérogénéité du contenu comme de la mise en page, ne nous permet pas d'utiliser des méthodes classiques de segmentation ou de caractérisation de l'information (approches descendantes, ascendantes ou mixtes).

Dans cet article nous présentons donc une nouvelle approche de caractérisation d'images de documents imprimés extraits d'un vaste corpus de la Renaissance et actuellement numérisés au CESR de Tours. Cette méthode est non seulement robuste aux bruits fréquemment rencontrés dans ce genre d'images (détériorations de l'encre, défauts de numérisation…) mais elle se veut aussi complètement indépendante de la typographie employée, de la taille des caractères, de la présence de parties graphiques… Ce travail est une contribution à l'analyse de la mise en page de documents par une séparation texte/graphique basée sur une localisation et un marquage en zones d'intérêt. L'originalité de ce travail vient de sa grande généricité (applicabilité à des documents de structures très hétérogènes), de l'absence quasi-totale de connaissances a priori sur les propriétés des textes, des graphiques et de leur agencement et de sa grande robustesse aux variations locales de contrastes, aux changements de résolution, aux phénomènes fréquents d'inclusion de texte dans les graphiques.

2. Un constat sur les méthodes d'analyse d'ouvrages anciens

Dans le domaine du document ancien, il n'y a pas un modèle universel. Il est cependant possible de dégager certaines caractéristiques communes à tous ces livres, concernant notamment les mises en page complexes

(plusieurs colonnes de taille irrégulière), les fontes spécifiques ou inconnues, l'utilisation fréquente d'ornements (enluminures, lettrines, cadres…), le faible espacement entre les lignes, ou encore l'espacement non constant entre les caractères et les mots, la superposition de couches d'information (bruit, notes manuscrites…)

Les méthodes d'analyse de documents dédiées aux documents anciens sont quasiment inexistantes, la plupart sont majoritairement dédiées aux documents contemporains. Nous avons cherché à les évaluer sur les ouvrages du patrimoine et avons pu constater qu'elles n'étaient pas directement transposables : nous avons tout d'abord testé des méthodes procédant par agglomération ou découpage de blocs dans les pages. Que ce soit les approches ascendantes (RLSA, extraction de composantes connexes [2,3]) ou les approches descendantes ou mixtes (projections horizontales, analyse multirésolution… [4,5,6]), ces deux catégories de méthodes demandent une grande quantité de connaissances a priori à introduire. Les algorithmes mis en place étaient spécifiquement adaptés à chaque famille de document voir même à chaque type de pages.

Nous avons ensuite testé des outils de caractérisation de textures qui permettent de séparer le texte des dessins sans avoir besoin de connaissances sur le modèle des documents traités. Parmi les travaux les plus pertinents on peut citer ceux se basant sur l'extraction de la matrice de transition des niveaux de gris ou l'application de filtres de Gabor [7]. Ces méthodes permettent d'exprimer l'aspect fortement texturé du texte et donc d'extraire du texte quelque soit ses caractéristiques typographique. Nous avons finalement décidé de nous appuyer sur une fonction d'autocorrélation. Cette dernière nous permet de caractériser, de manière unique, l'anisotropie de grande zones de texte.

3. Caractérisation des zones d'intérêt par extraction des orientations

3.1 Principe de notre méthode

L'objectif de notre méthode est de mettre en place un système d'indexation robuste adapté à un corpus d'images au contenu hétérogène. La première étape de notre méthode consiste à extraire visuellement la mise en page du document en séparant grossièrement les zones de texte (caractères d'imprimerie) des zones de dessin (lettrines, enluminures…). L'image étudiée est tout d'abord redimensionnée pour traiter par la suite des images de taille fixe et connue (900X600). Le fond est ensuite identifié à l'aide d'un algorithme basé sur l'étude de l'homogénéité des niveaux de gris selon la méthode de Sauvola et al [8]. Le marquage des pixels est réalisé à l'aide d'une fenêtre d'analyse de taille fixe que l'on déplace sur l'image. A chaque déplacement, les orientations caractéristiques de la zone sont extraites et analysées afin de labelliser chaque pixel inclus dans la fenêtre comme étant un pixel appartenant à du texte ou un dessin. Cette technique de marquage implique que les pixels peuvent être marqués plusieurs fois, c'est pourquoi un algorithme permettant de choisir le label final de chaque pixel a été mis en place.

3.2 Extraction des orientations

L'extraction des orientations d'une zone de l'image se fait à l'aide d'une fonction d'autocorrélation (form.1).

$$C_{xx}(k,l) = \sum_{k'=-\infty}^{+\infty} \sum_{l'=-\infty}^{+\infty} x(k',l').x(k'+k,l'+l) \quad (1)$$

La rose des directions, outil proposé par Bres [9], permet d'effectuer une analyse de cette fonction d'autocorrélation. Cette rose donne, avec une grande précision, les orientations privilégiées présentes dans le bloc étudié. La rose est en fait un diagramme polaire. Soit (u,v) le point central de l'autocorrelation et θi l'orientation étudiée, on calcule alors la droite Di tel l'ensemble de ses points (a,b) respecte la relation suivante : angle formé par les deux droites $(a,b)(u,v) = \theta_i$. Pour chaque orientation θi on calcule ainsi la somme des différentes valeurs de la fonction d'autocorrélation (form. 2).

$$R(\theta_i) = \sum_{D_i} C_{xx}(a,b) \quad (2)$$

Ces valeurs sont ensuite normalisées (form. 3) pour ne garder qu'un aspect relatif de la contribution de chaque orientation.

$$R'(\theta_i) = \frac{R(\theta_i) - R_{\min}}{R_{\max} - R_{\min}} \quad (3)$$

La figure 2 montre quelques exemples de roses des directions pour 3 images différentes.

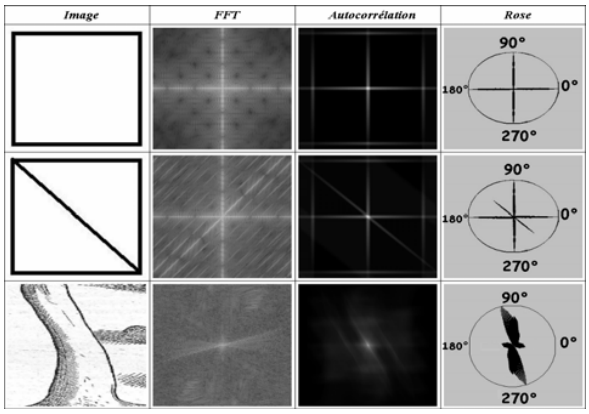


FIG. 2 : Exemples de roses des directions

3.3 Extraction de la mise en page

Compte tenu de sa définition mathématique, la rose possède des propriétés intéressantes. En effet, quand la zone étudiée possède une répartition homogène des directions, la rose produite est une boule parfaite (fig. 3.a). Si on ajoute une ligne à l'image précédente (fig 3.b), le fait que la rose soit normée génère deux caractéristiques invariantes : deux pics horizontaux de même intensité (pour 0° et 180°) et une même intensité pour toutes les autres directions (c'est ce qui donne cette impression de boule).

Une autre propriété intéressante est que la rose est robuste au changement de résolution. Nos tests ont confirmé que quelque soit la résolution à laquelle est calculée la rose, sa forme ne change pas (fig 4).

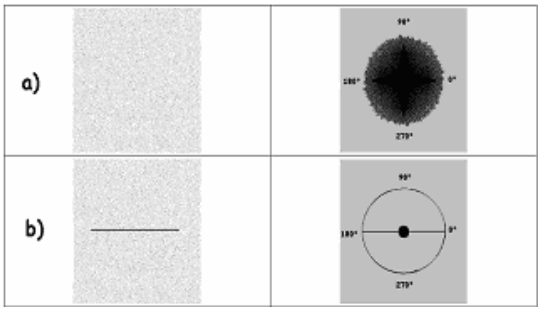


FIG 3 : Caractéristiques de la rose

Nos tests effectués sur le comportement de la rose face au contenu de notre corpus ont permis de déterminer 4 types caractéristiques de roses des directions.

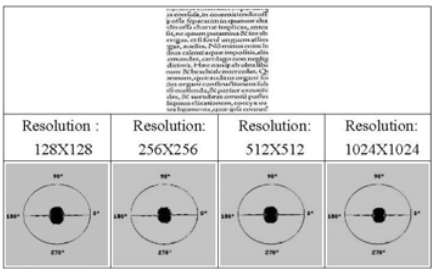


FIG. 4 : Comportement de la rose à différentes résolutions

Il y a deux types de roses qui permettent de caractériser du texte. Une zone de texte homogène (espace interligne constant, taille des caractères constants…) possède les mêmes propriétés que l'image b de la figure 3 : une répartition homogène des directions due aux formes des lettres et une forte corrélation horizontale due à l'orientation du texte (fig. 5.a). Il est donc possible de caractériser une zone homogène de texte par cette rose unique. Lorsque la zone n'est pas homogène (différentes taille de texte, espaces interligne variable…), la formule 3 implique que la taille de la « boule » peut diminuer voir même disparaître (fig. 5.b). Les deux autres formes de roses permettent d'identifier une zone de dessin. Ce type de rose est caractérisé soit par l'absence d'orientations horizontales (fig 5.c), soit par la présence d'une direction horizontale mais sans la présence de « pic » (fig 5.d). Toutes les autres formes de rose ne permettent pas de conclure sur la nature de la zone étudiée.

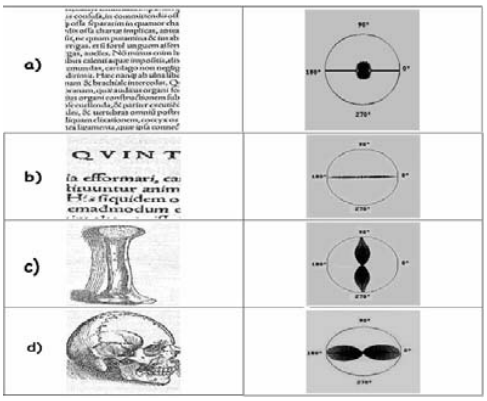


FIG. 5 : détermination de zones graphiques

Comme nous l'avons évoqué précédemment notre approche par fenêtre glissante implique qu'un pixel puisse être étiqueté plusieurs fois. La décision du label final de chaque pixel est prise une fois que le marquage est achevé. Pour décider du label final de chaque pixel de l'image, on comptabilise le nombre de fois ou il a été marqué comme texte par rapport au nombre de fois où il a été marqué comme graphique.

La taille de la fenêtre d'analyse est un paramètre important. Nous avons pour l'instant déterminé une taille optimale de la fenêtre pour analyser des images de documents ayant une résolution de 900X600. Expérimentalement nos tests ont montré que la taille de la fenêtre devait inclure un minimum de 5 à 7 lignes de texte pour permettre la détection d'une zone textuelle. Une alternative à ce choix consiste à laisser l'utilisateur, lors d'une phase d'initialisation, fixer lui-même la taille de cette fenêtre.

4. Résultats expérimentaux

4.1 Qualité de notre méthode

Le tableau 1 récapitule les résultats des tests effectués sur plus de 50 pages en provenant de 6 ouvrages différents de la Renaissance. Ces livres ont été choisis afin d'évaluer notre méthode sur des mises en page, des typographies et des résolutions variées. En ce qui concerne la qualité de la localisation des zones de texte, nous avons quantifié le pourcentage de lettres correctement identifiées, mal identifiées et enfin oubliées. En ce qui concerne la qualité de localisation de zones de dessin nous avons comptabilisé le nombre de pixels correctement identifiés, mal identifiés et enfin oubliés.

Tab. 1 Résultats (% de détection correcte / fausse / oubli)

Texte	Dessins
95/3/2	88/10/2

Quelques résultats sont présentés dans la figure 6.

FIG 6 : Quelques résultats (texte en bleu et dessin en rouge)

Dans la plupart des cas les différentes parties de l'image sont correctement détectées. Les erreurs principales de marquage proviennent de l'analyse de zones de transition entre textes et images mais aussi de titres contenant de gros caractères. Excepté ces deux cas, notre méthode donne de bons résultats et permet d'atteindre l'objectif principal qui est de mettre en évidence le contenu visuel du document sans forcément segmenter la page.

4.2 Utilisation de ce marquage

L'indexation de grosses quantités d'images au contenu hétérogène est un problème complexe. Nous pensons qu'il est possible, à l'aide de ce marquage, d'apporter plusieurs pistes de réponse à ce problème. Notre approche n'est pas une méthode de segmentation à proprement parler: elle a avant tout pour but de découper grossièrement l'image en zones labellisées. On peut trouver plusieurs intérêts à cette phase. Elle permet par exemple de mettre en évidence le contenu visuel du document et donc sa mise en page, elle permet également de simplifier l'application d'éventuels algorithmes de segmentation.

4.2.1 Catégorisation des pages d'un ouvrage

Le marquage obtenu (texte/graphique) peut servir de base pour comparer les pages d'un ouvrage entre elles afin de les regrouper en classes homogènes. La figure 7 démontre l'intérêt d'une telle approche en utilisant simplement comme critère de comparaison la quantité de pixels de dessin et de texte. Cette présentation a deux intérêts : d'une part elle permet d'illustrer la qualité de notre marquage et d'autre part, comme dans [10], nous voyons cette classification comme étant la première étape d'un travail d'indexation d'images de documents anciens. En analysant plus en profondeur les résultats de ce type de classification nous espérons pouvoir mettre en évidence une homogénéité intra-ouvrage (règles de mise en page, typographie spécifique…).

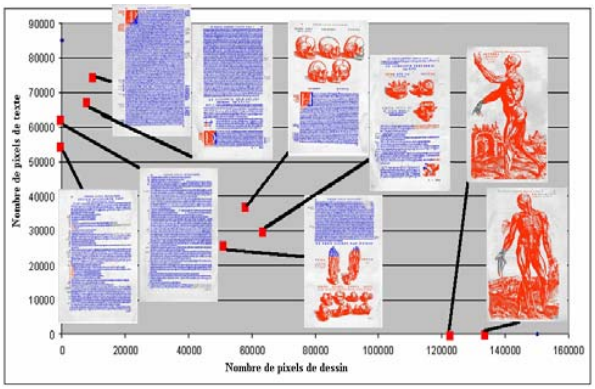


FIG 7 : Exemple d'une classification de pages d'un ouvrage

4.2.2 Segmentation fine des dessins

L'autre piste étudiée est celle d'une segmentation des illustrations basée sur les résultats du marquage. En effet, ce dernier fournis un a priori fort sur la localisation des dessins.

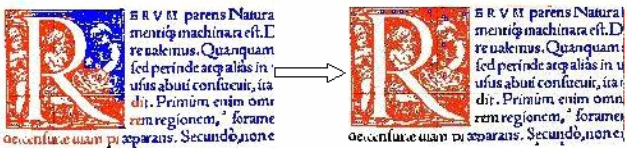


FIG. 8 : Segmentation basée sur l'étude du marquage

En se focalisant successivement sur toutes les régions rouges (graphiques) et en analysant le voisinage de chacune d'elles il est possible de localiser précisément les frontières des parties graphiques. Nos expérimentations montrent qu'il est possible, par exemple, de coupler l'information de notre marquage à un algorithme d'extraction de composantes connexes. Ainsi en étudiant l'information (dessin/texte) contenue dans chaque zone englobante, il est possible d'extraires de manière très fine les zones graphiques (fig. 8).

5. Conclusion

Dans cet article nous avons présenté une méthode de caractérisation d'images de documents imprimés anciens. L'originalité de notre approche tient à l'utilisation d'un outil qui permet de séparer, sans a priori, le texte des zones graphiques de différents types de documents. Le marquage obtenu constitue une représentation visuelle intéressante du contenu de chaque page et peut être utilisé par la suite aussi bien pour obtenir une segmentation physique fine que pour regrouper les pages d'un ouvrage en classes homogènes.

Références

[1] HJ Martin, La naissance du livre moderne, Editions du Cercle de la Librairie, 2000.

[2] Chew Lim Tan, Zheng Zhang Text block segmentation using pyramid structure. University of Singapore, 1998.

[3] L O'Gorman, The Document Spectrum for Page Analysis Layout, Trans IEEE One PAMI.15(11), 1993 P1162-1173.

[4] Jie Xi, J Hu, L Wu, Page segmentation of chinese newspaper Pattern recognition 2002 2695-2704

[5] D. Malerba, F Esposito Oronzo, Adaptive Layout Analysis of document. Università degli Studi di via Orabona Bari, ismis 2002.

[6] P. Duygulu, V. Atalay A Hierarchical Representation of Form Documents for Identification and Retrieval, International Journal on Document Analysis and Recognition IJDAR 5 2002 1, 17-27.

[7] P.B Pati,S Raju N Pali, AG Ramkrihnan. Gabor filters for document analysis in bilingual documents. DEE, Indian Institute of science. Bangalore 560 012 INDIA

[8] J. Sauvola, T. Seppänen, S. Haapakoski, and M. Pietikäinen. Adaptive Document Binarization. In ICDARecognition, volume 1, pages 147–152, 1997.

[9] S Bres, Contributions à la quantification des critères de transparence et d'anisotropie par une approche globale. PhD Thesis, 1994.

[10] G. Maderlechner, P. Suda, T. Bruckner, classification of documents by form and content, Siemens AG, Otto-Hahn-Ring, Munchen, Germany

Remerciements au CESR-CNRS (Tours – France) pour les images.http : www.cesr.univ-tours.fr

Appariement couleur de planches de bois par inférence floue.

E. SCHMITT[1, 2] V. BOMBARDIER[1] P. CHARPENTIER[1]

[1]Centre de Recherche en Automatique de Nancy, CNRS, UMR 7039
Faculté des Sciences – Bd des Aiguillettes – BP 239 – 54506 Vandoeuvre-lès-Nancy Cedex

[2]LuxScan Technologies, Z.A.R.E. Ouest L-4384 Ehlerange, Luxembourg

emmanuel.schmitt@cran.uhp-nancy.fr
vincent.bombardier@cran.uhp-nancy.fr
patrick.charpentier@cran.uhp-nancy.fr

Résumé – Cet article présente une méthode de classification de couleurs sur des planches de bois à partir d'un mécanisme à base de systèmes d'inférences floues implémenté à travers l'algorithme d'Ishibuchi-Nozaki-Tanaka. Le raisonnement flou permet de prendre en compte la subjectivité de la perception des couleurs au sein d'un produit « bois ». De plus, les informations extraites des images ne peuvent pas être considérées comme précises et certaines. L'utilisation de notre méthode améliore les taux de reconnaissance en apprentissage et en généralisation dans la classification des couleurs du bois, mais aussi sur le benchmark IRIS.

Abstract *– This article presents a method of color classification on wooden boards according to a fuzzy inference system mechanism implemented with Ishibuchi-Nozaki-Tanaka's algorithm. The fuzzy reasoning allows taking into account the subjectivity of color perception in the product « wood ». Moreover, the piece of information extracted from the images cannot be considered as accurate and certain. The use of our method improves the recognition rates in training and operating steps for the wood color classification, and for the IRIS database too.*

1. Introduction

L'amélioration de la qualité des produits des industries du bois (scieries, fabricants de meubles, ...) est devenue un enjeu majeur pour rester compétitif sur le marché actuel. En effet, les industriels ne peuvent plus se permettre de fournir des produits uniquement purgés de tous les défauts ; il faut aussi qu'ils soient esthétiquement appareillés.

Ainsi, nous abordons, dans cet article, le contrôle qualité de planches de bois, et plus particulièrement l'appariement couleur de ces planches. Le cadre général de l'étude est d'obtenir un procédé permettant de trier esthétiquement et visuellement des produits finis dans un environnement industriel. Le principal problème provient de l'aspect hautement subjectif de la perception des couleurs qui rend difficile son interprétation. En effet, l'Homme ne dispose pas d'un modèle universel pour définir un ensemble de couleur. Chaque individu possède ses propres caractéristiques visuelles et perceptives, l'œil humain n'étant pas un outil totalement fiable. C'est pourquoi, il est indispensable de rechercher des algorithmes robustes pour traiter des images couleur [13]. Pour tenir compte de la subjectivité de la perception des couleurs, nous avons choisi d'utiliser un système d'inférences basé sur des éléments de la théorie des ensembles flous. Ce système d'inférences floues (S.I.F.) repose en fait sur l'implémentation de règles linguistiques floues pour traiter le vecteur caractérisant la notion de couleur du bois.

Après une description du système de vision et des conditions d'acquisition des images, nous exposerons le principe de la méthode de classification. Enfin, nous évaluerons notre méthode par rapport à un classificateur bayésien, à partir de la base de données IRIS et des échantillons de bois, acquis sur des lignes de production.

Les travaux présentés dans cet article sont réalisés dans le cadre d'une collaboration université-entreprise entre le Centre de Recherche en Automatique de Nancy (CRAN), la société luxembourgeoise LuxScan Technologies.

2. Système de vision

Le premier point traité lors de notre étude consiste à pallier l'imperfection du système d'acquisitions actuellement utilisé (figure 1). En effet, un grand nombre de paramètres externes au système peuvent perturber la reconnaissance des couleurs. Pour exemple, nous pouvons citer la température dont une variation d'environ 10°C peut modifier totalement le spectre d'émission d'un éclairage, et donc modifier complètement la discrimination des couleurs (une pièce initialement reconnue dans la classe « brune » peut se retrouver dans la classe « rouge »). Une étude concernant ces paramètres a permis de fixer les conditions d'éclairage et de choisir le type d'éclairage moins sensible aux variations extérieures et plus stable au niveau de la référence de blanc. Pour des raisons de confidentialité, ces résultats ne peuvent être exposés.

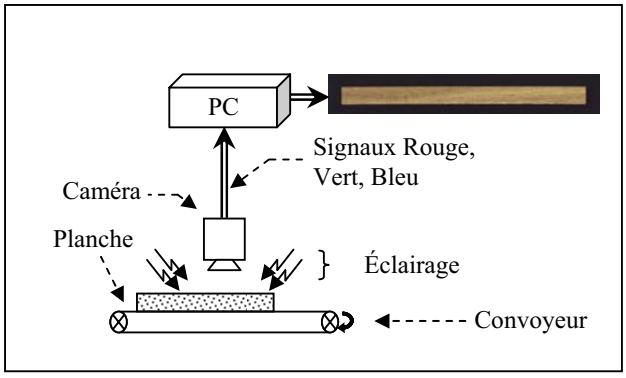


FIG. 1 : Système de vision

La reconnaissance des couleurs doit s'effectuer en temps réel sur des lignes de production pouvant atteindre des vitesses de 180 mètres par minute. Les capteurs utilisés renvoient alors 1500 lignes par seconde, chaque ligne étant composée de 900 pixels. Ainsi, nous pouvons évaluer la quantité de données à traiter, soit environ 1350000 pixels par seconde. Après une conversion des signaux Rouge, Vert et Bleu (RVB) dans le repère colorimétrique Lab conseillé par la Commission Internationale de l'Eclairage, nous pouvons effectuer la classification des données.

La couleur du bois n'étant pas homogène tout au long de la planche, nous travaillons sur des zones dont la taille a été déterminée de façon empirique. Ces Régions d'Intérêts (ROI) ne doivent pas être trop petite pour ne pas rendre la reconnaissance trop sensible aux détails, mais elles ne doivent pas non plus être trop grande pour permettre une discrimination efficace. Les ROIs nous permettent également de mieux traiter les transitions de couleur. Si une planche est composée de plusieurs zones de couleurs différentes, nous devons la classer dans une catégorie spécifique définie au préalable. C'est pourquoi, nous divisons les planches en quatre ROIs (figure 2) sur lesquelles nous calculons le vecteur caractéristique détaillé ci-après.

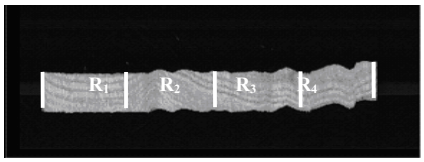


FIG. 2 : Image d'une face composée de 4 ROIs

Pour permettre un traitement rapide des données et pour répondre aux contraintes temps-réel du système industriel, il est nécessaire de caractériser la couleur à partir d'attributs simples. Nous avons choisi le moment statistique d'ordre 0 [12] qui défini le vecteur caractéristique V suivant.

$$V = \begin{pmatrix} m_L \\ m_a \\ m_b \end{pmatrix}. \qquad (1)$$

où m_i correspond à la valeur moyenne du paramètre i (i={L,a,b}).

Nous avons testé d'autres caractéristiques simples comme l'écart type des différentes composantes, ou plus complexes comme des paramètres de texture ou les matrices de cooccurrence. Mais bien souvent le gain obtenu en reconnaissance n'est pas suffisamment significatif par rapport à l'augmentation du temps de calcul qui en découle.

3. Méthodologie de classification couleur

Après avoir défini notre système d'acquisitions, nous devons aborder l'aspect décisionnel pour la classification des planches de bois. A ce niveau, nous avons à résoudre deux problèmes principaux. Tout d'abord, les données délivrées par le capteur peuvent être à la fois imprécises et incertaines ; l'imprécision étant due à la variabilité du capteur lui même, mais également de l'éclairage bien que nous aillons considérablement amélioré ce point. L'incertitude est due, quant à elle, au mode d'extraction et au choix des caractéristiques. Dans ce cas, nous ne sommes pas certains qu'elles soient totalement représentatives de l'information couleur qui nous intéresse. L'autre difficulté réside dans la subjectivité de la perception de la couleur. En effet, la définition des classes de sortie n'est pas stricte, elle peut varier d'un opérateur à l'autre et présenter des intersections non nulles. Il existe de nombreuses méthodes pour classer les planches de bois selon la couleur [4][7]. Nous avons décidé d'utiliser une méthode basée sur la théorie des sous ensembles flous [6] afin de faciliter la manipulation de telles données imprécises et incertaines (Soft Computing) et de pouvoir prendre en compte la subjectivité de la classification couleur (Fuzzy Information Engineering).

3.1 Système d'inférences floues (S.I.F.)

Il existe deux types de raisonnement utilisant des règles floues : le raisonnement abductif qui permet d'obtenir des informations sur une entrée X à partir d'une variable de sortie Y ; et le raisonnement déductif qui permet de déduire la sortie Y à partir des valeurs de l'entrée X. Dans notre cas, nous utilisons le Modus Ponens [14], qui correspond au deuxième cas. Les règles générées peuvent être classées en deux catégories : les règles conjonctives et implicatives. Dans le cas des règles implicatives, le raisonnement est dirigé par la connaissance. Plus nous aurons d'informations sur le produit, plus nous préciserons le résultat. Les informations, que nous devons manipuler dans le cadre de la classification couleur, ne sont pas assez « riches ». C'est pourquoi nous préférons utiliser les règles conjonctives qui sont mieux adaptés à ce type de problème d'analyse de données [2]. L'intérêt d'un tel principe réside notamment dans la rapidité de calcul et dans la certitude d'obtenir une base de règles cohérente. En effet, pour notre application, nous devons aboutir à une méthode simple à mettre en place pour rester en accord avec l'aspect « temps réel » des systèmes de production visés.

De nombreuses méthodes permettent d'obtenir automatiquement les règles conjonctives floues à partir d'un lot de données [1]. Il existe deux principaux modèles qui utilisent ce type de règles : celui de Larsen et celui de Mamdani [12]. La différence entre ces deux modèles est faite

à partir du choix de la T-norme pour l'opérateur d'implication. Finalement, notre choix s'est posé sur une méthode basée sur l'algorithme d'Ishibuchi-Nozaki-Tanaka [5][10] qui correspond à l'implémentation d'un modèle de Larsen. Nous avons choisi ce modèle car le ***Produit*** est plus adapté que le ***Minimum*** pour la T-norme. En fait, en utilisant le ***Produit***, nous autorisons un découpage de l'espace des variables d'entrée de façon non linéaire [11]. Après le traitement des prémisses, nous combinons les différents résultats partiels avec l'opérateur de disjonction (T-conorme), ***Maximum***. Ainsi, la sortie de notre classificateur à bases de règles linguistiques floues est un vecteur flou fournissant les différents degrés d'appartenance de l'échantillon traité (ROI) à chaque classe de sorties envisageable. Pour finir, la décision est prise à partir du maximum de ces degrés d'appartenance.

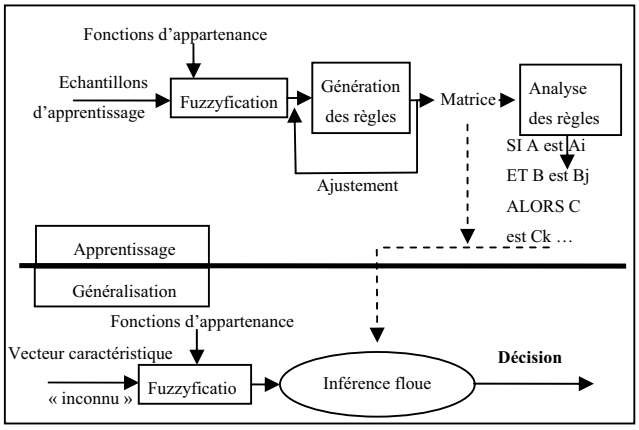


FIG. 3 : Module de reconnaissance des couleurs

Les fonctions d'appartenance (figure 3) sont construites suivant la connaissance des experts. En fait, nous divisons l'espace de représentation des différents attributs (moyennes e L, a et b) grâce à la répartition des couleurs.

3.2 Description de la méthode

Deux phases sont nécessaires pour exploiter en production notre méthode. Tout d'abord, la phase d'apprentissage des moteurs d'inférences, qui se décompose en trois étapes :

- La fuzzification des paramètres d'entrée, qui représente la traduction des données quantitatives en variables linguistiques en fonction d'un nombre de termes adéquats. Dans notre cas, nous cherchons à représenter une caractéristique par des termes répartis sur son univers de discours en fonction de sa plage de variation utile.
- La génération des règles floues du type « Si ... Alors ... » qui décrit ainsi la perception que le système a des couleurs qui sont à identifier.
- L'ajustement du modèle suivant le lot d'échantillons utilisés pour la phase d'apprentissage des couleurs. Cette dernière étape ajuste le découpage de l'espace de représentation des paramètres par itérations successives.

La base de règles floues est sauvegardée sous la forme d'une matrice numérique (figure 3) composée des attributs, des classes de sortie à reconnaître et des degrés de confiance. Ce degré associé à chaque règle peut être assimilé à un degré de pertinence de la règle en accord avec les données d'entrée. Les résultats, que nous obtenons en sortie de notre module de reconnaissance, sont décris sous la forme d'un degré de similarité à la classe de sortie reconnue [3]. En d'autres termes, nous nous intéressons à la ressemblance de la couleur inconnue avec la couleur la plus proche.

Nous proposons d'enchaîner plusieurs S.I.F. afin d'améliorer les temps de calcul. En effet, au lieu de travailler avec un S.I.F unique où toutes les caractéristiques sont entrées en même temps, nous préférons utiliser une structure arborescente, enchaînant plusieurs SIF plus simples. Le nombre total de règles se trouve ainsi réduit. Nous exposons dans la suite de l'article l'apport de notre méthode dans la classification des couleurs ainsi que son positionnement par rapport à d'autres méthodes classiquement utilisées.

4. Résultats

Le choix du nombre de termes de fuzzification pour définir une variable est souvent empirique mais peut être très important pour améliorer les taux de reconnaissance de notre méthode. Pour valider cette partie, nous avons utilisé la base de données IRIS afin de quantifier l'impact du découpage de l'espace de définition des paramètres d'entrées. Dans le cas de cette base composée de 150 échantillons, nous utilisons les 4 paramètres de la base de données (longueur du sépale, largeur du sépale, longueur du pétale et largeur du pétale) afin d'identifier trois classes en sorties (Setosa, Versicolor et Virginica). Le nombre de termes choisi permet de renseigner le nombre de règles linguistiques floues qui seront générées. Pour la base IRIS, nous obtiendrons par exemple 2^4 règles pour une fuzzification en 2 termes et 7^4 règles pour une fuzzification en 7 termes. La fuzzification peut être réalisée de deux manières différentes : suivant une équirépartition des termes ou suivant une répartition adaptée aux classes de sorties.

TAB. 1 : Taux de reconnaissance en généralisation sur la base IRIS pour différents types de fuzzification des entrées

	Classificateur Bayésien	Algorithme de Nozaki	
		2 termes équirépartis	3 termes équirépartis
Taux de reconnaissance	96.83%	52.67%	93.33%
	Algorithme de Nozaki		
	5 termes équirépartis	7 termes équirépartis	3 termes ajustés
Taux de reconnaissance	94.66%	98.00%	97.33%

Les résultats (tableau 1) montrent l'amélioration du taux de reconnaissance pour différentes fuzzifications dans le cas de la généralisation, ainsi qu'une comparaison par rapport à un classificateur bayésien. Dans le cas de la base IRIS, nous constatons que le choix du nombre de termes de fuzzification

est très important pour améliorer les résultats. Cependant, il faut faire attention à ne pas trop décomposer l'univers du discours des attributs car le nombre de règles générées augmente de manière rapide (150 fois plus de règles pour une décomposition en 7 termes par rapport à une décomposition en 2 termes). De plus, en comparaison aux méthodes classiquement utilisées, nous pouvons remarquer que notre méthode semble être justifiée.

Dans un deuxième temps, nous avons utilisé un lot de données plus complexes que le benchmark IRIS afin de tester une structure arborescente. Nous avons sélectionné 500 échantillons de couleur séparées en 6 classes (Brun foncé, Brun, Brun clair, Rouge foncé, Rouge et Rouge clair).

TAB. 2 : Comparaison des taux de reconnaissance et des temps de traitement pour deux types de SIF

	Classificateur bayésien	Algorithme de Nozaki	
	Inférence arborescente	Inférence unique 5 termes	Inférence arborescente 5 termes
Apprentissage	77.20%	91.07%	91.10%
Généralisation	71.50%	83.33%	84.70%
Temps de traitement	0.12 ms	0.10 ms	0.05 ms

Pour notre étude de la couleur de planches de bois, nous avons comparé notre méthode avec celle utilisée par l'industriel, qui prend essentiellement en compte des notions bayésiennes [4][7]. Les résultats exposés dans le tableau 2 montrent 3 atouts de notre méthode :

- une amélioration du taux de reconnaissance en utilisant une structure arborescente ;
- une amélioration du taux de reconnaissance par rapport à la méthode actuellement utilisée ;
- une amélioration du temps de traitement des données.

Tous ces atouts répondent entièrement aux critères imposés par l'environnement de travail sur une ligne de production devant réaliser le tri des planches de bois en temps-réel.

5. Conclusion

La perception de la couleur est une notion très subjective. Cet article présente une méthodologie permettant de tenir compte la perception humaine pour la classification des couleurs. Notre modèle est constitué de règles conjonctives floues, qui sont activées en parallèle et fusionnées par un opérateur de disjonction. Ce modèle est généré à partir de l'algorithme d'Ishibuchi-Nozaki-Tanaka, qui répond totalement aux contraintes liées à nos travaux.

Cette étude démontre l'avantage et l'utilité de la logique floue pour notre problème d'appariement de planches de bois selon la couleur. Outre l'amélioration des taux de reconnaissance, notre méthode permet de diminuer les temps de calculs. Nous notons une diminution de 50% du temps de traitement par rapport à une méthode bayésienne classique, ce qui s'avère non négligeable dans notre cadre d'application temps-réel. Néanmoins, nous pouvons encore améliorer notre méthode en modifiant le vecteur caractéristique d'entrée du S.I.F.. En effet, actuellement seules les moyennes des composantes couleurs calculées sont utilisées. Ce type de données ne permet pas de prendre en compte la grande variabilité du veinage au sein d'une planche de bois, celle-ci pouvant perturber la perception humaine de la couleur. Certaines personnes seront plus sensibles que d'autres aux fortes densités de veinage au sein d'un bois clair. Il serait donc intéressant de travailler sur des aspects de texture [8]. Ce projet a pour but de construire un capteur flou délivrant une information d'aspect du bois en fonction de sa texture et sa couleur [9].

Références

[1] O. Cordon, M.J. Del Jesus, F. Herrera. « A proposal on reasoning methods in fuzzy rule-based classification systems », *Int. Jour. of Approximate reasoning,* 20*,* 1999*,* pp. 21-45.

[2] D. Dubois, H. Prade. « Fuzzy rules in knowledge-based systems – Modelling gradedness, uncertainty and preference », *An introduction to fuzzy logic application in intelligent systems, Kluwer, Dordrecht,* 1992, pp. 45-68.

[3] D. Dubois, H. Prade. « The semantics of fuzzy sets », Fuzzy Sets and Systems, Vol. 90, 1997, pp.141-150.

[4] A. Hanbury, « Morphologie Mathématique sur le Cercle Unité avec applications aux teintes et aux textures orientées », *Thèse doctorale, Ecole Nationale Supérieure des Mines de Paris,* 2002.

[5] H. Ishibuchi, K. Nozaki, H. Tanaka, « Distributed representation of fuzzy rules and its application to pattern classification », *Fuzzy Sets and Systems,* Vol. 52, n° 1, 1992, pp. 21-32.

[6] A. Kaufmann, « Introduction à la théorie des sous-ensembles flous », *Édition Masson, 1975*.

[7] D.E. Kline, R.W. Conners, Q. Lu, P.A. Araman, « Automatic color sorting of hardwood edge-glued panel parts », *Hardwood Symposium Proceedings,* 1997.

[8] Topi Mäenpää, Matti Pietikäinen, « Classification with color and texture: jointly or separately? », Pattern Recognition, 2003.

[9] G. Maurice, E. Benoit, L. Foulloy, « Fuzzy sensors: another view », *Information Engineering,* 1997.

[10] K. Nozaki, H. Ishibuchi, H. Tanaka, « A Simple but powerful heuristic method for generating fuzzy rules from numerical data », *Fuzzy sets and systems*, Vol. 86, 1997, pp. 251-270.

[11] E. Schmitt, V. Bombardier, P. Charpentier, R. Vogrig, « Classificateur flou pour l'appariement couleur de pièces de bois », *LFA*, Nantes, 2004.

[12] E. Schmitt, C. Mazaud, V. Bombardier, P. Lhoste, P. Charpentier, R. Vogrig, « Fuzzy linguistic rules identification system for wood quality », *AISTA*, Luxembourg, 2004.

[13] G. Sharma, H.J. Trussel, « Digital color imaging », *IEEE Transactions on Image Processing,* 6(7), juillet 1997, pp. 901-932.

[14] L.A. Zadeh, « Outline of a new approach to analysis of complex systems and decision processes », *IEEE Trans. On Systems, Man and Cybernetics*, Vol 3, 1973, pp. 28-44.

Reconnaissance et détection robuste d'objets couleur

C. MAAOUI, H. LAURENT, B. EMILE

Laboratoire Vision et Robotique, UPRES EA 2078,
ENSI de Bourges - Université d'Orléans,
10 boulevard Lahitolle, 18020 Bourges Cedex
tel : 02-48-48-40-68, fax : 02-48-48-40-40
choubeila.maaoui@ensi-bourges.fr

Résumé – Dans cet article, nous présentons une nouvelle approche de reconnaissance d'objets basée sur l'utilisation des moments Zernike appliqués sur des images couleur. La détection d'objets repose sur une approche utilisant une segmentation couleur et l'étiquetage des composantes connexes. Une méthode basée sur les séparateurs à vaste marge est utilisée pour l'étape d'apprentissage et de reconnaissance. Nous illustrons la méthode proposée sur des images couleur incluant quelques objets de la base de données COIL-100.

Abstract *– In this paper, we propose an approach of object recognition based on Zernike moments applied on color images. For object segmentation, color segmentation method and the connected component labelling algorithm are used. A Support Vector Machine is used for the learning and recognition steps. Some experimental results on the COIL-100 database are presented.*

1. Introduction

Un des problèmes centraux en interprétation d'images concerne le choix d'une représentation pertinente permettant d'accéder à des primitives significatives et fiables traduisant le contenu de l'image. En général, ces descripteurs sont regroupés en 3 classes : les descripteurs liés à la couleur (histogramme) [5], les descripteurs de texture (matrice de cooccurrence, indices de direction principale et de rugosité, filtres de Gabor et ondelettes) [2], [10] et les descripteurs de formes (descripteurs de Fourier et moments invariants, points caractéristiques) [6],[7],[8]. L'extraction de primitives pertinentes est un problème qui ne connaît pas de solution dans le cas général car il semble qu'il n'existe pas d'attributs qui puissent modéliser une base selon tous les points de vues. Les systèmes actuels tentent de combiner des primitives variées pour améliorer la discrimination et la classification des objets.

Nous proposons ici une approche basée sur la combinaison de deux informations : la couleur et la forme pour la détection et la reconnaissance d'objets dans une image. La section 2 présente la méthode proposée qui est composée de plusieurs étapes. La première étape est la détection d'objets basée sur la segmentation couleur et l'étiquetage en composantes connexes. La deuxième étape concerne la reconnaissance d'objets utilisant les moments Zernike appliqués sur des images couleur. Les phases d'apprentissage et de reconnaissance sont réalisées par un séparateur à vaste marge. La section 3 présente quelques résultats expérimentaux obtenus sur la base d'images COIL-100 [12].

2. Méthode développée

La méthode de détection et de reconnaissance d'objets couleur est composée des étapes suivantes :

- Segmentation couleur afin de séparer l'objet du fond complexe,
- Etiquetage de l'image pour distinguer chaque objet,
- Description de chaque objet en utilisant les moments Zernike,
- Apprentissage et reconnaissance des objets à l'aide d'un séparateur à vaste marge.

2.1 Détection d'objets

Cette partie présente l'étape de détection d'objets.

2.1.1 Segmentation couleur

Afin de segmenter l'image couleur, nous avons utilisé un algorithme efficace proposé et décrit par D. Comaniciu dans [3]. Il est basé sur l'analyse d'espace de caractéristiques. Dans ce paradigme, les pixels sont groupés et exprimés dans un espace de couleur. Chaque groupe représente une région homogène dans l'image. L'algorithme utilise la procédure «Mean Shift» ou déplacement de la moyenne pour palier aux problèmes issus des techniques de clustering. L'algorithme de segmentation proposé ici utilise l'espace des couleurs L*u*v (au lieu de l'espace RGB) plus robuste pour calculer les distances dans l'hypersphère qui constitue la fenêtre de «Mean shift» [3].

2.1.2 Etiquetage de composantes connexes

L'extraction des composantes connexes est largement utilisée en reconnaissance de formes pour étiqueter les régions détectées par segmentation. La technique consiste à regrouper les pixels voisins dans un ensemble appelé composante connexe. Chaque ensemble est disjoint des autres et peut ensuite être aisément isolé [4].

2.2 Reconnaissance d'objets

Un objet est tout d'abord décrit par un ensemble d'invariants qui sont ensuite utilisés dans les étapes d'apprentissage et de reconnaissance.

2.2.1 Représentation d'objet

La représentation d'objets est une étape importante pour la phase de reconnaissance. Elle doit être invariante aux transformations géométriques (rotation, translation et facteur d'échelle) et robuste face à diverses perturbations (bruit, occultation, etc). La représentation d'objets que nous avons adoptée est basée sur les moments Zernike appliqués sur des images couleur.

Les moments Zernike sont définis comme des polynômes complexes qui forment un ensemble orthogonal complet du disque unité $x^2 + y^2 \le 1$. Présenté par Zernike (1934), cet ensemble polynomial ZP est défini comme suit :

$$ZP = \left\{ V_{nm}(x,y) \middle| x^2 + y^2 \le 1 \right\}$$
$$V_{nm}(x,y) = V(\rho,\theta) = R_{nm}(\rho)\exp(jm\theta) \qquad (1)$$

La base polynomiale de Zernike, d'ordre *n* et de répétition *m,* est donnée par :

$$V_{nm}(x,y) = R_{nm}(x,y)\exp\left(jm \arctan(y/x) \right) \qquad (2)$$

$n \in \mathrm{N}^+, m \in \mathrm{N}^+$ avec (n-|m|) pair et $|m| \le n$

Le rapport polynomial est défini comme suit :

$$R_{nm}(x,y) = \sum_{s=0}^{(n-|m|)/2} (-1)^s \frac{(n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n+|m|}{2}-s\right)!} \left(x^2+y^2\right)^{(n-2s)/2} \qquad (3)$$

Soit $\rho = \sqrt{x^2+y^2}$ la longueur du vecteur à partir de l'origine du pixel (x,y) et $\theta = a\tan 2(y/x)$ l'angle entre l'axe x et ce vecteur. $R_{nm}(\rho)$, la représentation en coordonnées polaires $(x = \rho\cos\theta, y = \rho\sin\theta)$ de $R_{nm}(x,y)$ est un polynôme de degré *n*.

$$R_{nm}(\rho) = \sum_{s=0}^{(n-|m|)/2} \frac{(-1)^s (n-s)!}{s!\left(\frac{n+|m|}{2}-s\right)!\left(\frac{n-|m|}{2}-s\right)!} \rho^{n-2s} \qquad (4)$$

Les moments Zernike sont les projections de la fonction d'image $f(x,y)$ dans la base orthogonale de fonctions $V_{nm}(x,y)$. Le moment Zernike, d'ordre *n* et de répétition *m,* est un nombre complexe défini comme suit :

$$A_{nm} = \frac{n+1}{\pi} \sum_{x=0}^{N-1} \sum_{y=0}^{M-1} f(x,y)\left[V_{nm}(x,y)\right]^* \qquad (5)$$

où $x^2 + y^2 \le 1$ et * définit le complexe conjugué.

Les moments Zernike sont bien connus pour être invariants à la rotation. Une normalisation d'image est nécessaire pour rendre ces moments invariants à la translation et au facteur d'échelle [1].

Zernike sur des images binaires

Théoriquement, les méthodes de reconnaissance d'objets basées sur les moments Zernike sont appliquées sur des images binaires où la fonction $f(x, y)$ est égale à 1 ou à 0.

FIG. 1 : Image binaire de deux objets de la base Coil-100 ayant la même forme mais avec des textures différentes.

La figure FIG. 2 représente les moments Zernike (d'ordre 10 et 11) appliqués sur des images binaires des objets présentés dans la Fig.1. On peut noter que pour les deux objets, nous avons des valeurs semblables des moments Zernike.

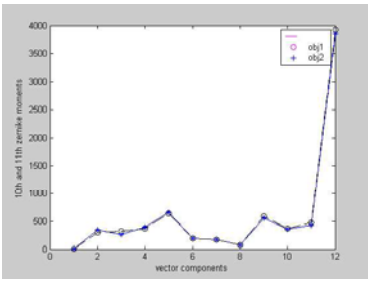


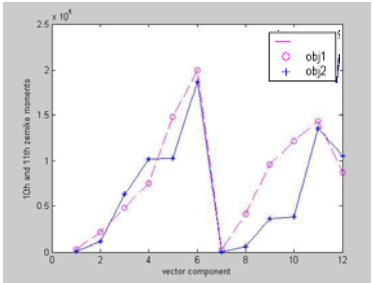
FIG. 2 : Les moments Zernike appliqués sur les images binaires de la FIG. 1.

Zernike sur des images couleur

Pour différentier les objets ayant la même forme, nous avons appliqué les moments Zernike sur des images couleur, où la fonction *f(x,y)* est le niveau de gris du pixel auquel on ajoute la moyenne des composantes couleur RVB (rouge, vert, bleu).

$$f(x,y) = 0.3R + 0.6V + 0.1B \qquad (6)$$

La figure FIG. 3 représente les moments Zernike (d'ordre 10 et 11) appliqués sur les images couleur. La distinction de deux objets de forme semblable peut alors être réalisée.



(c)

FIG. 3 : Les moments Zernike appliqués sur les images couleur de la FIG. 1.

2.2.2 Apprentissage et reconnaissance

Les moments de Zernike calculés précédemment ainsi que la classe de chaque objet forment un ensemble d'apprentissage $\{(x_i, y_i)\}$ $i=1..M$, où chaque $x_i \in \Re^d$ et $y_i \in \{1..N\}$ dans le cas où nous cherchons à reconnaître N classes différentes. Notre objectif est de construire une fonction $f(x)$ qui estime les dépendances entre les exemples x_i et les classes y_i et qui minimise le risque d'erreur de classification pour un point x donné n'appartenant pas à la base d'apprentissage. Dans le cas d'un problème à deux classes $y_i \in [-1,1]$, les séparateurs à vaste marge implémentent l'algorithme suivant. Tout d'abord, les points d'apprentissage x_i sont projetés dans un espace H (de dimension éventuellement infinie) à l'aide d'une fonction $\Phi(.)$. Il est important de noter que pour un même ensemble d'apprentissage, différentes transformations $\Phi(.)$ permettent d'obtenir différentes frontières de décision. En pratique, la transformation est réalisée de manière implicite grâce à un noyau $K(.,.)$ et la fonction de décision s'écrit :

$$f(\mathbf{x}) = \langle w, \Phi(\mathbf{x}) \rangle + b = \sum_{i=1}^{\ell} \alpha_i^* y_i K(\mathbf{x}_i, \mathbf{x}) + b \qquad (7)$$

$$\alpha_i^* \in \mathbb{R} \; \forall i$$

Le critère d'optimalité à maximiser dans le cas des SVM est la marge c'est-à-dire la distance entre l'hyperplan et le point $\Phi(x_i)$ le plus proche de l'ensemble d'apprentissage dans le cas où $\Phi(.)$ est la fonction identité. Les coefficients permettant d'optimiser ce critère s'obtiennent en résolvant le problème suivant :

$$\begin{cases} \max_{\alpha_i} \sum_{i=1}^{\ell} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{\ell} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{sous les contraintes,} \\ 0 \leq \alpha_i \leq C, \\ \sum_{i=1}^{\ell} \alpha_i y_i = 0. \end{cases} \qquad (8)$$

où C est un coefficient pénalisant les exemples se trouvant dans la marge et permettant d'effectuer un compromis entre le nombre de ces derniers et la largeur de la marge. Originellement, les SVM ont été conçus essentiellement pour les problèmes à 2 classes, cependant plusieurs approches permettant d'étendre cet algorithme aux cas à N classes ont été proposées. La méthode que nous utilisons dans cette communication, est une méthode dite "un contre un" [11].

3. Résultats expérimentaux

Nous avons testé notre approche de détection et de reconnaissance d'objets couleur sur la base de données COIL-100 [12]. Cette base de données contient des images couleur de 100 objets avec 72 vues différentes pour chaque objet.

La FIG. 5 présente quelques exemples objets de la base de données, alors que la FIG. 6 présente un objet de la base pour différentes vues.

FIG. 5 : Quelques objets de la base COIL-100.

FIG. 6 : Exemple d'un objet avec différentes orientations et changements d'échelle.

3.1 Résultats de la détection d'objets

La figure FIG. 6 montre un exemple de détection d'objets de COIL-100 sur un fond faiblement texturé en combinant la segmentation couleur et l'étiquetage des composantes connexes. Le fond est détecté comme la zone la plus dominante dans l'image. Il peut être détecté par calcul de l'histogramme de l'image résultant de la segmentation couleur. Les résultats obtenus sont satisfaisants.

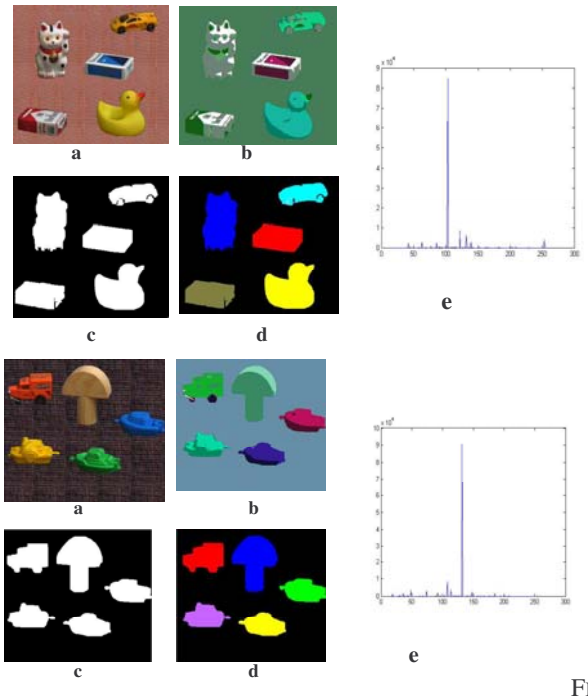


FIG. 6 : Détection d'objets : (a) image originale, (b) segmentation couleur de l'image (a), (c) binarisation de l'image (b), (d) étiquetage de l'image (c), (e) histogramme de l'image (b).

3.2 Résultats de la reconnaissance d'objets

Chaque objet détecté va être identifié en utilisant la description par les moments Zernike et la reconnaissance par SVM. Pour la reconnaissance d'objets, nous avons testé notre approche pour différentes tailles de la base d'apprentissage. Chaque échantillon d'un objet est décrit par un vecteur de dimension 73 (moments Zernike d'ordre 0 jusqu'à l'ordre 15) appliqués sur des images couleur. La figure FIG. 7 représente le taux de reconnaissance des objets de la base de données COIL-100 en fonction de la taille d'apprentissage. Dans ce cas, nous avons utilisé un pourcentage de vues de chaque objet pour l'apprentissage et le reste pour la reconnaissance. On peut noter que le taux de reconnaissance augmente considérablement quand l'information couleur est utilisée.

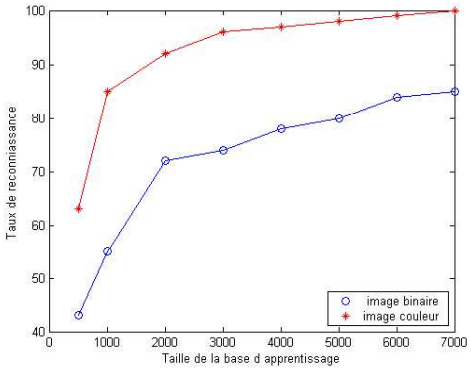


FIG. 7 : Evaluation de taux de reconnaissance en fonction de la taille de la base d'apprentissage.

Pour étudier la robustesse de notre système de reconnaissance, nous avons appliqué notre approche sur une base de données constituée de la base COIL-100 complétée par des images présentant différentes altérations, comme : le fond (noir, en niveau de gris et texturé), l'occultation partielle des objets, la variation de luminance et la présence de bruit blanc gaussien de variance 5, 10 et 20.

Le tableau TAB. 1 présente le taux de reconnaissance pour une base d'apprentissage constituée des 72 vues de chaque objet et d'une base de test constituée de 50 vues pour chaque altération. La robustesse obtenue est excellente vis à vis de la présence de bruit, bonne dans le cas de l'occultation et de la variation de luminance et moyenne dans le cas d'un fond bruité ou texturé.

TAB. 1 : Taux de reconnaissance (%) sur des images avec différentes altérations

N	FU	FB	FT	O-NG	O-FN	Lu	Br5	Br10	Br20
50	64	65	60	84	82	74	99.6	99.6	99.8

Avec FU : Fond Uniforme, FB : Fond Bruité, FT :Fond Texturé, O-NG : Occultation avec un fond en Niveau de Gris, O-FN : Occultation avec un Fond Noir, Lu : Luminance, Brx : Bruit blanc gaussien de variance 5, 10 et 20.

4. Conclusion

Nous proposons dans cet article un algorithme de détection et de reconnaissance d'objets basé sur les moments Zernike appliqués sur des images couleur. La combinaison de la méthode de segmentation couleur et l'étiquetage en composantes connexes permettent de détecter des objets dans une scène sur un fond faiblement texturé. Les résultats ont montré que les descripteurs de Zernike proposés utilisant l'information couleur augmentent le taux de reconnaissance d'objets ayant la même forme. En outre, cette approche est robuste face à différentes altérations : occultation, luminance et bruit.

Dans le futur, nous envisageons d'appliquer la méthode des contours actifs pour améliorer la phase de détection d'objets.

Références

[1] S. M. Abdallah, E.M. Nebot, et D.C. Rye " Object Recognition and Orientation via Zernike moments". In Chin, and Pong, T .C. editors, Proc. Computer Vision ACCV'98, volume 1 of LNCS 1351, pages 386-393. Springer Verlag, 1998.

[2] L. Balmelli, A. Mojsilovic, "Wavelet domain features for texture description, classification and replicability analysis". Proceedings of IEEE International Conference on Image Processing (ICIP), vol. 4, October, pp. 440-444. 1999.

[3] D. Comaniciu and P. Meer, "Robust analysis of feature spaces: color image segmentation", Proceedings of Computer Vision and Pattern Recognition, pp. 750-755, 1997.

[4] L. Di Stefano, A. Bulgarelli, "A simple and efficient connected components labeling algorithm". Image Analysis and Processing Conference, 27-29 Sept. 1999 pp : 322 – 327.

[5] J. Hafner, H.S. Sawhney, W. Equitz, M. Flicker, W. Niblack, "Efficient Color Histogram Indexing for Quadratic Form Distance Functions", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 17, No. 7, July 1995.

[6] M. Hu, "Visual pattern recognition by moment invariants", IRE Trans. On Information Theory, vol. IT-8, pp. 179-167, 1962.

[7] K.Mikolajczyk and C.Schmid. "Indexing based on scale invariant interest points". In International Conference on Computer Vision, 525-531, 2001.

[8] E. Person and K. S. Fu, " Shape Discrimination using Fourier Descriptors", IEEE Trans. Syst. Man Cybern. 7, pp, 170-179, 1997.

[9] J. Platt, "Support Vector Machines", http://research.microsoft.com/users/jplatt/svm.html, 2003.

[10] P. Kruizinga and N. Petkov, "Grating cell operator features for oriented texture segmentation", Proc. of the 14th Int. Conf. on Pattern Recognition, Brisbane, Australia, August 16-20, 1998, pp.1010-1014.

[11] V. Vapnik, "Support-Vector Network", Machine Learning, vol. 20, issue 3, September 1995, pp. 273-397.

[12] http://www1.cs.columbia.edu/CAVE/research/softlib/coil-100.html

Quantification des images couleur par un algorithme de classification crédibiliste

Anne-Sophie CAPELLE[1], Olivier COLOT[2], Christine FERNANDEZ-MALOIGNE[1],

[1]Laboratoire Signal Image Communications, FRE CNRS 2731
Université de Poitiers
Bât. SP2MI, Bd Marie et Pierre Curie
86962 Futuroscope-Chasseneuil, France

[2]Laboratoire d'Automatique, Génie Informatique & Signal, UMR CNRS 8146
Université des Sciences et Technologies de Lille
Cité Scientifique - Bâtiment P2
59655 Villeneuve d'Ascq Cedex, France

anne.sophie.capelle@univ-poitiers.fr
olivier.colot@univ-lille1.fr
fernandez@sic.sp2mi.univ-poitiers.fr

Résumé – Nous proposons un algorithme de quantification des images couleur fondé sur la théorie de l'évidence. Dans ce cadre, le problème de quantification est vu comme un problème de classification non supervisée dans l'espace couleur. Partant d'un nombre de classes égal au nombre de couleurs de l'image originale, l'algorithme de classification crédibiliste estime itérativement de nouvelles partitions crédales. L'originalité du travail présenté repose sur la définition d'un nouveau critère contrôlant la convergence de l'algorithme permettant ainsi d'obtenir le nombre de classes désiré. La nouvelle palette couleur est calculée à partir de la partition crédale obtenue à la convergence de l'algorithme. Cette approche de quantification couleur par une méthode de classification non supervisée présente l'intérêt d'être complètement indépendante de toute initialisation. La méthode de quantification proposée, comparée à des méthodes de quantification existantes a montré son efficacité. En particulier, les résultats en terme de distorsion vis-à-vis de l'image originale sont meilleurs que ceux obtenus avec les autres méthodes testées.

Abstract – This paper presents a novel approach of color quantization using an evidential clustering algorithm. Our method is a top-down approach which iteratively decreases the number of clusters associated with the image until having the desired K number of clusters. The final color palette is then obtained by computing the mean color vector of each cluster. This method is completely automatic which preserves the final result from any initial condition. Experiments on various images show that our algorithm is efficient for color quantization. Quantitative evaluation shows our method provides low distortion relatively to the original image. These results are confirmed by the high visual quality of the quantized images.

1 Problématique

La quantification d'une image couleur a pour objectif de réduire le nombre de couleurs qui compose une image tout en préservant sa qualité. Il existe différentes manières de classer les algorithmes de quantification couleur. Nous retiendrons ici la classification proposée dans [1] qui distingue les méthodes de *pré-clustering*, les méthodes de *post-clustering* et les méthodes *mixtes*. Les méthodes de pré-clustering découpent successivement l'espace couleur (le cube RGB, par exemple) jusqu'à obtenir K sous-espaces permettant ainsi la quantification de l'image en K couleurs. Cette approche a été largement explorée et regroupe les algorithmes de quantification tels que l'algorithme Median-Cut[2], l'algorithme Variance-Based [3] ou l'algorithme MPL [4]. Les méthodes de post-clustering vont quant à elles d'abord sélectionner K couleurs afin d'initialiser les K classes de l'espace couleur. Ces classes pourront être raffinées par la suite et fourniront la palette réduite des couleurs. Les techniques utilisées s'inspirent ainsi des méthodes de reconnaissance des formes et on trouve des méthodes de quantification fondées sur l'algorithme C-Means, l'algorithme Fuzzy-C-Means [5] ou encore sur des algorithmes génétiques [6]. Ces méthodes sont généralement coûteuses en calcul et en temps mais plus efficaces. Enfin, les méthode mixtes combinent les deux approches précédemment énoncées.

L'algorithme que nous proposons se classe parmi les méthodes de *post-clustering* et est issu d'un algorithme de classification crédibiliste [7]. Partant d'un nombre élevé de classes, ce dernier algorithme estime itérativement une partition crédale associée à l'espace des caractéristiques grâce à un critère fondé sur la fonction pignistique. Ce processus, qui est ré-itéré jusqu'à la convergence de l'algorithme, s'accompagne d'une décroissance progressive du nombre de classes. Nous proposons de modifier cet algorithme de classification en un algorithme de quantification en y adjoignant un critère supplémentaire qui permet de contrôler le nombre de classes de la partition. Ce nouveau critère permet ainsi d'obtenir une partition finale composée de K de classes. En associant à chaque classe un représentant, nous obtenons la nouvelle palette couleur. Notre méthode, proche des méthodes de reconnaissance des formes

pourra alors être utilisée préférentiellement dans le cadre d'un pré-traitement.
Après avoir rappelé le contexte crédibiliste dans lequel nous nous plaçons (section 2), nous présentons la méthode de quantification que nous proposons ainsi que les résultats obtenus (section 3).

2 Contexte crédibiliste

La théorie des fonctions de croyance, ou *théorie de l'évidence*, est issue des travaux de Dempster [8]. Formalisée par la suite par Shafer [9], elle est aujourd'hui considérée comme un cadre fédérateur des mesures de confiance.
Différentes interprétations de la théorie des fonctions de croyance ont été proposées. Dans le cadre de nos travaux nous considérons le *modèle des croyances transférables (MCT)* défini par Smets [10, 11] qui apporte une axiomatique et une interprétation non probabilistes des raisonnements incertains. Un aspect important de cette théorie repose de plus sur la séparation entre l'étape de modélisation de la connaissance et celle de la prise de décision.
Dans cette partie, nous rappelons brièvement quelques notions liées au MCT. Les lecteurs se réfèreront à [9, 12, 10, 11] pour plus de détails.

2.1 Modélisation de la croyance

Dans le cadre de la théorie des fonctions de croyance, l'ensemble exhaustif des p hypothèses répondant au problème de classification est appelé *ensemble de discernement* et est noté :

$$\Omega = \{\omega_1, \ldots, \omega_p\} \,. \tag{1}$$

Pour toute proposition $A \subseteq \Omega$, on peut alors définir une probabilité élémentaire $m(A)$, appelée *masse de croyance*, qui représente la véracité d'une proposition A. Cette masse de croyance vérifie :

$$\sum_{A \subseteq \Omega} m(A) = 1 \,. \tag{2}$$

La connaissance sur A est alors telle qu'il est impossible de redistribuer la connaissance sur un sous-ensemble quelconque de A.
Différents modèles ont été proposés pour construire des fonctions de croyance. Nous retenons en particulier le modèle fondé sur des distances que nous utiliseront par la suite [13].

Si on note $\mathcal{X} = \{x_i, i = 1, \ldots, n\}$ l'ensemble des vecteurs formes lié au problème de classification, il est possible de lui associer une *partition crédale* notée $\mathcal{P}$ et définie par :

$$\mathcal{P} = \{m_1, \ldots, m_n\} \,. \tag{3}$$

Cette partition est composée de n fonctions de croyance définies sur Ω où chaque fonction m_i décrit la connaissance relative à l'élément $x_i \in \mathcal{X}$.

2.2 Prise de décision

Suite à l'étape de modélisation de la connaissance au travers des masses de croyance, il est souvent nécessaire de prendre une décision. Généralement, celle-ci est faite en faveur d'une hypothèse singleton $\omega_q \in \Omega$ au moyen d'une probabilité appelée *probabilité pignistique* notée $BetP$ et définie par :

$$BetP(\omega_q) = \sum_{\omega_q \subseteq A} \frac{m(A|\Omega)}{|A|} \,. \tag{4}$$

À partir de cette probabilité, il est alors possible d'appliquer les règles classiques de la décision bayésienne. Dans le cas particulier où les coûts associés à une décision sont les coûts $\{0, 1\}$, la décision se fera en faveur de l'hypothèse maximisant la fonction de probabilité pignistique. Par la suite cette règle sera notée $\mathcal{R}_1$.

3 Quantification crédibiliste

La méthode de quantification que nous proposons est issue d'un algorithme de classification crédibiliste automatique [7, 14]. Nous revenons tout d'abord sur le principe de ce dernier algorithme (section 3.1) avant de décrire notre méthode de quantification (section 3.2).

3.1 Algorithme de classification crédibiliste automatique

L'algorithme de classification automatique sur lequel nous nous appuyons (figure 1) repose sur l'estimation itérative de partitions crédales jusqu'à l'obtention d'une partition stable.
Dans ce cadre, on note $\mathcal{X}$ l'ensemble des n individus à classifier, $\mathcal{P}_t$ la partition crédale obtenue à un instant t et Ω_t l'ensemble de discernement associé $\mathcal{P}_t$. Le nombre exhaustif d'hypothèses associées à $\mathcal{P}_t$ correspond au cardinal de l'ensemble de discernement, à savoir $|\Omega_t|$.
Considérant un ensemble d'apprentissage, une masse de croyance m_i est construite en utilisant la connaissance que nous possédons sur ses k plus proches voisins, selon la modélisation fondée sur des distances proposée par Denœux [13].
La partition initiale $\mathcal{P}_0$ est composée d'autant de classes que d'individus dans l'espace des caractéristiques, i.e à chaque individu x_i est associé une classe ω_i et $|\Omega_0| = n$. Il a été montré que ce processus itératif, reposant sur la règle de décision du maximum de la probabilité pignistique ($\mathcal{R}_1$), converge en un nombre fini d'itérations [15].

Conditions initiales :
$t \leftarrow 0$, partition $\mathcal{P}_0$ telle que $|\Omega_0| = n$
Répéter
  pour tout individu x_i de $\mathcal{X}$ **faire**
    Calculer la masse m_i
    Calculer la probabilité pignistique $BetP_i$ à partir de m_i
    Attribuer un label suivant $\mathcal{R}_1$ dans Ω
  fin pour
  Mise à jour de la partition :
    $t \leftarrow t + 1$:
    $\mathcal{P}_t \leftarrow \mathcal{P}_{t+1}$
    $\Omega_t \leftarrow \Omega_{t+1}$ avec $|\Omega_{t+1}| \leq |\Omega_t|$
jusqu'à l'obtention d'une partition stable

FIG. 1 – Classification crédibiliste

3.2 Quantification crédibiliste

3.2.1 Principe de l'algorithme

La quantification d'une image en K couleurs par une méthode de post-clustering repose sur la construction d'une partition de l'espace couleur en K classes. Dans la section 1, nous avons vu que les approches de post-clustering permettant l'obtention de ces partitions sont diverses. Toutefois, les résultats obtenus sont souvent liés à la définition d'un ensemble d'apprentissage ou des conditions initiales. Un des intérêts de la classification crédibiliste présentée dans la section 3.1 est son aspect non déterministe. L'apprentissage est réalisé à chaque itération de façon automatique et sans aucune condition initiale.

Nous proposons d'utiliser cet algorithme dans le cadre de la quantification des images couleur en adjoignant à cette approche un nouveau critère permettant de contrôler le nombre final de classes i.e. d'obtenir une classification en K classes. Ce critère est appliqué lorsque la convergence vers K classes n'est plus assurée, c'est-à-dire lorsque le nombre de classes obtenu après la prise de décision devient inférieur à K ou lorsque la partition se stabilise avant d'avoir atteint ce nombre. Par la suite nous notons $NbC = Cpt(\mathcal{P}, \mathcal{R})$ le nombre de classes composant la classification de individus $x_i \in \mathcal{X}$ après l'application de la règle de décision $\mathcal{R}$ sur $\mathcal{P}$.

À chaque itération (fig. 2), 4 situations peuvent se produire :

- Cas 1 : le nombre de classes reste supérieur au nombre de classes désiré i.e. $NbC > K$. On continue les itérations.
- Cas 2 : le nombre de classes est égal au nombre de classes désiré i.e. $NbC = K$. On aboutit à la partition recherchée et on peut calculer la palette réduite des couleurs.
- Cas 3 : le nombre de classes est devenu inférieur au nombre de classes désiré i.e. $NbC < K$. Dans ce cas particulier, nous proposons de revenir à la partition $\mathcal{P}_t$ précédente qui vérifie $NbC > K$ et d'appliquer la règle de décision $\mathcal{R}_2$. Cette règle, sur laquelle nous revenons ci-après, permet de définir un sous-ensemble de discernement $\Omega' \subset \Omega_t$ tel que $|\Omega'| = K$. Les individus $x_i \in \mathcal{X}$ anciennement étiquetés parmi les classes de $\Omega_t \backslash \Omega'$ seront alors étiquetés suivant la règles de décision $\mathcal{R}_3$.
- Cas 4 : la convergence de l'algorithme est obtenue et est telle que la partition finale Ω_f possède un nombre de classes supérieur à K. Dans ce cas, on applique la règle $\mathcal{R}_2$ sur Ω_f afin de déterminer Ω'.

3.2.2 Règles de décision $\mathcal{R}_2$ et $\mathcal{R}_3$

Comme nous venons de l'expliquer, l'objectif de la règle $\mathcal{R}_2$ est de définir un sous-ensemble de discernement Ω' à partir d'un ensemble de discernement Ω_t tel que $|\Omega'| = K$ avec $|\Omega_t| > K$.

Nous proposons de définir Ω' en gardant, parmi l'ensemble des classes composant Ω_t, les K classes les plus probables i.e. les K classes associées aux plus fortes valeurs du vecteur pignistique moyen. Plus formellement, ce vecteur de probabilité pignistique moyen $\overline{BetP}$ est défini par :

$$\overline{BetP} = \sum_{i=1}^{n} BetP(x_i), \quad x_i \in \mathcal{X} \ . \tag{5}$$

Il est possible d'ordonnancer les $\overline{BetP(\omega_q)}$ et de définir le sous-ensemble de discernement $\Omega' \subset \Omega_t$ par :

$$\Omega' = \{\omega_{q_i}, \text{pour } i = 1, \ldots, K \text{ et } \omega_{q_i} \in \Omega_t\} \tag{6}$$

avec

$$\begin{array}{l} \overline{BetP(\omega_{q_1})} \geq \overline{BetP(\omega_{q_2})} \geq \ldots \\ \geq \overline{BetP(\omega_{q_K})} \geq \ldots \geq \overline{BetP(\omega_{q_{|\Omega_t|}})} \end{array} \tag{7}$$

Ce sous-ensemble de discernement ainsi défini, il convient de ré-étiqueter les pixels anciennement étiquetées dans l'ensemble des classes supprimées i.e dans $\Omega_s = \Omega_t \backslash \Omega'$. Ils sont alors ré-étiquetés suivant la règle de décision $\mathcal{R}_3$:

$$\mathcal{R}_3(x_i) = \omega_q \quad \text{avec} \quad \omega_q = \arg \max_{\omega_j \in \Omega'} BetP(x_i) \tag{8}$$

À partir de la classification finale en K classes, nous obtenons la palette réduite des couleurs en associant à chaque classe sa couleur moyenne. L'image quantifiée est finalement obtenue en associant chaque pixel de l'image originale à sa plus proche couleur parmi les couleurs de la palette réduite.

Conditions initiales :
$t \leftarrow 0$, partition $\mathcal{P}_0$ telle que $|\Omega_0| = n$, $NbC = n$
Répéter
  pour tout pixel x_i de l'image **faire**
    Calculer la masse m_i
    Calculer la probabilité pignistique $BetP_i$ à partir de m_i
    Attribuer un label suivant $\mathcal{R}_1$ dans Ω_t
  fin pour
  Analyser de la partition : $NbC = Cpt(\mathcal{P}_t, \mathcal{R}_1)$
  $t \leftarrow t + 1$
jusqu'à $NbC \leq K$ **ou** obtention d'une partition stable.

si $NbC < K$ **alors**
  Revenir étape t :
    $\mathcal{P}_{t+1} \leftarrow \mathcal{P}_t$
    $\Omega_{t+1} \leftarrow \Omega_t$ avec $|\Omega_t| > K$
    Définir Ω' à partir de Ω_t selon $\mathcal{R}_2$
    Ré-labeliser suivant $\mathcal{R}_3$ dans Ω'
sinon si la partition est stable et $NbC > K$ **alors**
    Définir Ω' à partir de Ω_{t+1} selon $\mathcal{R}_2$
    Ré-labeliser suivant $\mathcal{R}_3$ dans Ω'
fin si
Calculer la palette réduite des couleurs
Affecter une couleur à chaque pixel de l'image

FIG. 2 – Quantification crédibiliste

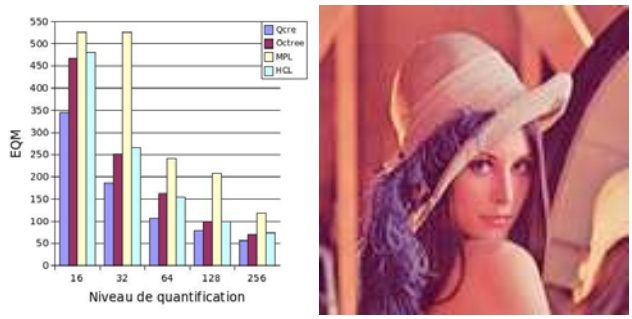
3.3 Résultats

Cette méthode de quantification a été testée sur une large base de données d'images couleur (images naturelles) ainsi que sur des images couleur de références (*lena*, *mandrill*, *peppers*). L'espace couleur utilisé est l'espace couleur RGB. Nous ne présenterons ici que les résultats concernant l'image *lena* au travers de la mesure de l'erreur quadratique moyenne EQM relativement aux images originales :

$$EQM = \frac{1}{n} \sum_{i=1}^{n} \left((R_i - \hat{R}_i)^2 + (G_i - \hat{G}_i)^2 + (B_i - \hat{B}_i)^2 \right) \tag{9}$$

où (R_i, G_i, B_i) représente les valeurs des composantes rouge, verte et bleu du pixel x_i dans l'espace RGB et $(\hat{R}_i, \hat{G}_i, \hat{B}_i)$

les valeurs des composantes rouge, verte et bleu du pixel dans l'image quantifiée en K couleurs.
Nous avons choisi de comparer la performance de notre méthode aux performances de trois approches de la littérature : la méthode Octree [16] qui fait référence dans la problématique de la quantification des images couleur, la méthode MPL [4] de type pré-clustering et la méthode HCL [6] de type post-clustering utilisant des réseaux de neurones.
Pour les quatre méthodes comparées, nous avons réalisé des quantifications en 256, 128, 64, 32 et 16 couleurs et calculé l'erreur quadratique moyenne (eq. 9). Les résultats sont regroupés dans le tableau de la figure 3-a.
Dans ce tableau, nous observons que quelque soit le niveau K de quantification, notre méthode (notée QCre) est celle qui fournit les plus faibles valeurs d'EQM. Ces résultats quantitatifs sont confortés par la qualité visuelle des images quantifiées obtenues. Les figures (fig. 3-c) et (fig. 3-d) représentent respectivement les résultats obtenus pour une quantification en 16 couluers par notre méthode et ceux obtenus par l'Octree (les 2 meilleurs algorithmes d'après le tableau 3-a). On observera que notre image quantifiée présente une très faible distorsion vis-à-vis de l'image originale alors que l'image Octree présente des dégradations en particulier au niveau des cheveux. Des résultats du même ordre sont obtenus sur l'ensemble des images testées.



(a) (b) originale

(c) Qcre (K=16) (d) Octree (K=16)

FIG. 3 – Comparaisons des résultats de quantification sur l'image Lena

4 Conclusion

Nous avons présenté une méthode de quantification crédibiliste des images couleur. L'approche proposée est une approche de *post-clustering* qui consiste à estimer une partition de l'espace couleur en K classes. Cette estimation est faite dans la cadre de la théorie des fonctions de croyance. L'approche proposée est une approche non supervisée : partant d'une partition composée initialement d'autant de classes que de couleurs dans l'image originale, un algorithme itératif permet d'estimer successivement de nouvelles partitions crédales tout en diminuant le nombre de classes. Nous avons défini un critère de contrôle du nombre de classes fondé sur la fonction pignistique. Celui-ci permet d'aboutir à une partition composée de K classes et ainsi d'en déduire une image quantifiée. Cet algorithme non supervisé a été testé sur des images de référence et sur une base de données d'images naturelles. Les résultats quantitatifs et qualitatifs sont très prometteurs.

Références

[1] Luc Brun. *Traitement d'images couleur et Pyramides combinatoires*. Hdr, Univesité de Reims, 2002.

[2] P. Heckbert. Color image quantization for frame buffer display. In *proceedings of SIGGRAPH'82*, pages 297–207, 1982.

[3] S. Wan, P. Prusinkiewicz, and S. Wong. Variance-based color image quantization for frame buffer display. *Color Research and Application*, 15(1) :52–58, 1990.

[4] M.C. Larabi, N. Richard, and C. Fernandez-Maloigne. A fast color quantization using a matrix of local pallets. In *AIPR'2000*, Washington, USA, October 2000.

[5] Dogan Ozdemir and Lale Akarun. A fuzzy algorithm for color quantization of images. *Pattern Recognition*, 35 :1785–1791, 2002.

[6] E. Scheunders. A genetic c-means clustering algorithm applied to color image quantization. *Pattern Recognition*, 30(6) :859–866, 1997.

[7] P. Vannoorenbergue, E. Lefevre, and O. Colot. Traitement d'images et théorie des fonctions de croyance. In *LFA, Rencontres francophones sur la logique floue*, pages 1–4, Tours, France, November 2003.

[8] A. Dempster. Upper and lower probabilities induced by multivalued mapping. *Annals of Mathematical Statistics*, 38 :325–339, 1967.

[9] G. Shafer. *A Mathematical Theory of Evidence*. Princetown University Press, 1976. Princetown New Jersey.

[10] P. Smets and R. Kennes. The transferable belief model. *Artificial Intelligence*, 66(2) :191–234, 1994.

[11] P. Smets. Pratical uses of belief functions. In *Uncertainty in Artificial Intelligence 15 (UAI99)*, pages 612–621, Stockholm, Sweden, 1999.

[12] T. Denœux and M. Masson. Clustering of proximity data using belief functions. In *Proceedings of IPMU'92*, pages 609–616, Annecy, France, 1992.

[13] T. Denœux. An evidence-theoretic neural network classifier. *IEEE International Conference on Systems, Man and Cybernetics*, 3 :712–717, October 1995.

[14] E. Lefevre, J.P Manata, and D. Jolly. Classification par la théorie de l'évidence pour la gestion de tournée de véhicules. In *RFIA'2004*, Toulouse, France, January 2004.

[15] T. Denœux and G. Govaert. Combined supervised and unsupervised learning for system diagnosis unsing dempster-shafer theory. In *CESA'96 IMACS Multiconference, Computational Engineering Applications Symposium on Control, Optimization and Supervision*, volume 1, pages 104–109, 1996.

[16] M. Gervaulz and W. Purgathofer. A simple method for color quantization : Octree quantization. *New Trends in Computer Graphics*, pages 219–231, 1988.

Comparaison expérience/calcul par association de courbes

Jean-Michel LAGRANGE[1], Roberto ARDON[2], Nicolas VILLAIN[2]

[1]CEA, Centre DAM Ile de France, BP 12, 91680 Bruyères-le-Châtel, France.

[2]Philips Medical Systems Research Paris, 51 rue Carnot BP 301, 92156 Suresnes Cedex, France

jean-michel.lagrange@cea.fr, roberto.ardon@philips.com, nicolas.villain@philips.com

Résumé – Nous nous intéressons à une méthode de comparaison entre les résultats de calculs numériques et des vérités terrains qui nous sont présentés comme des courbes planes fermées. Nous réalisons une mise en correspondance des deux formes en nous fondant sur les distances entre leurs points. De cette mise en correspondance nous déduisons une mesure de l'écart entre les deux courbes. En définissant une fonction de coût sur l'espace produit des deux intervalles de paramétrages, nous introduisons une nouvelle méthode qui, à partir d'une initialisation convenable, fournit une bijection réalisant l'association des deux paramétrisations. Notre initialisation est obtenue en utilisant un chemin minimal (par rapport à la fonction de coût) et notre association est recherchée comme une somme de fonctions presque sûrement croissantes. Nous montrons une application de notre méthode pour la déformation de courbes.

Abstract – In this paper we present a new approach for matching two planar curves based on the distance between their points. By defining a cost function on the product space of the two parameterization intervals, we generate a one to one function that associates the two parameterizations. From this function we deduce a mesure of the difference between the two curves. Our method is initialized by using a minimal path (with respect to the cost function) and our matching function is generated on a base of 'almost every where' increasing functions. Our matching can also be exploited for the deformation of curves.

1 Introduction

Les besoins actuels de validation des codes de calcul entraînent de vastes campagnes de comparaison entre les résultats de ces codes et des "vérités terrains". Chez Philips, nous développons des approches pour segmenter automatiquement des régions d'intérêt dans différentes modalités en imagerie médicale (organes, tumeurs, etc.) et nous souhaitons régler de façon robuste les paramètres de ces codes en comparant nos extractions avec des contours fournis par les médecins. Au CEA/DAM, cette même problématique apparaît lors de la confrontation entre des résultats expérimentaux référents (obtenus suite au dépouillement de radiographies X éclair d'objets en grande déformation) et leur restitution par des codes de calcul maison.

Dans les deux cas, les représentations mathématiques que nous manipulons sont des courbes que nous voulons comparer deux à deux de façon rapide, le résultat étant une mesure traduisant leur ressemblance. Ce problème a déjà été largement abordé par le passé. Younès et Bénard [1][2] ont proposé un cadre formel pour l'association optimale de courbes et sa résolution numérique. Cependant, cette dernière reste très lourde, coûteuse en temps de calcul et fournit une association non symétrique (fonction de la courbe choisie comme référence), pénalisante pour l'interprétation. Les travaux présentés dans [3], similaires à [6], fournissent une alternative très séduisante -car ne nécessitant pas d'échantillonage des courbes- qui génère une correspondance via le calcul de chemins minimaux entre deux courbes utilisant un potentiel déduit de leur cartes de distances signées. Néanmoins, cette approche fournit elle aussi une association non symétrique et, en pratique, des problèmes numériques apparaissent lorsque les courbes sont bruitées et très rapprochées. Dans [4], Frenkel et al. proposent un problème de calcul d'une association symétrique entre deux courbes $\mathcal{C}_1$ et $\mathcal{C}_2$ basées sur un précalcul de toutes les distances entre chaque point de ces deux courbes (fournissant une carte de potentiels) et un calcul d'association rapide utilisant la technique du Fast Marching [5]. Pour leur besoin en reconnaissance de formes, leurs distances sont des différences de courbure qui sont incompatibles avec notre besoin de comparaison expérience/calcul du fait des invariances par translation et changement d'échelle et pour les difficultés d'estimation à partir de courbes très bruitées. De plus, **la solution du Fast Marching ne génère pas une association croissante** si bien que des croisements de points associés demeurent possibles (voir un exemple sur le graphe 3(e)) et rendent l'interprétation délicate.

Dans ce papier, nous proposons une approche inspirée par les travaux décrits dans [4] car nous désirons tirer parti de l'association issue du Fast Marching (un chemin minimal) qui demeure une initialisation procurant à la fois rapidité, qualité et robustesse. Dans la section 2, nous fournissons une expression d'une carte de potentiels adaptée à notre problème de comparaison expérience/calcul. Dans la section 3, nous imposons la bijectivité de l'association symétrique sous la forme d'une contrainte sur la dérivée de la fonction associant les deux courbes dans un espace de représentation adapté. La section 4 enfin fournit une solution simple et efficace au problème de la déformation d'une courbe $\mathcal{C}_1$ vers une courbe $\mathcal{C}_2$ basée sur l'association précédente.

Dans ce qui suit, nous détaillons le calcul de la fonction

associant deux courbes $\mathcal{C}_1$ et $\mathcal{C}_2$ d'abscisses respectives s et S comprises dans l'intervalle $[0,1]$ et nous illustrons les résultats de nos algorithmes sur les données de la figure 3(a).

2 Initialisation de l'association : chemins minimaux

En reprenant l'idée donnée dans [4], nous calculons une carte de potentiel $\psi(s,S)$ de $[0,1]^2$ dans $[0,1]$ de sorte que l'association ϕ^{FM} soit un chemin minimal au sens du potentiel ψ. Par notre construction, ψ doit être proche de 1 pour les grands écarts de $\mathcal{C}_1$ et $\mathcal{C}_2$, nul aux croisements des courbes et proche de 0 pour les petits écarts. En définissant $d_{\mathcal{C}_1,\mathcal{C}_2}(s,S) = \|\mathcal{C}_1(s) - \mathcal{C}_2(S)\|^2$ la carte des distances entre $\mathcal{C}_1$ et $\mathcal{C}_2$, nous posons :

$$\psi(s,S) = 1 - e^{-\alpha.d_{\mathcal{C}_1,\mathcal{C}_2}(s,S)}, \forall (s,S) \in [0,1]^2 \qquad (1)$$

Comme dans les problèmes d'autofocalisation, α est un paramètre nous permettant de maximiser globalement le contraste et est pris solution du problème suivant :

$$\arg\max_{\alpha} \int_0^1 \int_0^1 \|\nabla\psi\|^2(u,v)dudv \qquad (2)$$

Etant donnée la carte de potentiel ψ, la carte d'action U est calculée par Fast Marching et l'association ϕ^{FM} est identifiée au chemin minimal sur le potentiel ψ obtenu par descente de gradient sur U à partir du point de coordonnées $(1,1)$. ϕ^{FM} et donc solution du problème suivant :

$$(\mathcal{P}) \quad \phi^{FM} = \arg\min_{\phi} \int_0^1 \psi(s,\phi(s))ds = \arg\min_{\phi} E(\phi) \qquad (3)$$

et il apparaît clairement que la fonction ϕ^{FM} n'est pas nécessairement croissante pour tout s dans l'intervalle $[0,1]$.

3 Association bijective de courbes

En s'inspirant des travaux décrits dans [1][2], la croissance de la fonction ϕ pourrait être obtenue en posant $\phi' = e^h$ et en recherchant la fonction h minimisant :

$$H(h) = \int_0^1 \psi\left(s, \int_0^s e^{h(u)}du\right) ds \qquad (4)$$

Par expérience [2], ce type de critère conduit à un résultat sous optimal (résultat de la descente de gradient $\frac{\partial h}{\partial t} = -\nabla_h H$) et dont la résolution est très lourde.

Nous proposons ici une solution sous optimale en réduisant l'espace de recherche de la fonction ϕ mais qui a l'intérêt d'avoir un faible coût calculatoire et d'intégrer la solution décrite dans la section 2 (ϕ^{FM} minimum global de l'énergie E du problème (P)) comme solution initiale tout en assurant $\phi' > 0$. Nous écrivons $\phi(s) = s + \sum_{k=0}^{K} a_k.P_k(s)$ où K est un nombre de modes de déformations (très inférieur au nombre de points des courbes $\mathcal{C}_1$ et $\mathcal{C}_2$) et nous résolvons le problème suivant :

$$(\mathcal{P}2) \quad \begin{cases} \arg \min_{\{a_k\}_{k\in\{0,..,K\}}} \int_0^1 \psi(s, s + \sum_{k=0}^{K} a_k.P_k(s))ds \\ (\text{Conditions}) \sum_{k=0}^{K} a_k.P_k'(s) > -1, \forall s \in [0,1] \end{cases} \qquad (5)$$

De façon identique aux développements décrits dans le chapitre 8 de [7], les modes P_k sont construits par Analyse en Composantes Principales sur un jeu d'exemples de fonctions croissantes de $[0,1]$ dans $[0,1]$ obtenues en intégrant un processus stochastique positif de façon très probable. Ainsi, pour des coefficients a_k proches de l'unité, la fonction $\phi(s) = s + \sum_{k=0}^{K} a_k.P_k(s)$ croît-elle presque sûrement, ce qui nous permet de résoudre le problème $(\mathcal{P}2)$ en réalisant un ICM sous contraintes sur les a_k.

Les figures 3(c) et 3(d) montrent un exemple de ϕ croissante suite à une initialisation avec une projection de ϕ^{FM} sur les P_k. La figure 3(g) illustre la non bijectivité de ϕ^{FM} (présence de valeurs négatives pour la courbe en trait plein bleu donnant $\frac{d\phi^{FM}}{ds}$) alors que notre solution (tracée en pointillé rouge) est telle que $\phi' > 0$. L'influence locale de la contrainte de positivité de la dérivée est nettement visible sur les figures 3(e) et 3(f).

La mesure d'écart entre les courbes $\mathcal{C}_1$ et $\mathcal{C}_2$ est finalement donnée par $\int_0^1 d_{\mathcal{C}_1,\mathcal{C}_2}(s,\phi(s))ds$.

4 Application à la déformation de courbes

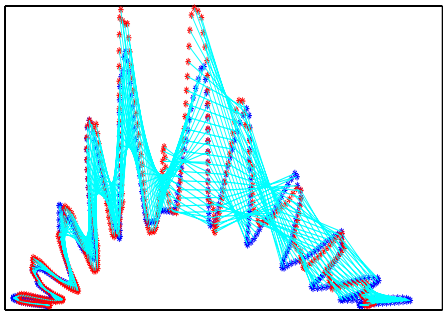
Outre la mesure obtenue à l'optimum du problème $(\mathcal{P}2)$, cette association ϕ strictement croissante peut être utile pour calculer des chemins de déformation "harmonieux" entre $\mathcal{C}_1$ et $\mathcal{C}_2$. A partir de la nappe solution du problème de Dirichlet suivant :

$$\Delta U = 0 \quad \begin{array}{l} \text{avec } U(C_1(s)) = s \\ \text{et } U(C_2 \circ \phi(s)) = s, \forall s \in [0,1] \end{array} \qquad (6)$$

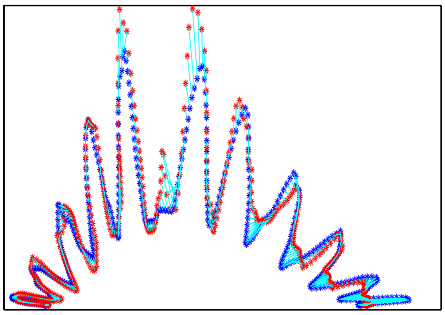
on peut générer des trajectoires $l(s,t)$ (paramétrées par t et pour chaque paramètre s) entre les points $C_1(s)$ et $C_2 \circ \phi(s)$ en recherchant les lignes de niveaux de U, c'est à dire les courbes vérifiant : $\forall (s,t) \in [0,1], U\big(l(s,t)\big) = s$. Sur le problème donné par la figure 2(a) et grâce à l'association solution de $(\mathcal{P}2)$ illustrée sur la figure 2(b), nous obtenons un ensemble de trajectoires qui conduisent à la déformation données sur la figure 2(c).

5 Travaux futurs

Tout d'abord, nous travaillons sur une exploitation de la mesure des écarts $\int_0^1 d_{\mathcal{C}_1,\mathcal{C}_2}(s,\phi(s))ds$ entre un ensemble de "vérités terrain" $\mathcal{C}_1$ et de restitutions ou de segmentations $\mathcal{C}_2$, le but étant de discriminer entre tel ou tel code de calcul et tel ou tel jeu de paramètres.



(a) association obtenue avec $\phi = Id$



(b) association croissante

FIG. 1 – Association de deux courbes, la première en mettant en correspondance les points avec la même abscisse curviligne, l'autre en utilisant notre méthode.

Dans le futur, nous aimerions nous attaquer au problème plus ardu de l'association de surface métrée, c'est-à-dire, à l'extension 3D de l'approche décrite dans cet article.

Références

[1] L. YOUNÈS, R. AZENCOTT , F. COLDEFI A distance for elastic matching in object recognition. Proceedings of 12^{th} ICPR, 1996. 1, 2

[2] M. BÉNARD. Comparaison élastique de courbes à l'aide de distances bâties sur des modèles stochastiques et déterministes. Thèse de doctorat de l'ENS Cachan soutenue le 29 Janvier 1997. 1, 2

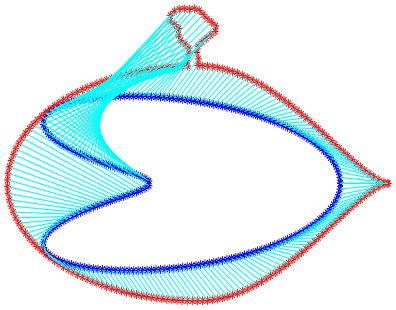
[3] I. COHEN , I. HERLIN. Curves matching using geodesic paths. IEEE Conference on CVPR, pp. 741-746, 1998. 1

[4] M. FRENKEL , R. BASRI. Curves matching using the fast marching methods. EMMCVPR, pp. 35-51, 2003. 1, 2

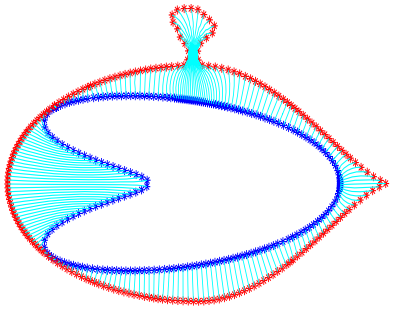
[5] S. OSHER , J.A. SETHIAN. Fronts propagating with curvature dependent speed : algorithms based on the Hamilton-Jacobi formulation. Journal of Computational Physics, vol 79, 12-49, 1988. 1

[6] R. ARDON , L.D. COHEN. Fast Constrained Surface Extraction by Minimal Paths. to appear in IJCV 2005. 1

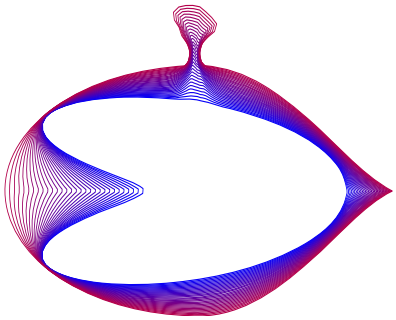
[7] J. M. LAGRANGE. Reconstruction tomographique à partir d'un petit nombre de vues. Thèse de doctorat de l'ENS Cachan soutenue le 16 Décembre 1998. 2



(a) exemple synthétique : élongation d'abscisse

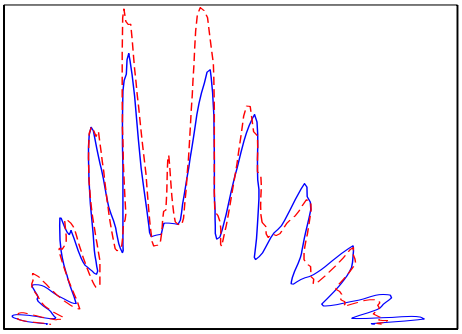


(b) ϕ croissant et chemins d'associations.

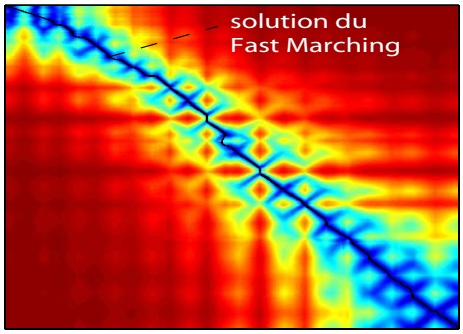


(c) exemple de déformation suivant les chemins déduit par la mise en correspondance.

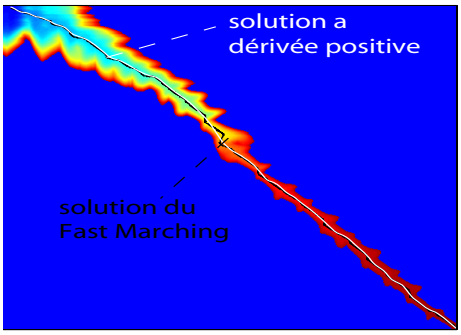
FIG. 2 – Association de deux courbes fermées, la première en mettant en correspondance les points avec la même abscisse curviligne, l'autre en utilisant notre méthode. Nous montrons aussi la déformation qui peut se déduire à partir de notre mise en correspondance.



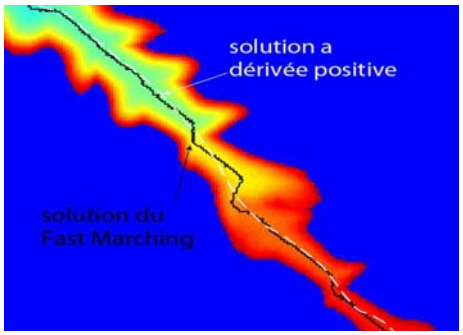
(a) Courbes C_1 et C_2 à associer



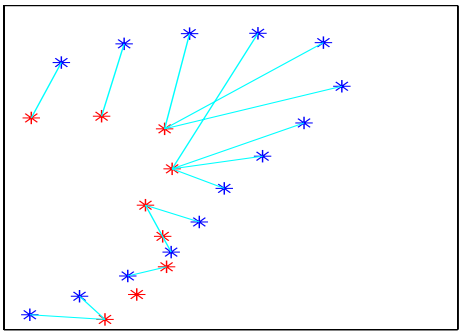
(b) Potentiel basé sur les distances.



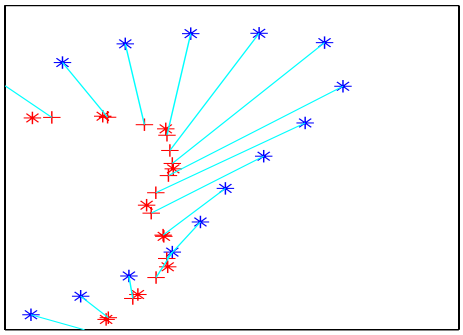
(c) Carte d'action U.



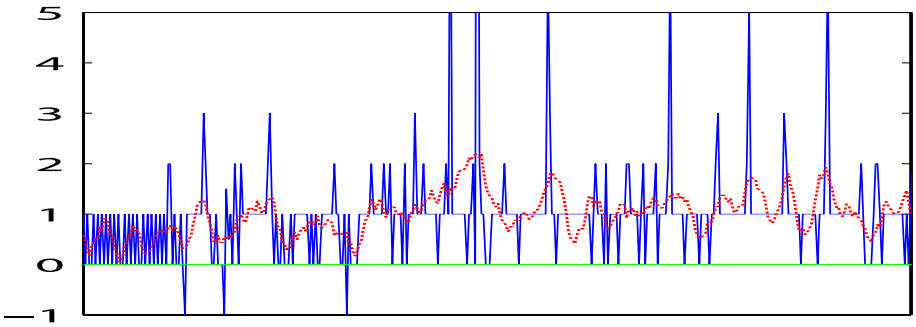
(d) Agrandissement montrant l'écart entre la solution du chemin minimal et la courbe reconstruite par notre méthode.



(e) Association ϕ^{FM}, possibilité d'associations multiples et croisements.



(f) Association croissante générée par notre méthode.



(g) Dérivées de ϕ^{FM} et de ϕ, la première n'est pas strictement croissante.

FIG. 3 – Principales étapes du calcul de ϕ

Une Nouvelle Méthode de Classification en Grande Dimension pour la Reconnaissance de Formes

Charles BOUVEYRON[1,2], Stéphane GIRARD[1], Cordelia SCHMID[2],

[1]LMC–IMAG, BP 53, Université Grenoble 1,
38041 Grenoble Cedex 9 – France

[2]INRIA Rhône-Alpes, projet LEAR, 655 avenue de l'Europe,
38334 Saint-Ismier Cedex – France
Charles.Bouveyron@imag.fr, Stephane.Girard@imag.fr,
Cordelia.Schmid@inria.fr

Résumé – Nous proposons une nouvelle modélisation gaussienne adaptée aux données de grande dimension pour la discrimination et la classification automatique. Notre modélisation est basée sur l'hypothèse que les données de grande dimension vivent dans des sous-espaces dont la dimension intrinsèque est inférieure à la dimension de l'espace. Pour ce faire, notre approche recherche les sous-espaces spécifiques dans lesquels vivent chacune des classes. De plus, nous régularisons les matrices de covariance des classes en supposant que les classes sont sphériques à la fois dans leur espace propre et son supplémentaire. Nous utilisons ensuite ce nouveau modèle en analyse discriminante et en classification automatique dans le cadre de la reconnaissance d'objets dans des images naturelles.

Abstract – We propose a new Gaussian model to classify high-dimensional data in both supervised and unsupervised frameworks. Our approach is based on the assumption that high-dimensional data live in low-dimensional subspaces. Our model therefore finds the specific subspace and the intrinsic dimension of each class to correctly fit the data. In addition, our approach regularizes the class conditional covariance matrices by assuming that classes are spherical both in their eigenspace and in its supplementary. We thus obtain a robust clustering method for high-dimensional data. Our approach is then applied to recognize object in real images and its performances are compared to classical methods.

1 Introduction

La reconnaissance de classe d'objets est un des problèmes les plus difficiles en vision par ordinateur. Les approches les plus efficaces utilisent une description locale des images et des méthodes statistiques de classification pour l'étape de reconnaissance. Cependant, les descripteurs locaux fournissent souvent des données de grande dimension et la classification de telles données est un problème difficile.

En effet, dans des espaces de grande dimension, les performances des méthodes d'apprentissage souffrent du phénomène du *fléau de la dimension* [1] qui est dû au fait que la taille de l'ensemble d'apprentissage est trop faible comparé à la dimension de l'espace. Nous proposons une nouvelle modélisation gaussienne qui localise les sous-espaces dans lesquelles vivent les données de grande dimension et qui régularise les matrices de covariance des classes dans ces sous-espaces. Cette modélisation donne naissance à une nouvelle méthode de discrimination, nommée analyse discriminante de haute dimension, et une nouvelle méthode de classification automatique basée sur l'algorithme d'estimation EM.

Cet article est organisé de la façon suivante. Le paragraphe 2 rappelle le cadre général de la classification dans le contexte gaussien. Nous présentons au paragraphe 3 une nouvelle modélisation des données de grande dimension ainsi que les méthodes de classification associées. Ces méthodes de classification sont utilisées au paragraphe 4 pour la reconnaissance d'objets en vision par ordinateur et des résultats numériques sont présentés au paragraphe 5.

2 Classification et modèles gaussiens

La classification regroupe deux approches distinctes : la classification supervisée (analyse discriminante) et non-supervisée (classification automatique ou *clustering* en anglais). Dans les deux cas, la modélisation gaussienne des classes est classiquement employée.

L'analyse discriminante Le problème de la discrimination est de prédire l'appartenance d'un individu x, décrit par p variables explicatives, à une classe parmi k classes C_1, ..., C_k définies *a priori*. Pour ce faire, nous disposons d'un ensemble d'*apprentissage* $\mathcal{A}$:

$$\mathcal{A} = \{(x_1, c_1), ..., (x_n, c_n),\ x_i \in \mathbb{R}^p,\ c_i \in \{1, ..., k\}\}.$$

La règle de décision optimale, dite *règle de Bayes*, affecte l'observation x à la classe C_{i^*} qui a la probabilité *a posteriori* maximum (MAP) :

$$x \in C_{i^*} \text{ if } i^* = \underset{i=1,...,k}{\operatorname{argmin}} -2\log(\pi_i f_i(x)),$$

où π_i et $f_i(x)$ sont respectivement la probabilité *a priori* et la densité de la classe C_i. La méthode la plus communément employée, l'analyse discriminante linéaire (LDA), suppose que les distributions des classes sont normales, de moyennes μ_i et de même matrice de covariance Σ. Si les matrices de covariance des classes Σ_i ne sont pas supposées communes, alors nous nous trouvons dans le cadre de l'analyse discriminante quadratique (QDA).

La classification automatique La classification automatique a pour objectif d'organiser en classes homogènes un ensemble de données à partir de la seule connaissance des valeurs prises par les individus sur p variables explicatives. Les modèles de mélange, qui supposent que chaque classe est caractérisée par une densité de probabilité, sont classiquement utilisés pour la classification automatique. Dans un modèle de mélange, on considère que les données sont un échantillon de réalisations d'une variable aléatoire $X \in \mathbb{R}^p$ dont la densité s'écrit : $f(x) = \sum_{i=1}^{k} \pi_i f_i(x)$. Les modèles de mélange les plus utilisés en pratique sont les mélanges gaussiens, *i.e.* chaque classe est modélisée par une distribution normale, et l'estimation des paramètres de ces modèles est en général réalisée par l'algorithme itératif EM.

3 Une nouvelle méthode de classification des données de grande dimension

Les méthodes paramétriques de classification souffrent, dans des espaces de grande dimension, du phénomène bien connu du fléau de la dimension. En analyse discriminante, les méthodes telles que QDA et LDA ne sont plus efficaces quand la taille de l'échantillon n est trop petite comparée au nombre de paramètres à estimer qui dépend de la dimension de l'espace p. Il est donc nécessaire de diminuer le nombre de paramètres pour accroître les performances des méthodes de classification. Pour cela, il est possible, soit de réduire la dimension des données (par Analyse en Composantes Principales ou en utilisant l'Analyse Factorielle Discriminante), soit d'utiliser des modèles parcimonieux en faisant des hypothèses supplémentaires sur les modèles (on peut citer par exemple la décomposition spectrale des matrices de covariance des classes de Celeux et Govaert [4]). Cependant, ces méthodes ne permettent pas de résoudre efficacement le problème de la classification en grande dimension car les données sont composées de groupes qui sont généralement cachés dans différents sous-espaces de l'espace initial.

3.1 Le modèle

Le phénomène de l'*espace vide* [6] nous permet de supposer que les données de grande dimension vivent dans des sous-espaces dont les dimensions sont inférieures à p. Afin d'adapter le modèle gaussien aux données de grande dimension, nous proposons de travailler pour chaque classe indépendamment dans des sous-espaces de dimension inférieure à p. De plus, nous supposons que les classes sont de forme sphérique dans ces sous-espaces et leur orthogonal, *i.e.* les matrices de covariance des classes ont seulement deux valeurs propres distinctes. Comme en analyse discriminante classique, nous supposons que les densités des classes sont normales, *i.e.* $\forall i = 1, ..., k$, $f_i(x) = \phi(x, \theta_i)$ où ϕ est la densité d'une loi normale multi-variée de paramètres $\theta_i = \{\mu_i, \Sigma_i\}$. On définit Q_i la matrice orthogonale composée des vecteurs propres de Σ_i. La matrice de covariance Δ_i est définie dans l'espace propre de Σ_i par : $\Delta_i = Q_i^t \Sigma_i Q_i$. Ainsi, la matrice Δ_i est diagonale et constituée des valeurs propres de Σ_i. Nous supposons de plus que Δ_i n'a que deux valeurs propres distinctes $a_i > b_i$. Soit $\mathbb{E}_i$ l'espace affine engendré par les vecteurs propres associés à la valeur propre a_i et tel que $\mu_i \in \mathbb{E}_i$. Soit $\mathbb{E}_i^{\perp}$ tel que $\mathbb{E}_i \oplus \mathbb{E}_i^{\perp} = \mathbb{R}^p$ avec $\mu_i \in \mathbb{E}_i^{\perp}$. Soient respectivement $P_i(x) = \tilde{Q}_i \tilde{Q}_i^{\,t}(x - \mu_i) + \mu_i$ et $P_i^{\perp}(x) = (Q_i - \tilde{Q}_i)(Q_i - \tilde{Q}_i)^t(x - \mu_i) + \mu_i$ les opérateurs de projection sur $\mathbb{E}_i$ et $\mathbb{E}_i^{\perp}$, où $\tilde{Q}_i$ est composée des $d_i < p$ premières colonnes de Q_i complétées par des zéros. Avec ces notations, les paramètres du modèles sont donc $\theta_i = \{\mu_i, a_i, b_i, \tilde{Q}_i, d_i\}$ et le modèle sera noté $[a_i b_i Q_i d_i]$ dans la suite.

3.2 Analyse discriminante de haute dimension

Nous allons construire la règle de décision associée au modèle présenté ci-dessus de la même manière que les méthodes classiques d'analyse discriminante, c'est à dire en utilisant la méthode du *maximum a posteriori* (MAP). En utilisant le modèle $[a_i b_i Q_i d_i]$, la règle de Bayes donne alors lieu à une nouvelle règle de décision δ^+ qui consiste à affecter x à la classe C_{i^*} qui minimise la quantité K_i :

$$\begin{aligned} K_i(x) &= \frac{1}{a_i}\|\mu_i - P_i(x)\|^2 + \frac{1}{b_i}\|x - P_i(x)\|^2 \\ &\quad + d_i \log(a_i) + (p - d_i)\log(b_i) - 2\log(\pi_i). \end{aligned}$$

En permettant ou non à certains des paramètres de varier, nous obtenons 24 modèles particuliers incluant QDA et LDA dans le cas de classes sphériques et dont certains peuvent être interprétés de façon géométriques (voir [2] pour plus de détails). Les estimateurs du maximum de vraisemblance des paramètres du modèle $[a_i b_i Q_i d_i]$ sont alors :

$$\hat{a_i} = \frac{1}{d_i}\sum_{j=1}^{d_i} \lambda_{ij} \text{ et } \hat{b_i} = \frac{1}{(p - d_i)}\left(\text{tr}(\hat{\Sigma_i}) - \sum_{j=1}^{d_i}\lambda_{ij}\right),$$

où λ_{ij} est la $j^{\text{ème}}$ plus grande valeur propre de $\hat{\Sigma_i}$ qui est estimée de façon classique par $\hat{\Sigma_i} = \frac{1}{n_i}\sum_{x_j \in C_i}(x_j - \bar{x_i})(x_j - \bar{x_i})^t$, avec $n_i = \text{Card}(C_i)$. On peut remarquer que la règle de décision δ^+ ne nécessite pas de projeter sur l'espace $\mathbb{E}_i^{\perp}$ et par conséquent il n'est pas nécessaire d'estimer les vecteurs propres associés aux $(p - d_i)$ plus petites valeurs propres de Σ_i. Les d_i premières colonnes de Q_i sont estimées par les vecteurs propres associés aux d_i plus grandes valeurs propres de $\hat{\Sigma_i}$. Cela engendre une économie importante du nombre de paramètres à estimer et régularise d'autant l'estimation de Δ_i. D'autre part, l'estimation des dimensions intrinsèques d_i des classes est faite grâce au *scree-test* de Cattell [3] qui est basé sur l'ébouli des valeurs propres de la matrice $\hat{\Sigma_i}$. On ferra référence dans la suite à cette méthode sous le nom de HDDA.

3.3 Clustering des données de haute dimension

Le modèle génératif que nous avons présenté au paragraphe 3.1 peut-être utilisé en classification automatique. Nous rappelons que l'algorithme EM permet d'estimer les paramètres $\theta = \{\pi_1, ..., \pi_k, \mu_1, ..., \mu_k, \Sigma_1, ..., \Sigma_k\}$ du modèle de mélange en construisant une suite $(\theta^{(q)})_q$ qui maximise la *vraisemblance complétée*. L'algorithme EM prend la forme suivante pour la classification des données de grande dimension :

- Initialisation : choix arbitraire d'une solution initiale $\theta^{(0)}$,
- Itération jusqu'à convergence :

* **Étape E** : calcul des probabilités conditionnelles $t_{ij}^{(q)} = t_i^{(q)}(x_j)$ d'appartenance à la classe C_i

$$t_{ij}^{(q)} = \pi_i^{(q-1)} f_i(x_j, \theta^{(q-1)}) / \sum_{l=1}^{k} \pi_l^{(q-1)} f_l(x_j, \theta^{(q-1)})$$
$$= 1 / \sum_{l=1}^{k} \exp\left[\frac{1}{2}\left(K_i^{(q-1)}(x_j) - K_l^{(q-1)}(x_j)\right)\right].$$

* **Étape M** : maximisation de la vraisemblance conditionnellement aux $t_{ij}^{(q)}$,

$$n_i^{(q)} = \sum_{j=1}^{n} t_{ij}^{(q)},\ \hat{\pi}_i^{(q)} = \frac{n_i^{(q)}}{n},\ \hat{\mu}_i^{(q)} = \frac{\sum_{j=1}^{n} t_{ij}^{(q)} x_j}{n_i^{(q)}},$$
$$\hat{\Sigma}_i^{(q)} = \frac{1}{n_i^{(q)}} \sum_{j=1}^{n} t_{ji}^{(q)} (x_j - \hat{\mu}_i^{(q)})(x_j - \hat{\mu}_i^{(q)})^t,$$

et les paramètres a_i, b_i et Q_i s'estiment à l'étape q de la même façon qu'au paragraphe précédent. Cette méthode de clustering sera dénommée HDDC dans la suite.

4 Application à la reconnaissance de classes d'objets

La plupart des approches efficaces en reconnaissance d'objets utilisent une description locale des images : des régions d'intérêt invariantes à l'échelle sont détectées sur les images et sont ensuite caractérisées par un descripteur local. Dans cet article, nous utiliserons la même approche que [5] (Harris-Laplace + SIFT) pour décrire les images. Nous considérons également la reconnaissance de classe d'objets dans un cadre faiblement supervisé, *i.e.* l'objet à identifier n'est pas segmenté dans les images d'apprentissage. L'ensemble d'apprentissage $\mathcal{A}$ contient donc des descripteurs locaux extraits dans des images contenant l'objet (notés P) et dans des images ne contenant pas l'objet (notés N). Notre approche va combiner les méthodes de classification supervisée et non-supervisée que nous avons présentées pour apprendre de façon automatique les parties discriminantes de l'objet. Ainsi, nous pourrons decider si une nouvelle image contient ou non l'objet.

4.1 Apprentissage

Clustering L'étape de clustering organise les descripteurs du jeu d'apprentissage en k groupes homogènes grâce à la méthode de clustering présentée au paragraphe 3.3. D'un point de vue théorique, les descripteurs x sont les réalisations d'une variable aléatoire $X \in \mathbb{R}^p$ dont la densité est :

$$f(x) = \underbrace{\sum_{i=1}^{k} R_i \pi_i \phi(x, \theta_i)}_{\text{Objet}} + \underbrace{\sum_{i=1}^{k} (1 - R_i) \pi_i \phi(x, \theta_i)}_{\text{Fond}},$$

où $R_i = \mathbf{1}_{\{C_i \in O\}}$. Notre méthode de clustering HDDC fournit les estimateurs des paramètres π_i et θ_i, $\forall i = 1, ..., k$. Ainsi, il ne nous reste plus qu'à estimer les paramètres R_i.

Sélection des classes discriminantes Cette étape consiste à identifier les classes discriminante de l'objet en calculant les estimateurs des paramètres R_i. Malheureusement, l'estimateur du maximum de vraisemblance de R_i n'a pas de formulation explicite et doit être calculé grâce à une procédure itérative. Toutefois, nous avons remarqué que l'estimateur de R_i est proche en pratique de l'estimateur des moindres carrés $\hat{R}$:

$$\hat{R} = (\Psi^t \Psi)^{-1} \Psi^t \Phi,$$

où $\Psi_{ij} = P(x_j \in C_i | x_j)$ et $\Phi_j = P(x_j \in O | x_j)$. On suppose d'autre part que $\forall x_j \in P$, $P(O|x_j) = 1$ et $\forall x_j \in N$, $P(x_j \in O|x_j) = 0$. Ainsi, $\hat{R}$ nous fournit une mesure du pouvoir discriminant de chacune des k classes définies par l'HDDC.

4.2 Reconnaissance

On identifie deux types d'applications de la reconnaissance d'objets dans des images naturelles : la localisation de l'objet dans une image et la classification d'images contenant l'objet.

Localisation de l'objet L'objectif est d'identifier avec précision les descripteurs de l'image qui appartiennent à l'objet considéré. L'approche probabiliste que nous proposons ici nous permet de connaître la probabilité de classification correcte de chacun des descripteurs de l'image. Ainsi, pour localiser l'objet, il suffit de considérer les descripteurs ayant la plus grande probabilité d'appartenir à une des sous-classes de l'objet. La formule de Bayes nous permet d'obtenir la probabilité $P(x \in O|x)$:

$$P(x \in O|x) = \sum_{i=1}^{k} R_i P(x \in C_i|x) = \sum_{i=1}^{k} \frac{R_i \pi_i \phi(x, \theta_i)}{f(x)},$$

où la probabilité *a posteriori* $P(x \in C_i|x)$ que le descripteur x d'une nouvelle image appartienne à la classe C_i est donnée par notre méthode d'analyse discriminante HDDA présentée au paragraphe 3.2.

Classification d'images Cette tâche vise à décider si l'objet considéré est présent ou non dans une image. Des travaux précédents [5] utilisent une méthode empirique basée sur le nombre de détections positives dans l'image pour décider de la présence ou non de l'objet dans cette image. Ce genre d'approche ne prend pas en considération la probabilité avec laquelle chaque descripteur est classé. Nous proposons donc d'utiliser cette information dans notre technique de décision en calculant pour une nouvelle image I un score $S \in [0, 1]$. Plus le score de l'image est proche de 1, plus il est vraisemblable que l'image contienne l'objet. Le score S est calculé de la façon suivante :

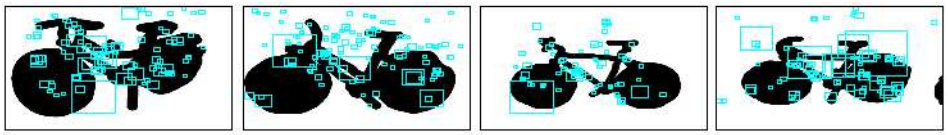
$$S = \frac{1}{m} \sum_{x_j \in I} P(x_j \in O|x_j),$$

où m est le nombre de descripteurs extraits dans l'image I. On décidera alors qu'une image I contient l'objet si son score est supérieur à un certain seuil.

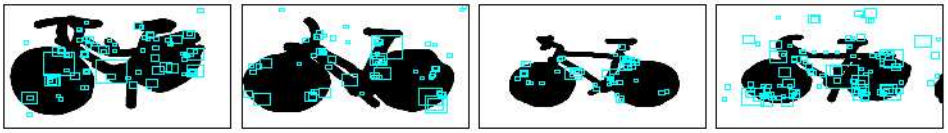
5 Résultats

Nous avons évalué nos méthodes de classification sur un jeu de données[1] proposé récemment. Il faut noter que les images de cette base sont complexes et que l'objet (vélo) est arbitrairement localisé dans les images. Le jeu d'apprentissage contient 40 images ce qui représente $n = 20000$ descripteurs locaux en dimension $p = 128$.

1. Données disponibles sur *http://www.emt.tugraz.at/ pinz/data/*.



(a) Modèle gaussien diagonal et localisation empirique (similaire à [5]).



(b) Modèle $[a_i b_i Q_i d_i]$ et localisation probabiliste.

FIG. 1: *Localisation de l'objet "vélo". Le même nombre de descripteurs est affiché pour les deux méthodes.*

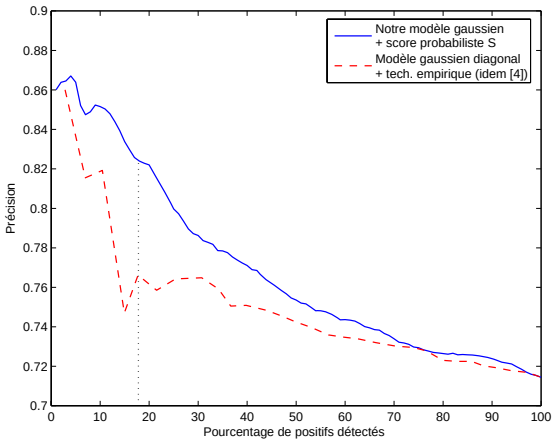


FIG. 2: *Comparaison des résultats de localisation de l'objet.*

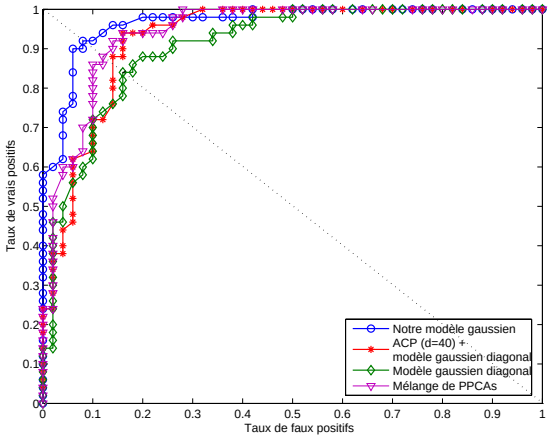


FIG. 3: *Comparaison des résultats de classification d'images.*

5.1 Résultats de localisation

La figure 1 montre les résultats de la localisation de l'objet sur des images segmentées manuellement : (a) approche de [5], utilisant un modèle gaussien diagonal et utilisant uniquement les classes les plus discriminantes de l'objet, (b) notre approche qui utilise le modèle gaussien $[a_i b_i Q_i d_i]$ et basée sur les probabilités *a posteriori* d'appartenance à l'objet. Pour cette illustration, nous avons affiché le même pourcentage de descripteurs classés comme appartenant à l'objet (18%, correspondant à la ligne verticale pointillée de la figure 2). La figure 2 présente les résultats numériques de classification correcte des descripteurs locaux avec ces deux méthodes. On peut observer sur les deux figures que notre méthode est plus efficace et permet de localiser un nombre important de points sur l'objet avec une précision satisfaisante.

5.2 Résultats de classification d'images

La figure 3 présente les résultats de classification d'images obtenus avec différentes méthodes. Nous comparons la méthode de classification associée au modèle $[a_i b_i Q_i d_i]$ à d'autres méthodes de classification. La méthode basée sur un modèle gaussien diagonal est la plus simple mais est également celle qui donne les résultats les moins bons (0,84 sur la diagonale). La méthode qui combine réduction de dimension par ACP et classification basée sur un modèle diagonal confirme l'intérêt de réduire la dimension mais ne fournit toutefois pas des résultats satisfaisants (0,86). La méthode des mélanges de PPCA [7], basée sur une modélisation similaire à la notre mais qui considère que les dimensions des sous-espaces sont communes et qui ne régularise pas les matrices de covariance, donne des résultats meilleurs (0,88) que les méthodes précédentes mais nettement moins bons que ceux obtenus avec notre méthode (0,92).

Références

[1] R. Bellman, *Dynammic Programming*, Princeton University Press, 1957.

[2] C. Bouveyron, S. Girard et C. Schmid, *Analyse Discriminante de Haute Dimension*, Rapport de Recherche INRIA, n° 5470, Janvier 2005.

[3] R. Cattell. *The scree test for the number of factors*, Multivariate Behavioral Research, 1, 2, pp. 245-76, 1966.

[4] G. Celeux et G. Govaert, *Gaussian Parsimonious Clustering Models*, Pattern Recognition, 28, 5, pp. 781-793, 1995.

[5] G. Dorko and C. Schmid, *Object class recognition using discriminative local features*, Technical Report, INRIA, n° 5497, 2004.

[6] D. Scott and J. Thompson, *Probability density estimation in higher dimensions*, Fifteenth Symposium on the Interface, pp. 173-179, Holland, 1983.

[7] M. Tipping and C. Bishop, *Mixtures of probabilistic principal component analysers*, Neural Computation, 11, 2, pp. 443-482, 1999.

Localisation de défauts de capteur pour la validation des mesures hydrométriques dans réseaux d'assainissement

F. HAMIOUD[1,2], C. JOANNIS[1] et J. RAGOT[2]

[1]LCPC, Laboratoire Central des Ponts et Chaussées, Route de Bouaye BP 2941. 44341 Bouguenais cedex. France.

[2]Institut National Polytechnique de Lorraine, Centre de Recherche en Automatique de Nancy, UMR 7039 CNRS-UHP-INPL, 2, Avenue de la forêt de Haye. 54516 Vandoeuvre cedex. France.

Collaborateur : Groupement pour l'évaluation des mesures et des composants en eaux et assainissement, 149, rue Gabriel PERI. 54500 Vandoeuvre Les Nancy. France.

hamioud@lcpc.fr; Joannis@lcpc.fr; Jose.Ragot@ensem.inpl-nancy.fr

Résumé : Dans ce travail, nous proposons une approche de diagnostic hors ligne permettant d'identifier les capteurs en défauts parmi les capteurs installés sur le réseau de mesure du réseau d'assainissement. Cette approche est basée sur un raisonnement flou avec prise en compte du concept de persistance de défaut.

Abstract— *In this study we represent an approach of diagnosis in differed times adapted of this kind of the sewer network to isolate the sensors in faults among the set of sensors on the measurement network. This approach is based on fuzzy concepts with taking into account the fault persistence parameter in our reasoning.*

1. Introduction

Les mesures dans les réseaux d'assainissement sont difficiles et perturbées. Les résultats faux peuvent considérablement perturber un bilan. Il est donc nécessaire d'éliminer des analyses, les valeurs erronées. La validation des mesures comme traitée dans [1]-[6], permet d'analyser la représentativité de ces mesures et de contrôler leur consistance. Une méthode efficace pour valider ces mesures est de multiplier l'équipement de mesure. Cette redondance appelée matérielle introduit plus de complexité dans le système et peut être très coûteuse.

L'objectif de ce travail est de développer une approche automatique qui permet le diagnostic des défauts quand ceux ci se présentent simultanément sur un ou plusieurs capteurs dans le réseau de mesure. Cet objectif est atteint, en développant des modèles pour l'ensemble des données pour générer des relations de redondance de l'information (redondance analytique) [7] ; et en appliquant par la suite, une approche de diagnostic pour isoler chaque défaut qui peut se présenter. Dans ce travail, la méthode de diagnostic à base de modèles est considérée. La première étape de cette méthode est la génération des résidus [8], [9]. Ils sont généralement utilisés dans différentes méthodes de diagnostic, spécialement dans celles basées sur les concepts du raisonnement flou [10]-[14]. Une approche de diagnostic basée sur le raisonnement flou est présentée dans ce papier. Elle est développée pour résoudre un problème de diagnostic concret appliqué au réseau de mesure défini dans la section suivante. Dans les autres sections, l'ensemble des modèles linéaires établi pour générer les résidus et la matrice des signatures des défauts expérimentale seront présentés. L'approche de diagnostic sera détaillée; elle consiste à la «fuzzification » des résidus et la définition du paramètre de persistance introduit dans le raisonnement flou. Et avant de conclure, les résultats expérimentaux obtenus seront présentés et discutés.

2. Réseau de mesure de Nantes

Le système de mesure installé sur le réseau d'assainissement de Nantes est représenté dans la figure 1.

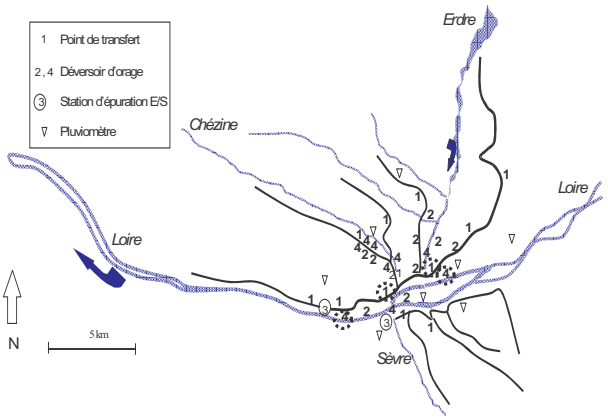


FIG. 1 : Système de mesure de Nantes

Il opère depuis 2001 et comporte 25 sites de mesure de débit et de 8 pluviomètres, pour un total de 53 capteurs utilisés. Chaque site de mesure est constitué d'un ou de plusieurs capteurs (mesure de hauteur et(ou) mesure de vitesse d'eau dans le réseau).

Le présent travail est appliqué sur un ensemble de sites de mesure indiqués par des cercles en pointillés sur la figure 1. Ces sites comportent quatre variables de mesure de hauteur d'eau et une variable de mesure de vitesse d'eau. Ce travail peut être appliqué sur une partie du système ou sur tout le système. La formulation du problème traité reste similaire, seulement la taille des données à traiter change.

3. Modèles pour le réseau de mesure

Dans le cadre de cet article, un modèle représente une description mathématique des relations qui existent entre les différentes variables (mesures) sur le réseau de mesure. L'ensemble de modèles permet ainsi, de créer les relations de redondance analytique [7]. Celles ci permettent d'exprimer une variable selon une ou plusieurs autres variables, qui sont en corrélation physique[1] et statistique avec cette variable.

La forme générale de ces modèles linéaires est donnée par :

$$Ax(k) = 0, \quad (1)$$

$A \in \Re^{l.n}, x \in \Re^n$, où A est la matrice des paramètres des modèles et $x(k)$ est l'état général du système à l'instant k.

Dans le cas où toutes les variables sont mesurées, l'équation de mesure est donnée comme suit :

$$y(k) = x(k) + \varepsilon(k) + S\delta(k), \quad (2)$$

$y \in \Re^n$, $\varepsilon \in \Re^n$, $\delta \in \Re^q$, $S \in \Re^{n.q}$, où $y(k)$ est le vecteur de mesure, $\varepsilon(k)$ est le vecteur de bruit de mesures supposé aléatoire et suit une loi Gaussienne de moyenne nulle et de matrice de covariance $V_\varepsilon \in \Re^{n.n}$, $\delta(k)$ est le vecteur de défauts qui peuvent affecter quelques capteurs et S est la matrice qui exprime la direction des défauts.

La nature des variables d'entrées du modèle (les variables explicatives), est fixée par la nature de la variable à expliquer, Ainsi, on a des relations entre

- les hauteurs et les vitesses locales (même site de mesure);
- les hauteurs dans le site global (dans tout le réseau);
- les débits en aval et en amont traduis en hauteurs et vitesses.

Ces modèles sont établis à différents pas de temps, on a exploité ceux à pas de temps d'une minute.

Pour le réseau de mesure global, l'état $x(k)$ est défini par :

$$x(k) = [h_1(k), \ldots, h_{n_1}(k) \quad v_1(k), \ldots, v_{n_2}(k)] \quad (3)$$

où $n_1 + n_2 = n$, $h_i(k)$ et $v_j(k)$ sont respectivement, la hauteur et la vitesse d'eau dans le réseau, n_1 et n_2 sont les nombres de hauteurs et de vitesses d'eau respectivement.

4. Génération des résidus

Les résidus sont des signaux indicateurs de défauts [15]. La méthode utilisée pour générer ces signaux est basée sur l'exploitation des modèles établis. En éliminant les variables inconnues $x(k)$ des deux équations (1) et (2), on obtient :

$$Ay(k) = A\varepsilon(k) + AS\delta(k). \quad (4)$$

À partir de cette équation, on peut obtenir les deux relations suivantes :

$$r(k) = Ay(k), \quad (5)$$

$$r(k) = A\varepsilon(k) + AS\delta(k), \quad (6)$$

où $r(k) = [r_i(k)]_{i=1:l}$.

La relation (5) permet de calculer le vecteur résidu à partir de mesures, et la relation (6) explique l'influence des erreurs de mesure et des défauts. En présence de(s) défaut(s) ($\delta(k) \neq 0$), $r(k)$ est non nul. Dans ce cas, on doit résoudre le problème inverse ; c. à. d. identifier à partir des valeurs des résidus les variables affectées par le(s) défaut(s). Le résidu doit être statistiquement nul pendant le fonctionnement normal du réseau et dévie de zéro dans le cas de présence de défaut. Ainsi, toute déviation affectant le vecteur résidu est un révélateur de présence d'un évènement anormal.

Les valeurs des résidus ne sont pas suffisantes pour isoler le(s) défaut(s), il est nécessaire aussi de connaître l'influence de chaque variable sur les valeurs des résidus ; c. à. d. la connaissance de la relation entre les symptômes et leurs causes. La matrice des signatures des défauts théorique Λ est établie. Elle défini l'influence des défauts sur les valeurs des résidus. C'est une matrice ternaire (contient des valeurs de code ternaire) et est définie comme suit :

$$\begin{aligned} \Lambda_{ij} &= -1 \ \textit{if} \ A_{ij} < 0, \\ \Lambda_{ij} &= \ 0 \ \textit{if} \ A_{ij} = 0, \\ \Lambda_{ij} &= +1 \ \textit{if} \ A_{ij} > 0, \end{aligned} \quad (7)$$

où $i = 1:l$, $j = 1:n$ et $\Lambda_{.j}$ la matrice de signature de défaut affectant les mesures du $j^{ème}$ élément de $y(k)$.

5. Isolation des défauts

5.1 Principe de la méthode.

Cette méthode est basée sur les concepts élémentaires reliés à la théorie des ensembles flous [10], [16], [17]. Un ensemble flou est défini par :

$$B = \{(x, \mu_B(x)) / \ x \in X\}, \quad (8)$$

où $\mu_B(x)$ est la fonction d'appartenance de l'élément x à l'ensemble flou B, à partir de l'ensemble de définition X et prend des valeurs dans l'intervalle [0 1] appelé μ_B. On défini trois ensembles flous (modalités) r_i^-, r_i^0 and r_i^+ correspondent à des résidus « négatif », « nul » et « positif » respectivement. Dans le cas général, on peut écrire $r^g(k) = [r_i^g(k)]_{i=1:l}$, avec $g = 0, -$ ou $+$.

Les fonctions d'appartenance à ces modalités sont définies comme suit :

$$\mu_{r_i^0}(k) = 1 - \frac{(r_i(k)/\tau_i)^\rho}{1 + (r_i(k)/\tau_i)^\rho}, \quad (9)$$

$$\mu_{r_i^+}(k) = \begin{cases} 0 \ \text{if} \ r_i(k) \leq 0 \\ 1 - \mu_{r_i^0}(k) \ \text{if} \ r_i(k) > 0 \end{cases}, \quad (10)$$

$$\mu_{r_i^-}(k) = \begin{cases} 1 - \mu_{r_i^0}(k) \ \text{if} \ r_i(k) \leq 0 \\ 0 \ \text{if} \ r_i(k) > 0 \end{cases}, \quad (11)$$

où $\mu_{r_i^0}(k)$ représente la valeur de la fonction d'appartenance associée à l'ensemble flou r_i^0 à l'instant k, ρ est le facteur de forme qui doit être pair (il doit être pair pour obtenir des formes sigmoïdes) et τ_i est la tollérence[2] associée au résidu $r_i(k)$. On défini à chaque instant k, le vecteur $\mu_{r^g}(k) = [\mu_{r_i^g}(k)]_{i=1:l}$, avec $g_i = 0, +$ ou $-$.

La localisation des défauts est basée sur l'analyse conjointe des résidus et la matrice des signatures de défauts théorique. Elle permet de déduire à partir de la valeur des différents

[1] Degré d'influence réel qui existe entre les variables (capteurs), il peut être plus ou moins significatif selon la dispersion des capteurs sur le réseau de mesure.

[2] Limite d'acceptante des valeurs du résidu (les valeurs acceptées sont celles supposées saines, non affectées par des erreurs). La tolérance est fixée en fonction de la variance du résidu

résidus, la variable affectée par les défauts. On a ainsi une base de règles [19] (règles d'inférence flou) entre symptômes (déviation des résidus) et causes (présence de défaut). Chaque règle comprend une partie prémisse donnée en fonction des différents résidus et une partie liée à la décision. La règle est donc donnée par :

$$\text{si } r_1(k) \text{ is } r_1^{g_1} \text{ and } r_2(k) \text{ is } r_2^{g_2} \text{ and } \ldots r_l(k) \text{ is } r_l^{g_l} \quad \text{alors } \textit{decision}, \tag{12}$$

avec $g_i = 0, +$ ou $-$, $i = 1:l$, (valeurs de g_i sont reliées aux valeurs de la signature du défaut $\Lambda_{.j}$), où

$$\text{decision} \equiv \begin{cases} \text{défaut(s) sur un ou plusieurs capteur(s)} \\ \text{pas de défaut sur le(s) capteur(s)} \end{cases}$$

L'inférence floue (proposition) est donc donnée par (12).
La quantification des différentes inférences floues est effectuée en fonction des valeurs de la fonction d'appartenance (appelé aussi degré d'appartenance). Le coefficient $\mu_{r_i^{g_i}}(k)$ indique la quantification de la proposition. Pour chaque instant et chaque règle, le degré de véracité noté $\mu_f(k)$ d'une conséquence doit être évalué selon la validité des prémisses. Le « et » dans une proposition est un opérateur de conjonction ou d'agrégation, qui permet la combinaison des degrés d'appartenance des différentes propositions. Différents opérateurs de conjonction peuvent être utilisés [11]. Les opérateurs « minimum », « moyenne » et « Mandel », sont en cours d'utilisation [11], [15]. On a remarqué lors de l'utilisation de ces opérateurs que l'opérateur « Mandel » est équivalent à l'opérateur « moyenne » quand, pour un défaut spécifique on a $g_j \in \{-,+\}$. Par conséquent, un autre opérateur (13) est défini. Il augmente la robustesse de la méthode d'isolation appliquée sur les données du réseau de mesure

$$\mu_f(k) = \min\left(\frac{\sum_{i \in \{i / g_i \in \{0\}\}}^{n} \mu_{r_i^{gi}}}{\left| \{i / g_i \in \{0\}\} \right|}, \min\left(\mu_{r_i^{gi}} / g_i \in \{-,+\} \right) \right), \tag{13}$$

où $\left|\{e_i\}\right|$ représente le cardinal de l'ensemble $\{e_i\}$, f désigne le défaut affectant la variable $y(k)$ ($f_j^-(k), f_j^+(k)$ sont des défauts négatif et positif qui peuvent affectés chaque variable $y_j(k)$) et $g_i = 0, -$ ou $+$ selon l'influence du défaut f sur $r_i(k)$. Les résultats d'isolation sont basés sur les degrés de véracité. Ainsi, le défaut le plus vraisemblable est celui décrit par le degré de véracité le plus élevé.

5.2 Persistance temporelle du défaut

La persistance temporelle des symptômes observés est un indicateur important qui permet de distinguer un vrai défaut d'une fausse alarme. Cela signifie qu'on ne doit pas considérer seulement l'amplitude de ces symptômes pour conclure sur la présence du défaut. En effet, quand seulement l'amplitude des défauts est considérée, un faible degré de véracité est systématiquement associé à ce défaut. Mais, la la considération de l'amplitude du défaut et de sa persistance, permet d'obtenir un degré de véracité élevé pour un défaut de faible amplitude, en condition que ce défaut soit persistant.
Le paramètre persistance intervient lors de l'évaluation des degrés d'appartenance des résidus aux différentes modalités. Pour chaque résidu, $r_i(k)$ sont associés trois indicateurs de persistance de défaut $p_i^-(k), p_i^0(k)$ et $p_i^+(k)$. Chaque indicateur correspond au rapport nombre des valeurs du degré d'appartenance dépassant un seuil s_p sur L (largeur d'une fenêtre défini). Les degrés d'appartenance qui considèrent la persistance des défauts sont donnés par le calcul des moyennes entre les degrés d'appartenance (amplitudes des symptômes) et les indicateurs de persistance. Ces fonctions d'appartenance sont données comme suit :

$$\mu p_{r_i^-}(k) = \frac{p_i^-(k) + \mu_{r_i^-}(k)}{2}, \tag{16}$$

$$\mu p_{r_i^0}(k) = \frac{p_i^0(k) + \mu_{r_i^0}(k)}{2}, \tag{17}$$

$$\mu p_{r_i^+}(k) = \frac{p_i^+(k) + \mu_{r_i^+}(k)}{2}. \tag{18}$$

Le poids donné pour chaque élément dans les relations (16), (17) et (18) est de 0.5. Il représente un choix arbitraire (spécifique à notre application). Dans ce cas, plus les symptômes du défaut observé sont persistants, plus les degrés d'appartenance sont élevés. Trois paramètres doivent être pris en considération pour une méthode d'isolation des défauts efficace : La tolérance τ_i associé au résidu, la valeur de s_p et la largeur de la fenêtre L défini ci-dessus. Dans ce papier, on a considéré l'approche défini dans (5.1), en remplaçant les degrés d'appartenance $\mu_{r_i^-}(k), \mu_{r_i^0}(k)$ et $\mu_{r_i^+}(k)$ par les degrés d'appartenance $\mu p_{r_i^-}(k), \mu p_{r_i^0}(k)$ et $\mu p_{r_i^+}(k)$ pour calculer les degrés de véracité correspondant à chaque défaut.

6. Résultats de simulation

Une application de l'approche d'isolation sur l'ensemble des variables de mesure $y(k)$ est présentée dans cette section. Ces variables sont collectées des sites du réseau de mesure de Nantes montré dans la figure 1 (sites entourés par des cercles en pointillés). On défini de la relation (3), le vecteur $x(k) = \left[h_1(k), h_4(k), h_{15}(k), h_{37}(k), v_5(k)\right]$.
Dans une première étape, on établi l'ensemble des modèles linéaire pour ces variables, où leurs paramètres sont identifiés en appliquant la méthode des moindres carrées à l'ensemble $y(k)$ dans le cas de fonctionnement normal du système (non présence de défauts). Cela est effectué en considérant la période du temps pendant laquelle les variables de mesure sont saines (pas de défauts sur les mesures). Les valeurs des résidus sont générées comme défini dans la relation (5) et sont données comme suit :

$$\begin{aligned} r_1(k) &= h_4(k) - 0.99h_1(k) + 0.014, \\ r_2(k) &= h_{37}(k) - 0.32h_1(k) - 0.68h_{15}(k) - 0.007, \\ r_3(k) &= v_5(k) - 0.44h_1(k) + 0.17h_4(k) - 0.28. \end{aligned} \tag{19}$$

On établi ensuite, la matrice des signatures des défauts (20) correspondante aux défauts envisagés.

$$\Lambda = \begin{array}{|c|c|c|c|c|} \hline -1 & +1 & 0 & 0 & 0 \\ \hline -1 & 0 & -1 & +1 & 0 \\ \hline -1 & +1 & 0 & 0 & +1 \\ \hline \end{array} \tag{20}$$

La matrice Λ est utilisée par la suite, pour générer une base de règles d'inférence. On procède dans une deuxième étape à

la « fuzzification » des résidus. On calcule les valeurs des fonctions d'appartenance $\mu p_{r_i^-}$, $\mu p_{r_i^0}$ et $\mu p_{r_i^+}$ pour chaque résidu. On a fixé le seuil s_p à 0.7 (pour minimiser le nombre de fausses alarmes) et la longueur de fenêtre L à 400 qui correspond approximativement à 7 heures (maximum de durée de défaut dans le cas de cette application). Les degrés de véracité de chaque hypothèse d'existence de défaut (correspondent à chaque variable) sont évalués par la relation (13) et quelques-uns sont représentés dans la figure 2

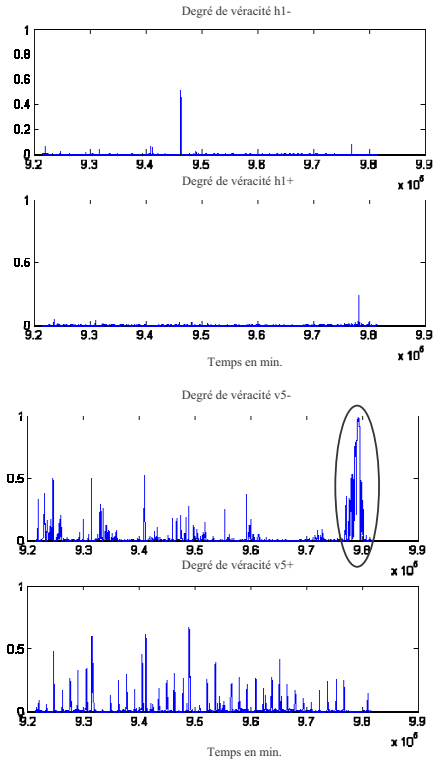


FIG. 2 : Degrés de véracité d'une anomalie + ou – affectant h_1 et v_5

L'analyse de ces degrés de véracité pour l'hypothèse correspondant à chaque variable montre que la défaillance (vrai défaut) est négative et affectant la variable v_5 (où le degré de véracité est élevé (proche de 1) comme montré dans la figure 2). Ce vrai défaut est confirmé par le superviseur (validation par expertise). On a testé la procédure sur d'autres sites dans le même réseau de mesure et pour d'autres défauts réels. Les résultats obtenus confirment son efficacité.

7. Conclusion

Dans ce papier, on a utilisé un ensemble de modèles linéaires établi pour générdr un ensemble de résidus et produire la matrice des signatures des défauts expérimentale. Par la suite, on a appliqué une méthode de diagnostic hors ligne basée sur les concepts du raisonnement flou qui est élaborée et implémentée pour la surveillance du réseau de mesure de la communauté urbaine de Nantes.

Dans le futur, on utilisera les modèles établis à pas de temps d'une heure et d'un jour pour améliorer l'ap`roche d'isolation ; et on complétera ce diagnostic en introduisant les règles de l'expertise et les variables de pluie pour donner les vrais causes de la présence de défauts sur les mesures. Cela nous donnera la possibilité de prendre une décision concernant la solution à proposer pour que le système de surveillance continu sa mission correctement.

Références

[1] J. Ragot, M. Darouach, D. Maquin et G. Bloch, *Validation de données et diagnostic,* traité des nouvelles technologies, série diagnostic et maintenance, Hermès, Paris, 1990.

[2] F. Berrada, S. Bennis et L. Gagnon, « Validation des données hydrométriques par des techniques univariées de filtrage », *Journal Canadien de Génie Civil*, vol. 23, pp. 872-892, 1996.

[3] J. Ma, J. E. Amadi-Echendu et Y. Yan, « Non-redundancy based approach to validate a measuring device », *Proc. IEEE Instrum. and Meas. Technol. Conf.,* Ottawa, pp. 540-543, 1997.

[4] M. Bourouis, L. P. Pascal F. S. Domenech et D. M. K. Al-Gobaisi, « Simulation and data validation in multistage flash desalination plants », *Desalination,* vol. 115, n°. 1, pp. 1-14, 1998.

[5] S. Bennis et N. Kang, « Multivariate technique for validating historical hydrometric data with redundant measurement », *Nordic Hydrology*. vol. 31, n°. 2, pp. 107-126, 2000.

[6] M. Mourad et J. L. B. Krajewski, « Procédure automatisée d'aide à la validation de données environnementales en hydrologie urbaine », *2ème Colloque Automatique et Environnement*, Saint-Etienne, 2001.

[7] G. Olsson and B. Newel, *Modelling, Diagnostic and Control*, IWA Publishing, London, 1999.

[8] D. Maquin, et J. Ragot, « Comparison of gross errors detection methods in process data », *Proc. of the 30th IEEE Conf. on Decision and Control,* Brighton, vol. 3, pp. 2253-2261, 1991.

[9] C. Combastel, S. Gentil et J-P Rognon, « Génération de résidus adaptative pour le diagnostic de la machine asynchrone », *Conf. Internationale Francophone d'Automatique*, Nantes, 2002.

[10] M. Ulieru, « A fuzzy logic based computer assisted fault diagnosis system », *European Journal of Diagnosis and Safety in Automation*, Hermès, vol. 4, n°. 3, pp. 405-441, 1994.

[11] D. Mandel, D. Maquin et J. Ragot, « Process supervision based on fuzzy reasoning. Application to the urban water supply network », 3rd IFAC Workshop on Online Fault Detection and Supervision in the Chemical Processes Industries", Solaize, 1998.

[12] C. K. Mechefske, « Objective machinery fault diagnosis using fuzzy logic », *Mechanical Systems and Signal Processing,* vol. 12, n°. 6, pp. 855-862, 1998.

[13] B. M. Izadi-Zamanabadi et S. Katebi, « Cheap diagnosis using structural modelling and fuzzy-logic-based detection », *Cont. Eng. Practice*, vol. 11, n°. 4, pp. 415-422, 2003.

[14] H. D. Cheng, J. Wang et X. Shi, « Microcalcification detection using fuzzy logic and scale space approaches », *Pattern Recognition*, vol. 37, n°. 2, pp. 363-375, 2004.

[15] D. Maquin et J. Ragot, *diagnostic des systèmes linéaires*, Hermès, Paris, 2000.

[16] K. T. Atanassov, *Intuitionistic fuzzy sets*, Springer Phisica-Verlag, Heidelberg, 1999.

[17] G. Chen, Q. Wie et E. Kerre, « Fuzzy data mining : discovery of fuzzy generalized association rules », *G. Bordogna, G. Pasi*, Springer Phisica-Verlag, pp. 45-66, 2000.

Classification d'EEG pour les Interfaces Cerveau-Machine

Alain RAKOTOMAMONJY, Vincent GUIGUE, Gregory MALLET, Victor ALVARADO

Laboratoire Perception Systèmes et Informations CNRS FRE2645,
INSA de Rouen, 76801 Saint Etienne du Rouvray , France
Alain.Rakotomamonjy@insa-rouen.fr

Résumé – Cet article propose une méthodologie pour la classification de signaux électro-encéphalogrammes (EEG) dans le cadre des interfaces cerveau-machine. L'algorithme que nous proposons est basé sur une approche mettant en oeuvre un mélange de classifieurs SVM linéaire. Chaque SVM est entrainé sur une partie des données d'apprentissage provenant d'une même session d'acquisition des données. Ainsi, le mélange de SVM permet de prendre en compte la variabilité des EEGs lors des differentes sessions de mesures. Combiner à une méthode permettant de sélectionner automatiquement les canaux EEGs pertinents, nous montrons que notre algorithme constitue l'état de l'art pour les données provenant de la compétition BCI 2003.

Abstract – This paper proposes a methodology for classifying EEG signals in the context of Brain-Computer Interfaces. The proposed algorithm is based on a mixture of linear Support Vector Machines. Each SVM has been trained on part of the dataset which has been acquired during a single session. Hence, the mixture of SVM is able to cope with the variability of the EEG signals for different acquisition sessions. Combined with a channel selection algorithm, we show that our result is the state of the art result for the data we used which is part of the BCI 2003 competition .

1 Introduction

Certaines personnes atteintes de maladies neurologiques peuvent être totalement paralysées car elles ne peuvent plus contrôler leurs muscles. Pour ces personnes, seule l'utilisation de leur activité cérébrale et les signaux qui en découle permettent de communiquer. L'objectif de la recherche sur les interfaces cerveau-machine est le développement de la technologie pouvant améliorer cette communication.

Dans cet article, nous nous intéresserons aux interfaces cerveau-machine permettant d'épeler un mot. Ces interfaces sont basées sur l'apparition de potentiels évoqués dans les électro-encéphalogrammes (EEG) en réponse à certains stimuli visuels. Cet interface génère donc des stimuli auquels le patient est supposé répondre, et c'est grâce à ces réponses que la communication cerveau-machine se fait. Le problème réside donc dans la reconnaissance des signaux EEG contenant ou non des potentiels évoqués en réponse aux stimuli.

Notre objectif ici, est de proposer une méthodologie permettant de discriminer ces signaux EEG. Sachant que la difficulté réside dans le faible rapport signal sur bruit et dans la variabilité des types de réponses, notre approche essaye de tenir compte de ces écueils . Ainsi, nous étudions de manière systématique les canaux EEG utiles pour la discrimination et développons une technique de classification utilisant un mélange de classifieurs SVM.

2 Méthodologie

L'objectif de cette étude est de développer un système de reconnaissance des potentiels évoqués permettant de tenir compte de la variabilité des signaux, mais tout d'abord nous allons décrire le principe de l'interface cerveau-machine ainsi que les signaux issus de cet interface.

2.1 L'interface cerveau-machine et les potentiels évoqués

L'interface cerveau-machine que nous utilisons est le *BCI P300 speller* qui est un interface permettant d'épeler des mots. Originalement développé par Farwell et al. [FD88], il est basé sur le protocole suivant : il est demandé à un patient de fixer une matrice 6×6 composée de caractères. Plusieurs stimuli visuels sont alors engendrés par l'illumination dans un ordre aléatoire des 12 lignes et colonnes. Le patient doit alors compter le nombre de fois où la ligne ou la colonne du caractère qu'il souhaite épeler s'illumine ; le fait de compter ainsi génère un potentiel évoqué avec un phénomène apellé P300 (apparaissant 300 ms après le stimulus visuel). D'après ce principe, le caractère désiré est illuminé 2 fois parmi les 12 illuminations de la matrice (ligne et colonne). Notre objectif est donc de reconnaitre parmi les 12 signaux la présence ou non d'un P300 dans les signaux réponses à un stimulus. Une illustration de la matrice d'épellation est présenté dans la figure (1).

Etant donné que les EEG ont un faible rapport signal sur bruit, il est nécessaire de rendre plus robuste le protocole d'épellation à travers un processus répétitif. Ainsi, le protocole suivi est d'épeler 15 fois le même caractère avant de passer au caractère suivant.

Les signaux utilisés dans cet article sont issus d'expériences d'épellation de mot réalisées au sein du Wadsworth Institute [SMH$^+$04] et constituent la base de données d'une compétition sur les interfaces cerveau-machine [BMC$^+$04]. Ces signaux ont été acquis grâce à un scalp de 64 canaux échantillonnés à une cadence de 240 Hz et correspondent à l'épellation par un même individu, lors de 3 différentes sessions, de respective-

FIG. 1 – La matrice d'épellation des mots avec une colonne illuminée

ment 5, 6 et 8 mots.

Les signaux correspondant aux deux premières sessions d'acquisition sont utilisés pour apprendre le modèle de reconnaissance des signaux P300 tandis que la troisième session sera réservée pour le test. Ainsi, la base d'apprentissage est composée de 7560 signaux réponse à un stimulus, décrits sur 64 canaux. Dans la mesure où le phénomène que l'on cherche à reconnaitre apparait autour de 300 ms après le stimulus visuel, nous ne nous intéressons ici qu'à la fenêtre temporelle de 666 *ms* après le stimulus. Par ailleurs, afin de réduire le rapport signal sur bruit, chaque canal a été filtré par un filtre passe-bande d'ordre 8 et de fréquences de coupure de 0.1 Hz et 20 Hz. Le signal est ensuite décimé en accord avec la fréquence de coupure haute. Ainsi, la représentation d'un signal une réponse à un stimulus est un vecteur de 896 éléments constitué par la concaténation des 64 signaux issus des canaux après filtrage et décimation.

Les figures (2) et (3) donnent un exemple de représentations des signaux obtenus pour un canal donné après ce premier prétraitement. La figure (2) représente les variations des échantillons à l'aide de boites à moustaches [Sap90]. Il est possible de distinguer la présence des P300 dans les signaux à travers les variables 7 et 8 correspondant aux échantillons se situant autour de 300 ms. En effet, la valeur médiane de ces variables sont clairement plus élévées en présence des P300. La figure (3) propose une représentation bidimensionnelle des signaux à l'aide de ces 2 variables. Cette nouvelle représentation corrobore le fait que les variables 7 et 8 ont tendance à être plus élevées pour les signaux où le P300 est présent. D'un autre côté, il confirme le fait que le problème de discrimination est difficile car les 2 ensembles de signaux sont fortement superposés.

2.2 Discrimination des signaux EEG

Après ce premier prétraitement, les signaux peuvent être présentés directement à un classifieur qui dans notre cas sera un Séparateur à Vaste Marge (SVM) [CST00]. Cette approche est celle proposée par Kaper et al. [KMG+04], et les résultats obtenus sont très satisfaisants dans la mesure où l'ensemble des mots de la base de test sont correctement reconnus en utilisant seulement 5 séquences d'illuminations. Cependant, cette approche ne permet pas de tenir compte de la variabilité des signaux EEG, ni de celle des potentiels évoqués et des conditions d'acquisitions. Un exemple de cette variabilité est illustré sur la figure (4). Cette figure représente les mêmes signaux que ceux de la figure (3) avec en superposition des signaux sans P300 issus d'une autre session d'acquisition. Cette figure

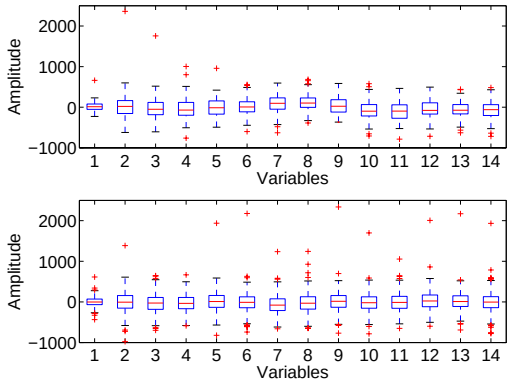


FIG. 2 – Boites à moustaches correspondant aux 14 échantillons d'un canal donné dans le cas de signaux avec P300 (en haut) et sans P300 (en bas).

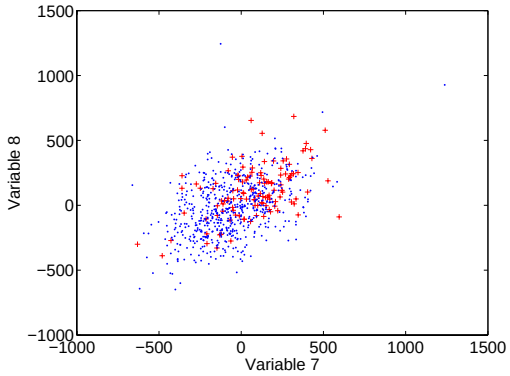


FIG. 3 – Répresentations des signaux avec P300 (en '+' rouge) et sans P300 (en '.' bleu) en fonction des échantillons 7 et 8.

illustre que mélanger des signaux de différentes sessions d'acquisitions aura tendance à mélanger les 2 classes de signaux et donc à rendre le problème de discrimination plus difficile encore.

Nous proposons donc une méthode différente de celle de Kaper et al. [KMG+04] dans la mesure où nous allons developper une fonction de discrimination associée à chaque mot (et donc session d'acquisition) de la base d'apprentissage. Ainsi, la reconnaissance d'un nouveau signal sera effectué par l'ensemble des classifieurs dont les réponses sont ensuite fusionnées pour décider de la classe finale du signal.

Ainsi, pour les signaux $\{x_i, y_i\}_{i=1,\cdots,\ell}$ associés à l'épélation d'un mot donné, la fonction de discrimination SVM est obtenue en suivant les étapes suivantes.

Tout d'abord les signaux d'entrée sont normalisés de sorte qu'ils soient centrés et de variance unitaire. Ensuite un processus de selection de canaux est mis en oeuvre.

Parmi l'ensemble des 64 canaux, Kaper et al. [KMG+04] utilisent un nombre restreint et prédéfini de canaux. Cette approche suppose que l'activité cérébrale associé à la réponse aux stimuli soit localisée sur les mêmes parties du cerveau quelles que soient les conditions d'acquisitions. Dans ce papier, nous nous proposons de tenir compte de cette variabilité en sélectionnant automatiquement les canaux pertinents. Cette sélection se fait grâce à un algorithme d'élimination successive des canaux basé sur un critère de sélection.

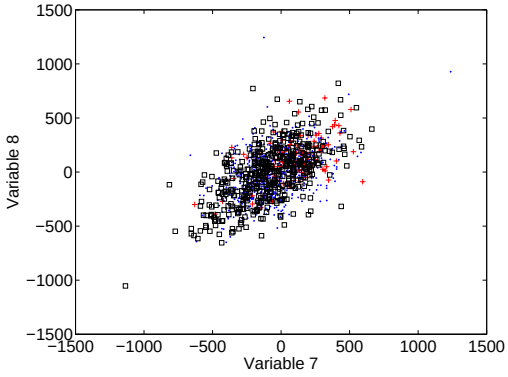


FIG. 4 – Répresentations des signaux avec P300 (en '+' rouge), sans P300 (en '.' bleu) pour lors le l'épélation d'un même en fonction des échantillons 7 et 8. D'autres signaux sans P300 provenant d'une autre session d'acquisitions sont également représentés en carré noir.

La méthodologie suivie est la suivante : un SVM linéaire est appris à partir des exemples de signaux EEG $\{x_i, y_i\}$ relatifs au mot concerné. La performance de ce SVM est evaluée sur les signaux des 10 autres mots de la base d'apprentissage. L'élimination de canal se fait en supprimant un canal de la description des signaux, en ré-apprenant un SVM et en évaluant les performances de ce nouveau classifieur à travers le critère :

$$C_{sel} = \frac{VP}{VP + FP + FN}$$

où VP, FP et FN sont respectivement le taux de vrai positifs, faux positifs et faux negatifs. Le canal qui est éliminé est celui dont la suppression permet de maximiser le critère C_{sel}. Ce processus est ainsi répété tant que tout les canaux ne sont pas éliminés. Ainsi, cet algorithme permet de trier les canaux par ordre de pertinence en fonction du critère C_{sel} en fonction de leur ordre d'élimination. La figure (5) représente pour deux sessions d'acquisitions différentes les variations du critère C_{sel} lors du processus d'élimination en fonction du nombre de canaux. Comme prévu, l'allures de ces courbes sont globalement concaves dénotant ainsi le fait que plusieurs canaux ont tendance à bruiter le problème de discrimination. Par ailleurs, ces courbes corroborent également la variabilité du problème puisque qu'elles montrent clairement qu'en fonction de la session d'acquisition (et donc des signaux utilisés pour l'apprentissage), les performances optimales nécessitent plus ou moins de canaux.

Finalement, la fonction de décision associée à un mot de la base d'apprentissage est obtenue en utilisant les données $\{x_i, y_i\}$ du mot concerné et les N canaux les plus pertinents au sens du critère C_{sel} , ce nombre N étant déterminé de manière *ad hoc* ou automatique en choisissant les canaux qui maximisent le critère.

Chaque fonction de décision $f_k(x)$ (k allant de 1 à 11 ce qui correspond aux 11 mots des sessions d'acquisition de la base d'apprentissage) permet de reconnaitre la présence ou non un phénomène de P300 dans un EEG. Cependant, la reconnaissance d'un caractère nécessite une procédure plus complexe. En effet, un signal x (un ensemble de signaux EEG obtenus sur plusieurs canaux) associé à l'illumination de la ligne ou colonne i de la matrice est présenté après traitement à l'entrée de chaque SVM. Si on dénote respectivement par $S_{Ligne}(i)$ et

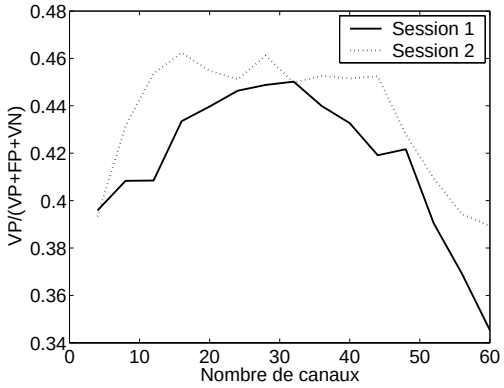


FIG. 5 – Variations du critère C_{sel} pour deux sessions d'acquisition différentes en fonction du nombre de canaux lors du processus d'élimination de canal. Dans le premier cas, le critère est maximisé en utilisant 32 canaux et 16 dans le deuxième cas.

$S_{Col}(i)$ le score associé à la i-ème ligne et la i-ème colonne de la matrice, après chaque illumation de cette ligne ou colonne, ce score est incrémenté en fonction des réponses des différentes fonctions de décision selon :

$$S_{Ligne}(i) = S_{Ligne}(i) + \sum_k f_k(x)$$

Ces scores sont itérativement mis à jour lors des 15 séquences d'illuminations de la matrice disponibles pour la reconnaissance d'une lettre. A la fin des 15 séquences d'illumination, la ligne (respectivement colonne) présentant un score maximal est considérée comme étant la ligne (respectivement colonne) contenant le caractère désiré.

3 Résultats

Cette méthodologie a donc été appliquée aux données de la compétition BCI 2003 [BMC+04] et les résultats sont présentés dans le tableau (1). La première ligne de résultats présente les performances obtenues par un des meilleurs participants à la compétition . L'approche de Bostanov consiste à représenter les signaux EEG sur un plan temps-échelle grâce à une transformée en ondelettes continues, à extraire ensuite des caractéristiques de ce plan et à classer ces caractéristiques par une analyse discriminante linéaire. La deuxième ligne de résultats correspond à un algorithme similaire à celui de Kaper et al [KMG+04] : après avoir choisi a priori des canaux et filtré les signaux en utilisant le même filtre que celui décrit ci-dessus, une frontière de décision unique obtenue par un SVM linéaire est apprise en utilisant l'ensemble des données d'apprentissage (toutes sessions confondues). Les résultats montrent que les algorithmes de Bostanov et al. et Kaper et al. ont une performance équivalente. Par ailleurs, en prenant l'ensemble des canaux pour l'algorithme de Kaper, on constate que les résultats ont tendance à être moins bons. Cela justifie donc le fait que les nombres de canaux doit être effectivement réduit.

Les 4ème et 5ème ligne de résultats prennent en compte la variabilité des signaux et des conditions d'acquisition des données d'apprentissage, à travers une aggrégation de SVM comme décrit dans la partie précédente. Il est interessant de voir que, en utilisant les canaux choisis par Kaper, notre ap-

	Nombres de séquences utilisées							
Algorithmes	1	2	3	4	5	6	7	10
Bostanov [Bos04]	11	5	2	1	1	0	0	0
10 canaux preselectionnées et un SVM unique	14	6	6	0	1	0	0	0
tout les canaux et un SVM unique	14	10	9	5	5	5	1	0
10 canaux preselectionnées et un SVM par session	13	8	3	1	2	0	0	0
tout les canaux et un SVM par session	7	4	3	0	0	0	0	0
4 canaux pertinents et un SVM par session	8	7	4	0	1	0	0	0
10 canaux pertinents et un SVM par session	8	5	5	1	0	1	0	0
26 canaux pertinents et un SVM par session	4	2	0	0	0	0	0	0
30 canaux pertinents et un SVM par session	5	3	0	0	0	0	0	0
Nb de canaux pertinents optimal pour C_{sel} et un SVM par session	4	2	1	0	0	0	0	0

TAB. 1 – Nombre de caractères erronés dans les mots de la base test en fonction du nombre de séquences utilisées et les algorithmes.

proche ne permet pas d'ameliorer les performances de reconnaissances. Sans sélection de canal, tous les caractères de la base de test sont reconnus avec seulement 4 séquences. Par rapport à ceux donnés précedemment, ces résultats démontrent clairement que la prise en compte de la variabilité permet d'améliorer considérablement les résultats et que fixer a priori les canaux pertinents est une approche trop peu flexible vis à vis d'un problème aussi variable que la classification d'EEG.

Les résultats de notre algorithme sont présentés sur les dernières lignes du tableau (1). Après avoir sélectionné les canaux pertinents et effectué l'apprentissage pour chaque SVM associé à une session d'acquitision, nous avons testé les performances en fonction du nombre de canaux pertinents utilisés. Si ce nombre de canaux est fixé a priori, on constate que les performances peuvent être excellente (aucune erreur de reconnaissance pour 3 séquences) tandis que si ce choix est fait de manière automatique c'est-à-dire en choisissant les canaux qui maximisent C_{sel}, l'ensemble des caractères de test sont reconnus sur 4 séquences et le nombre d'erreurs pour 1,2, et 3 séquences est très correct et constitue l'état de l'art sur ce problème.

4 Conclusions

Cet article a présenté une méthodologie pour la classification de signaux EEG issus d'un interface cerveau-machine. Le problème de classification est intimement lié à l'interface et a consiste ici à reconnaître des signaux contenant des P300 en réponse à un stimulus visuel. La difficulté majeure d'un tel problème est la variabilité des données qui est induite par les variations des conditions d'acquisitions et de l'état d'esprit de l'utilisateur. Dans notre approche, nous avons tenu compte de cette variabilité grâce à un mélange de classifieurs et à une sélection automatique des canaux EEG pertinents. Les résultats ont démontré que notre algorithme est très prometteur et nos perspectives vont maintenant nous orienter vers la prise en compte de la variabilité des signaux entre différents utilisateurs.

Remerciements : Ce travail est financé en partie par le programme IST de la Communauté Européenne, avec le réseau d'excellence PASCAL, IST-2002-506778. Cette publication reflète uniquement le point de vue des auteurs.

Références

[BMC+04] B. Blankertz, K-R Müller, G. Curio, T. Vaughan, G. Schalk, J. Wolpaw, A. Schlögl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schröder, and N. Birbaumer. the BCI competition 2003 : Progress and perspectives in detection and discrimination of eeg single trials. *IEEE Trans. Biomed. Eng*, 51(6) :1044–1051, 2004.

[Bos04] V. Bostanov. BCI competition 2003-data sets Ib and IIb : feature extraction from event-related brain potentials with the continuous wavelet transform and the t-value scalogram. *IEEE Transactions on Biomedical Engineering*, 51(6) :1057–1061, 2004.

[CST00] N. Cristianini and J. Shawe-Taylor. *Introduction to Support Vector Machines*. Cambridge Univeristy Press, 2000.

[FD88] L. Farwell and E. Donchin. Talking off the top of your head : toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6) :510–523, 1988.

[KMG+04] M. Kaper, P. Meinicke, U. Grossekathoefer, T. Lingner, and H. Ritter. BCI competition 2003 - dataset IIb : Support vector machines for the P300 speller paradigm. *IEEE Transactions on Biomedical Engineering*, 51(6) :1073– 1076, 2004.

[Sap90] G. Saporta. *Probabilités, Analyse de données et Statistique*. Editions Technip, 1990.

[SMH+04] G. Schalk, D. McFarland, T. Hinterberger, N. Birbaumer, and J. Wolpaw. BCI2000 : a general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6) :1034–1043, 2004.

Nouveau noyau de séquences pour la vérification du locuteur

Jérôme Louradour, Khalid Daoudi

Institut de Recherche en Informatique de Toulouse - CNRS UMR 5505
118, route de Narbonne, 31062 Toulouse cedex 04, France
louradou@irit.fr, daoudi@irit.fr

Résumé – En utilisant la théorie des Espaces de Hilbert à Noyau Reproduisant, nous concevons un nouveau noyau de séquences, qui mesure la similarité entre deux séquences d'observations. Nous appliquons ce noyau à une tâche de vérification du locuteur (campagne d'évaluation NIST 2004). Les résultats montrent qu'incorporer notre nouveau noyau de séquences dans une architecture SVM non seulement fournit des résultats bien meilleurs qu'un classifieur UBM-GMM de base, mais aussi donne de meilleures performances que le classifieur utilisant un noyau GLDS (Generalized Linear Discriminant Sequence kernel). De plus, notre noyau opère dans un espace de plus faible dimension, tout en permettant un large choix de noyaux.

Abstract – Using the framework of Reproducing Kernel Hilbert Spaces, we develop a new sequence kernel that measures similarity between sequences of observations. We then apply it to a text-independent speaker verification task using the NIST 2004 Speaker Recognition Evaluation database. The results show that incorporating our new sequence kernel in an SVM training architecture not only yields performance significantly superior to those of a baseline UBM-GMM classifier but also outperforms the Generalized Linear Discriminant Sequence (GLDS) Kernel classifier. Moreover, our kernel maps to a relatively low dimensional feature space while allowing a large choice for the kernel function.

1 Introduction

La vérification du locuteur est un problème de classification binaire de séquences. Pour ce genre de problème, les méthodes à noyaux, comme les Machines à Vecteurs Supports (SVM), donnent des performances égales ou meilleures que les autres classifieurs. Alors que les algorithmes de base de ces méthodes sont maintenant bien maîtrisés [1], et que leurs propriétés ont été largement étudiées, trouver la façon optimale de représenter les données en entrée de ces algorithmes reste un problème ouvert.

Dans le cas de la vérification du locuteur, la décision est binaire, et les entités à classifier sont des séquences (dans nos expériences, séquences de vecteurs acoustiques). En pratique, les données d'apprentissage du locuteurs (classe +1), en plus d'être bruitées, sont éparpillées au sein même du nuage des données du monde, que l'on assimilera aux données d'apprentissage imposteurs (classe -1). La géométrie du problème est de surcroît complexifiée par le biais éventuel entre les conditions d'apprentissage et celles de test (combinés téléphoniques, canaux de transmissions, fond sonore).

Pour appliquer les SVM à la tâche de vérification, une idée simple consisterait à apprendre pour chaque locuteur une frontière de décision (non linéaire) dans l'espace des vecteurs d'observation, et moyenner les sorties sur chaque vecteur en phase de test, comme il a été fait dans [2] pour une tâche d'identification. Mais à cause des raisons invoquées dans le paragraphe précédent, cette méthode donne de mauvaises performances en vérification du locuteur. Aussi, la finalité du problème étant de classifier des séquences, il est plus adéquat d'optimiser les critères d'apprentissage au niveau des séquences plutôt que des vecteurs.

Plusieurs études récentes ont porté sur la conception de noyaux pour les séquences, de tels noyaux visant à quantifier la similarité entre séquences vis-à-vis d'un problème donné. On peut distinguer trois types d'approches :

- Une première tendance pour comparer deux séquences consiste à entraîner deux modèles statistiques à partir de chacune d'entre elles, et à définir le noyau comme la similarité entre les deux distributions estimées (divergence de Kullback-Leibler [3], affinité de Bhattacharyya [4], distance du χ_2, etc.). Dans notre cas, la courte durée des séquences d'apprentissage ne permet pas d'estimer un modèle de façon suffisamment fiable.
- Une seconde tendance consiste à travailler dans l'espace des scores probabilistes, ou encore l'espace dérivé via le noyau de Fisher [5, 6]. Ces méthodes manquent d'efficacité : elles restent lourdes à mettre en œuvre, et n'améliorent que modérément les performances par rapport aux systèmes fondés sur une décision bayésienne.
- Une autre stratégie [7] revient à apprendre un modèle sur une séquence et à attribuer un score en conséquence sur une autre pour évaluer le noyau entre ces deux séquences. Dans le cas d'une modélisation discriminante par expansion polynomiale, apprendre sur une séquence A et tester sur une séquence B revient au même que d'apprendre sur B et tester sur A. Ceci conduit à l'élaboration d'un noyau symétrique. Cette catégorie de méthodes fournit des résultats très prometteurs : les performances sont meilleures que les systèmes à base de modèles génératifs, pour une complexité moindre.

Dans cette dernière approche, le noyau de séquence se calcule en projetant explicitement les séquences dans un espace de dimension fixe D, via une expansion polynomiale, et en appliquer un produit scalaire après normalisation. En notant d la dimension des vecteurs d'entrée, et k l'ordre de l'expansion polynomiale, la dimension du *feature space* est $D = \frac{(k+d)!}{K!d!}$, Pour des valeurs de k supérieures à 3, elle devient trop grande (*e.g.* $D = 20.475$ pour $d = 24$ et $k = 4$) et rend le problème intraitable. Le choix du noyau est de ce fait très limité. C'est pour combler ce manque de flexibilité que nous avons conçu un nouveau noyau de séquence, en adoptant la même philosophie.

2 Conception du noyau

Cette section introduit le raisonnement théorique qui conduit à la formulation du nouveau noyau de séquence.

2.1 Apprentissage via l'espace de Hilbert à Noyau Reproduisant

Considérons un corpus d'apprentissage $(\tau_n, s_n)_{n=1...N}$, constitué de vecteurs $\tau_n \in \Re^d$ et d'étiquettes binaires $s_n \in \{0, 1\}$ qui valent 0 pour un ensemble fixé de données du monde $C = (c_m)_{m=1...M}$ éparses, et 1 pour les vecteurs d'une séquence $A = (a_t)_{t=1...T_A}$ produite par un locuteur cible donné. On cherche une fonctionnelle $f : \Re^d \to \Re$ qui minimise la fonction objectif :

$$\min_{f \in \mathcal{H}} \sum_{n=1}^{N} \parallel f(\tau_n) - s_n \parallel_2 \qquad (1)$$

où $\mathcal{H}$ est l'espace de recherche de la solution $\hat{f}$.

Une sous-classe importante de problèmes de la forme (1) sont générés par des fonctions noyau définies positives $K(x, y)$. Le sous-espace de fonctions correspondant, $\mathcal{H}_K$, est appelé *espace de Hilbert à noyau reproduisant*. Supposons que K ait une expansion de la forme :

$$K(x, y) = \sum_{i=1}^{D} \gamma_i^2 \Psi_i(x) \Psi_i(y) = \langle \gamma\Psi(x), \gamma\Psi(y) \rangle \qquad (2)$$

avec $\gamma = diag(\gamma_1, ..., \gamma_D)$ et $\Psi : \Re^d \to \Re^D$ une fonction à haute dimension (D peut être infini). Alors tout élément de $\mathcal{H}_K$ est de la forme $f(x) = \sum_{i=1}^{D} \epsilon_i \Psi_i(x)$.

Dans [8], G. Whaba montre que la solution à (1) est de la forme :

$$f(x) = \sum_{n=1}^{N} \omega_n K(x, \tau_n) \qquad (3)$$

Dans la suite, notre *mapping* de séquence pour le calcul du noyau (6,8,9) sera basé sur les fonctions de base $x \mapsto K(x, \tau_n)$. En vue de concevoir un système stable et efficace, il est préférable de concevoir un noyau qui soit indépendant du locuteur cible. C'est pourquoi nous cherchons en fait les solutions sous la forme :

$$f(x) = \sum_{n=1}^{M} \omega_m K(x, c_m) \qquad (4)$$

La résolution des équations normales avec quelques approximations (*cf.* [9] pour plus de détails) donne la solution :

$$\hat{\omega}_A = [\hat{\omega}_1, ..., \hat{\omega}_M]^{\mathrm{T}} = M.{\mathbf{K}_C}^{-2} \overline{\varphi_{\mathbf{C}}}(A) \qquad (5)$$

où $\mathbf{K}_C = (K(c_m, c_n))$ est la matrice de similarité symétrique, et où l'on définit les transformations :

$$\begin{cases} \overline{\varphi_{\mathbf{C}}}(x_1, ..., x_T) = \frac{1}{T} \sum_{t=1}^{T} \varphi_{\mathbf{C}}(x_t) \\ \varphi_{\mathbf{C}}(x) = [K(x, c_1), ..., K(x, c_M)]^{\mathrm{T}} \end{cases} \qquad (6)$$

2.2 Similarité entre deux séquences

Supposons avoir appris à partir d'une séquence A un modèle discriminant de la forme (5). La fonction de similarité entre un vecteur x et la séquence A est donnée par :

$$\hat{f}(x) = \langle \hat{\omega}_A, \varphi_C(x) \rangle \qquad (7)$$

Pour étendre cette mesure à une séquence $B = (b_t)_{t=1...T_B}$, nous moyennons la similarité de chacun de ses vecteurs :

$$\begin{aligned} similarité(B|A) &= \frac{1}{T_B} \sum_{t=1}^{T_B} \hat{f}(b_t) \\ &= M.\overline{\varphi_{\mathbf{C}}}(B)^{\mathrm{T}} {\mathbf{K}_C}^{-2} \overline{\varphi_{\mathbf{C}}}(A) \end{aligned}$$

Au facteur multiplicatif M près, cela conduit à notre nouveau noyau de séquence symétrique :

$$\kappa(A, B) = \langle \overline{\Phi_C}(A), \overline{\Phi_C}(B) \rangle \qquad (8)$$

où l'on définit le *mapping* de séquence :

$$\overline{\Phi_C}(x_1, ...x_T) = {\mathbf{K}_C}^{-1} \overline{\varphi_{\mathbf{C}}}(x_1, ..., x_T) \qquad (9)$$

Dans [9], nous montrons que ce *mapping* est équivalent à une projection, dans l'espace de dimension D défini par Ψ dans (2), de l'expansion moyenne $\frac{1}{T} \sum_{t=1}^{T} \Psi(x_t)$, sur la base des expansions des vecteurs du monde $(\Psi(c_m))_{m=1...M}$.

3 Implémentation du noyau

L'implémentation du noyau de séquences (8) se fait en projetant chaque séquence dans un espace de dimension M, et en calculant un produit scalaire. Ceci est délicat à mettre en œuvre lorsque M est trop grand. C'est pourquoi nous avons supposé en 2.1 que les données imposteurs c_m étaient éparses. Cependant, en pratique, la quantité de données du monde disponibles est très grande. Dans ce qui suit, l'ensemble C est obtenu par quantification vectorielle de données du monde, et les c_m sont des vecteurs *codebook* représentatifs de ces données.

Le but de cette section est de présenter comment appliquer notre noyau de séquence, pour un problème de vérification du locuteur.

3.1 Notion de séquence

Dans notre problématique, une séquence est définie comme un ensemble de vecteurs produits par un même locuteur dans les mêmes conditions d'enregistrement.

Des expériences ont montré que même si une seule séquence était disponible par locuteur, il était inutile de la diviser en plusieurs. Un tel artifice n'améliore pas les performances. Par contre, si l'on dispose de plusieurs séquences provenant de différentes sessions d'enregistrement, alors il est préférable de les conserver distinctes, afin de garder l'information sur la variabilité due aux conditions d'enregistrement.

3.2 Mapping de séquence

Le *mapping* de séquence $\overline{\Phi_C}$ est entièrement déterminé par un ensemble de vecteurs *codebook* C et une fonction K satisfaisant les conditions de Mercer. L'implémentation du *mapping* d'une séquence $X = (x_1, ..., x_T)$ est montrée dans la figure 1.

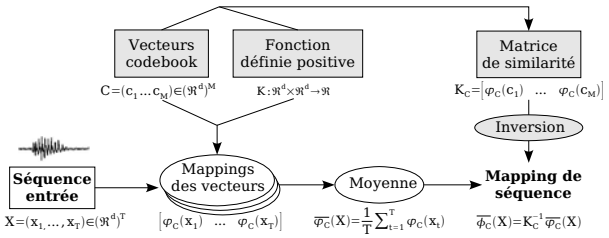


FIG. 1: *Implémentation du mapping de séquences. Les blocs grisés désignent les étapes préliminaires (phase de développement).*

3.3 Apprentissage de modèle discriminant

Une fois le *mapping* défini, nous pouvons pré-calculer les *mappings* des séquences imposteurs d'un corpus du monde, et entraîner un modèle SVM par locuteur cible, avec la procédure *un-contre-tous* schématisée dans la figure 2. Dans nos expériences, Les modèles SVM ont été estimés avec SVM Torch [10]. Il est important de noter que les composantes du *mapping* sont normalisées de façon à avoir une variabilité unitaire sur chaque entrée. La normalisation est donnée par :

$$\overline{\Phi_C}(X) \mapsto \frac{\overline{\Phi_C}(X) - \mu}{\sigma} \tag{10}$$

où μ et σ sont les estimations respectives de la moyenne et l'écart-type, sur le corpus du monde, des vecteurs projetés $\overline{\Phi_C}(X_{imp})$. Cette normalisation est une précaution classique pour les SVMs, qui ne sont pas invariants aux transformations linéaires.

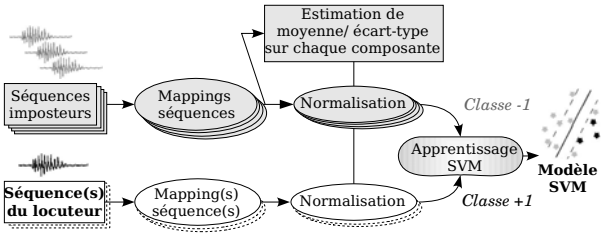


FIG. 2: *Apprentissage de modèle locuteur par SVM*

3.4 Test d'une séquence

Modulo un seuil de décision, la sortie d'une SVM sur une séquence Y est de la forme :

$$score(Y) = \sum_i \alpha_i y_i \langle \overline{\Phi_C}(Y), \overline{\Phi_C}(S_i) \rangle \tag{11}$$

où les (α_i) sont des poids positifs appris, les (S_i) sont les séquences d'apprentissage, et les $y_i = \pm 1$ sont les étiquettes correspondantes ($+1$ pour les séquences du locuteur cible, -1 pour les séquences du monde).

Étant donné la linéarité du produit scalaire, le calcul du score peut être simplifié :

$$score(Y) = \langle \overline{\Phi_C}(Y), \sum_i \alpha_i y_i \overline{\Phi_C}(S_i) \rangle = \langle \overline{\Phi_C}(Y), \Omega_{loc} \rangle \tag{12}$$

Comme dans [7], compacter les séquences support en un seul modèle Ω_{loc} permet d'économiser de l'espace mémoire pour le stockage des modèles locuteur, et de réduire les temps de calcul en phase de test.

Finalement, la décision binaire *locuteur/imposteur* pour une séquence Y est prise en comparant $score(Y)$ à un seuil pré-déterminé.

4 Expériences

4.1 Données de développement et de test

Nous avons appliqué notre nouveau noyau de séquences à la campagne d'évaluation de reconnaissance du locuteur NIST 2004, dans les conditions standards [11]. Dans ces conditions, une séquence contenant plus ou moins deux minutes de parole est disponible pour entraîner chaque modèle locuteur. Pour le corpus du monde, nous avons utilisé un millier de séquences extraites des données de la campagne NIST 2001.

4.2 Pré-traitement

A partir d'une séquence de parole, 12 MFCC sont extraits sur des fenêtres de Hamming de 16 ms, avec un pas régulier de 10 ms. Les 12 coefficients dérivés et la dérivée de l'énergie en échelle logarithmique sont aussi rajoutés. Un détecteur d'activité vocale rejette alors les zones de faible énergie.

Enfin, les vecteurs de dimension 25 ainsi calculés sont normalisés par *feature warping* [12], sur une fenêtre glissante de 300 échantillons. Une telle normalisation a pour but de réduire l'effet combiné du bruit additif et des distortions dues aux canaux de transmission.

4.3 Systèmes de référence

Pour valider le nouveau système que nous avons conçu, nous le comparons à deux systèmes de références à partir de mêmes données de développement et de test.

Le premier, qui a été présenté par l'IRIT à l'évaluation NIST 2004, est un système UBM-GMM [13]. Il estime deux modèles Mélanges de Loi Gaussiennes, pour chacun

des genres homme/femme, en appliquant un algorithme EM sur les données du monde. Chaque modèle locuteur est alors dérivé du modèle du monde correspondant en adaptant les vecteurs moyennes avec un critère MAP. En phase de test, le score d'une séquence est la moyenne des rapports de vraisemblance.

Le second est fondé sur le noyau GLDS décris dans [7]. Le principe est similaire à notre approche, mais le *mapping* de séquence est différent. Il s'agit d'une expansion polynomiale moyenne, suivi d'une normalisation dont les paramètres sont aussi estimés à partir des données du monde. L'ordre polynomial maximal est fixé à 3 et la taille du *mapping* est alors $\frac{(25+3)!}{25!3!} = 3276$.

4.4 Résultats

Les courbes DET relatives aux deux systèmes de référence et à un système basé sur notre nouveau noyau de séquence, sont tracés dans la figure 3 (chaque point de la courbe correspond à un seuil de décision). La fonction de coût de détection considérée, à minimiser, est celle définie par NIST SRE [11], comme somme pondérée des probabilités de faut rejet et de fausse acceptation (*resp.* P_{fr} and P_{fa}) :

$$DCF = (0.1 \times P_{fr}) + (0.9 \times P_{fa}) \qquad (13)$$

Pour notre système, nous avons choisi un noyau polynomial de degré 7 : $K(x, y) = (1+\langle x, y\rangle)^7$. Un noyau gaussien donne des performances similaires. $M = 2048$ vecteurs *codebook* ont été estimés par quantification vectorielle sur les données du monde. Nos expériences ont montré qu'augmenter le nombre M améliorait les performances en général.

On peut voir que notre système donne de meilleures performances que les autres sur toutes les régions de la courbe DET. Des expériences sur l'évaluation NIST 2003, avec exactement les mêmes choix techniques, ont confirmé cette tendance. Par rapport à méthode utilisant un noyau GLDS, notre *mapping* est de dimension inférieure.

5 Conclusion

Nous avons introduit un nouveau noyau de séquences à l'origine d'un nouveau système SVM pour la vérification du locuteur. Toutes les expériences que nous avons menées ont montré que ce système fournit des résultats bien meilleurs qu'un classifieur UBM-GMM. De plus, il donne des performances comparables au classifieur SVM avec noyau GLDS, tout en travaillant dans un espace de dimension plus faible. Aussi, la flexibilité de notre nouveau noyau offre de belles perspectives pour des améliorations futures, et peut être appliqués à d'autres tâches de classification.

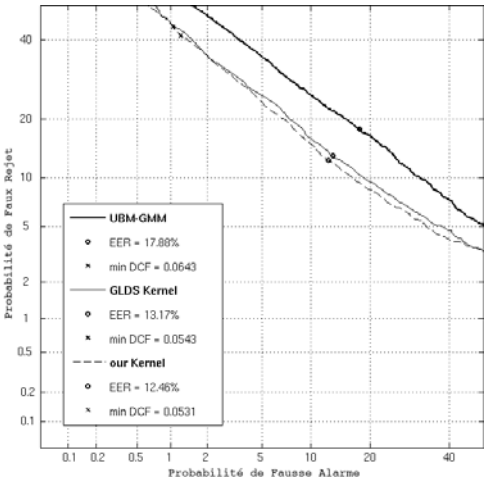


FIG. 3: *courbes DET - Comparaison de la nouvelle approche avec deux systèmes de références. Les cercles indiquent les EER, les croix montrent les points de fonctionnement correspondant au minimum de la fonction coût définie en (13)*

Références

[1] B. Schölkopf and A.J. Smola, *Learning with kernels : Support Vector Machines, regularization, optimization and beyond*, MIT Press, 2001.

[2] M. Schmidt and H. Gish, "Speaker identification via support vector machines," in *Proc. ICASSP*, 1996.

[3] P. Moreno and P. Ho, "A new svm approach to speaker identification and verification using probabilistic distance kernels," in *Proc. Eurospeech*, 2003.

[4] R. Kondor and Jebara T., "A kernel between sets of vectors," in *Proc. ICML*, 2003.

[5] T. Jaakkola and D. Haussler, "Exploiting generative models in discriminative classifiers," *Advances in Neural Information Processing Systems 11*, 1998.

[6] V. Wan and S. Renals, "Speaker verification using sequence discriminant support vector machines," *IEEE Trans. on Speech and Audio Processing*, 2004.

[7] W.M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Proc. ICASSP*, 2002.

[8] G. Wahba, *Applied Mathematics*, vol. 59, chapter Spline Models for Observational Data, CBMS-NSF Regional Conference Series, 1990.

[9] J. Louradour, "a new sequence kernel and its application to speaker verification," Irit research report, www.irit.fr/∼jerome.louradour/papers/, 2005.

[10] R. Collobert and Bengio, "Svmtorch : Support vector machines for large-scale regression problems," *Journal of Learning Machine Research*, 2001.

[11] "Nist speaker recognition 2004 evaluation plan," http://www.nist.gov/speech/tests/spk/2004/.

[12] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, 2001.

[13] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, 2000.

Espaces hybrides couleur – texture adaptés au comptage d'épis de blé

F. COINTAULT[1], D. GUERIN[1]

[1]ENESAD, UMR CPAP, 21, Bd. O de Serres, 21800 QUETIGNY
f.cointault@enesad.fr

Résumé – Une méthode d'analyse d'image texture – couleur basée sur la détermination d'un espace hybride couleur a été développée dans le cadre d'une étude de faisabilité d'un système de comptage (semi)-automatique des épis de blé dans un champ, pour une prédiction précoce de rendement. A cette fin, des analyses couleur et de texture (paramètres d'Haralick) sont utilisées pour permettre une nouvelle représentation des images prises dans un espace spécifique (espace hybride) construit à partir de connaissances a priori sur les images. Les méthodes classiques de classification et de segmentation des images, combinées à des informations morphologiques sur les épis, sont appliquées pour le comptage. Le taux de reconnaissance des épis est compris entre 73% et 85% sur les quelques images prises pendant le stade de floraison. Couplées avec le temps, la luminosité et en accord avec les modèles de développement du blé, les informations générées à partir de l'algorithme implémentés dans cette étude semblent significatives pour pouvoir évaluer le troisième apport d'engrais azotés qui se déroule au début de la floraison.

Abstract *– A colour and texture image analysis method based on colour hybrid space was developed for a feasibility study concerning (semi-)automatic wheatears counting, for yield prediction. To detect ears, textural (Haralick features) and colour analyses are used to give a new representation of the images in a specific space (hybrid space) built with a priori knowledge of the images. Classical methods of image segmentation and classification, combined with morphological information of heads, are applied to the new images for the counting, for which the recognition is included between 73% and 85%. Coupled with weather, lighting, and in accordance with wheat development models, information generated from this algorithm developed in this investigation may be relevant to the evaluation of nitrogen requirements for the third application (head emergence and flowering).*

1. Introduction et contexte

Le rendement est un facteur primordial en agriculture qui permet l'établissement de cartes afin d'adapter certaines pratiques des agriculteurs les années suivantes (apports d'intrants, applications de produits phytosanitaires...), en fonction de la variabilité intra-parcellaire.

Les capteurs de rendements développés sur les moissonneuses-batteuses sont relativement efficaces mais ne fournissent le rendement instantané et en continu que lors de la récolte, avec une précision de 3 à 4%. L'objet de notre recherche est l'étude de faisabilité d'un système ((semi-)automatique) de comptage d'épis de blé par imagerie pour une estimation de rendement : l'utilisation de l'imagerie peut alors offrir des possibilités d'analyses étendues (évaluation précoce du rendement), et est en outre un atout par rapport aux protocoles de dénombrement manuel dont disposent les agronomes. Peu de travaux ont cependant été réalisés sur l'étude de rendement du blé par vision. Le principal projet (analyse essentiellement texture) est l'étude de faisabilité qui s'est développée à l'ENITA de Bordeaux, dans laquelle Germain *et al.* ([1]) réussissent à extraire les épis et à les compter sur une surface assez réduite, dans des conditions de prise d'images idéales (images noir et blanc), sans pour autant exploiter les résultats.

La période d'observation des échantillons est un point crucial mais reste délicate à choisir; néanmoins, la fin de la floraison (mai à juillet) semble être plus appropriée que le gonflement ou l'épiaison pour l'observation et la prise d'images. Dans un souci de réduction des coûts et de facilité d'utilisation, un appareil photo numérique a été préféré à une caméra (figure 1).

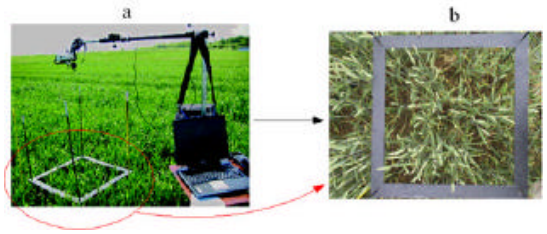


FIG. 1 : (a) : Dispositif expérimental pour l'acquisition d'images au champ ; (b) : Une image type

La variabilité intra-parcellaire nécessite un suivi d'échantillons caractéristiques de la parcelle. La variabilité de l'éclairage et des conditions météorologiques (ombres, vent) demandent l'adaptation de l'appareillage lors de la prise d'images ([2]). Un cadre noir posé au sol (taille 50x50 cm²) sert à délimiter la zone de l'image traitée (figure 1) : il est supprimé par morphologie mathématique.

2. Méthode d'extraction des épis de blé

2.1. Objectif et principe

Pour discerner des épis dans une scène en milieu naturel, l'étude de la couleur s'avère trop sensible à la variabilité de la luminosité sur un échantillon d'observation ([3]). Nous la complétons par une étude de texture qui lui donne plus de robustesse, nécessitant la construction d'un espace hybride qui allie des informations texture obtenues sur des composantes couleur. Son principe repose sur une analyse discriminante qui détermine le pouvoir séparateur des paramètres de texture, évalués sur toutes les composantes couleurs choisies pour leur utilisation dans la reconnaissance d'adventices. La construction de l'espace est tirée de la généralisation de l'espace couleur hybride de Vandenbroucke ([4]).

2.2. Construction de l'espace couleur

L'analyse discriminante est basée sur un échantillon d'apprentissage constitué d'individus (pixels) dont la classe est connue a priori. L'analyse de texture est déroulée sur une fenêtre d'apprentissage autour de ces pixels. Un système à 3 classes est retenu pour différencier {terre, déchets de végétaux, cailloux, feuilles malades}, {feuilles saines et tiges} et {épis}. Le nombre de points d'apprentissage (identique pour chaque classe) est fixé à 20. Pour chaque fenêtre, l'ensemble des composantes couleurs utilisées dans la construction de l'espace hybride peut être regroupé au sein d'un espace de couleurs global formé d'un repère de $N_\pi = 23$ vecteurs :

$$!I-R,G,B,r,g,b,r-g,g-b,\left|\frac{g-b}{r-g}\right|,2g-r-b,H,S,L,L*,a*,b*,u*,v*,I,V1,V2,D,B$$

L'analyse de texture complémentaire repose sur l'étude de $N_{texture} = 7$ paramètres de texture : 2 paramètres statistiques sur les intensités P des pixels (la moyenne μ_v et la covariance cov_v); 5 paramètres d'Haralick obtenus à partir d'une matrice de cooccurrence M_{cooc} (l'énergie (traduite par le second moment angulaire), le moment des différences inverses, l'entropie, la variance et le contraste). Il existe cependant dans la littérature d'autres paramètres (paramètres de textures couleurs) non encore utilisés par manque de temps ([5]). L'analyse de texture doit évaluer la capacité discriminante de $N_{attributs} = N_{texture} \times N_\pi = 161$ attributs, données de textures colorées, nécessitant une normalisation pour la construction des matrices de cooccurrence (valeurs d'intensité de pixel comprises dans l'intervalle [0,N-1]).
La construction de l'espace hybride est itérative : elle sélectionne au pas d'exécution d = {1, …, $N_{attributs}$} l'espace discriminant de dimension d (le meilleur au sens du critère choisi et selon sa construction), constitué de l'espace de dimension d-1 trouvé au pas précédent et d'un nouvel attribut A_d : $E_{hybride}(d) = E_{hybride}(d-1) \times A_d$. Le choix du nouvel espace au rang d est effectué parmi ($N_{attributs} - d + 1$) espaces potentiels. Il est aussi utile de limiter la recherche à des attributs A_d qui ne seront pas trop corrélés (réduire les redondances d'information entre attributs) : $cor(A_d, A_i) < 0.80$. Cette corrélation est calculée sur les attributs normalisés puisqu'elle est calculée pendant l'espace hybride, qui utilise ces mêmes attributs.
Le critère d'arrêt du processus est choisi pour $dim(E_{hybride}) = D$: la dimension D est déterminée en évaluant un critère informationnel basé sur le calcul des matrices de compacité et de séparabilité des classes ([6]).

3. Résultats de la classification et de la segmentation des pixels

Chacun des critères définit des espaces hybrides permettant une classification des pixels d'épis. Pour des raisons de temps de calcul, les critères s'appuyant sur des opérations d'additions comme la trace, moins complexes que des calculs de déterminant (critère de Hotelling ou de Wilks), sont préférés. $E_{hybride}$ a une structure d'espace euclidien ce qui permet ensuite d'utiliser des méthodes de segmentation par mesure de distances, mesures effectuées par les classifieurs à distance minimale déjà définis dans R^D que sont les distances Euclidienne ou de Mahalanobis.
Les résultats obtenus sont tout à fait satisfaisants (figure 2) et il s'avère que la distance de Mahalanobis génère moins d'erreurs de classification (bruits issus des effets de bord et feuilles malades).

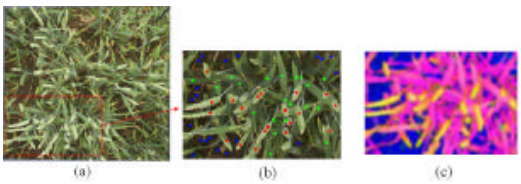


FIG. 2 : (a) : Image sans cadre; (b) : Portion de l'image et sélection des classes; (c) : Image dans le nouvel espace de représentation (en jaune-orange les épis)

La segmentation obtenue est complétée par des algorithmes de comptage d'épis (individualisation) utilisant des outils de morphologie mathématique, en particulier la squelettisation (figure 3).

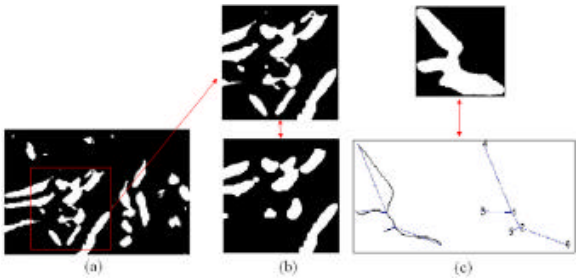


FIG. 3 : (a) : Segmentation par mesure de distance (Mahalanobis); (b) : Exemple d'utilisation de la morphologie mathématique; (c) : Exemple d'utilisation de la squelettisation (dans cet exemple : 23 épis reconnus sur 27 (85% de reconnaissance))

La méthode de construction de l'espace hybride est supervisée et ne permet donc pas une automatisation de la technique actuellement. Néanmoins, en faisant varier les paramètres du système (tableau 1) nous avons observé une certaine redondance pour les espaces hybrides obtenus, qui doit être validée sur d'autres images.

TAB. 1 : Paramètres utilisés

Paramètres	Valeurs testées
Points d'apprentissage	{10; 15; 20; 25; 30; 35; 40}
Dimension de l'espace	{2; 3; 4}
Coefficient de corrélation	{0,7; 0,75; 0,8; 0,85; 0,9}
Niveau de quantification	{8; 16; 32; 64; 128; 256}
Taille des fenêtres (en pixels)	{11; 13; 15; 17; 19; 21}

Les analyses statistiques menées sur seulement 5 images sont ouvertes aux critiques. Les résultats nécessitent d'être confirmés par de nouvelles observations. En outre, l'espace hybride semble indépendant de la luminosité. Au niveau du comptage, les améliorations devraient être possibles par l'optimisation des algorithmes de squelettisation.
Une nouvelle approche pour la classification de pixels serait d'utiliser les réseaux de neurones à réponse binaire. Les modèles de perceptrons multicouches et de réseaux de

neurones à base radiale floue sont deux techniques qui justifient un développement. Ensuite il serait intéressant d'évaluer l'influence, sur notre classification, du choix du système d'apprentissage et de la structure du réseau ([7]).

4. Conclusion

A travers cette étude de faisabilité d'un système de comptage d'épis de blé dans un champ, nous avons étudié les différentes composantes couleur utilisées en agronomie et qui permettent d'extraire des plantes du sol. Afin de minimiser les contraintes dues à la prise d'images en conditions naturelles, nous avons complété cette étude couleur par une étude de texture en construisant un espace d'attributs texture nécessitant 7 paramètres de texture.
L'utilisation de méthodes de segmentation basées sur des mesures de distance fournit des résultats satisfaisants qui doivent être validés sur de nombreuses autres images et évidemment pour différentes variétés de blés. La technique de comptage doit être améliorée (avant 2008) pour obtenir une méthode au champ capable de prédire le rendement de blé en temps-réel.
Les prespectives d'un tel projet sont nombreuses et essentiellement tournées sur la conception d'un outil de développement utilisé par les instituts de recherches pour réaliser des comptages au champ. Cet appareil pourrait également être utilisé pour évaluer les corrélations entre les composantes du rendement et l'état sanitaire de la parcelle, pour compter le nombre de plantes levées par rapport aux plantes semées, pour mesurer l'influence des traitements et la compétition entre la culture et les mauvaises herbes et pour évaluer les bénéfices économiques liés au troisième apport d'engrais.

Références

[1] Germain, C., Rousseau, R. and Grenier, G. 1995. Non destructive counting of wheatears with picture analysis. Image Processing and Its Application 435-439.
[2] El-faki, M., Zhang, N. and Peterson, D. 2000. Factors affecting colour-based weed detection. Transactions of the ASAE **43**(4) 1001-1009.
[3] Lu, J.W. 2003. Segmentation d'images couleur et application à la séparation des oignons et des adventices (Colour image segmentation and application to onions-weeds separation). Ph-D Thesis, University of Burgundy, Dijon, France.
[4] Vandenbroucke, N. 2000. Segmentation d'images couleur par classification de pixels dans des espaces d'attributs colorimétriques adaptés. Application à l'analyse d'images de football (Colour image segmentation by pixel classification in colorimetric attributes spaces. Application to football image analysis). Ph-D Thesis, University of Sciences and Technologies, Lille 1, France.
[5] Foucherot, I., Gouton, P., Devaux, J., Truchetet, F. 2004. New methods for analysing colour texture based on the karhunen-loeve transform and quantification". Pattern Recognition.
[6] Guérin, D., Cointault, F., Gée C. and Guillemin, J.P. 2004. Etude de faisabilité d'un système de comptage d'épis de blé par vision (feasibility study of a wheatears counting system per vision). Revue Traitement du Signal, Vol. 21, N° **5**, pp. 549-560.
[7] Foucher, P., Revollon, P. and Vigouroux, B. 2001. Segmentation d'images couleur par réseaux de neurones : application au domaine végétal (Image segmentation by neural networks : application to plant domain). ORASIS 2001 Poster.

Codage entropique à base de règles de ré-écriture

Hervé JÉGOU[1], Christine GUILLEMOT[2]

[1]IRISA/Université de Rennes 1

[2]IRISA/INRIA
Campus Universitaire de Beaulieu, 35042 RENNES CEDEX
herve.jegou@irisa.fr, christine.guillemot@irisa.fr

Résumé – Des codes de compression sans perte basés sur l'utilisation d'un ensemble de règles de ré-écriture sont introduits et leur performance théorique en compression est analysée. Le formalisme des règles de production permet d'obtenir un degré de liberté important lors de la création de codes. En particulier, il est possible de trouver des codes qui compressent sous le bit ou d'obtenir des codes dont la probabilité marginale bit est uniforme.

Abstract – This article introduces a new set of lossless source codes based on re-writing rules. Theoretical compression efficiency is analytically derived. These production rules allows to obtain an higher degree of freedom in the design of variable length codes. Hence, it is possible to design some codes generating less than 1 bit per encoded symbol or codes such that the marginal bit distribution is uniform.

1 Introduction

Cet article présente des codes de compression définis à partir de règles de production de la forme $a\overline{l} \rightarrow \overline{b}$, où $a \in \mathcal{A}$ représente un symbole de l'alphabet $\mathcal{A}$ et où $\overline{l}, \overline{b}$ sont des petites suites de bits. Contrairement aux algorithmes basés sur un dictionnaire (par ex. [1]), ou aux codes proposés dans [2], les règles de production sont ici pré-définies en accord avec les probabilités de source, ce qui rapproche ces codes des codes à longueur variable (CLV), tels que les codes de Huffman [3]. Ces derniers sont naturellement englobés par le formalisme proposé. Le degré de liberté résultant de l'extension des CLV permet d'obtenir des codes ayant une longueur de description plus courte que celle des codes de Huffman, pour la même complexité d'encodage et de décodage. En effet, l'encodage est effectué en utilisant des automates d'encodage et de décodage. Dans cet article, nous présentons ces codes dans la section 2 ainsi que la méthode de construction des automates d'encodage et de décodage. L'analyse de leur efficacité en compression est fournie dans la section 3. Nous décrivons ensuite, dans la section 4, une méthode de construction qui permet d'obtenir des codes tels que la probabilité marginale des bits émis est de $\frac{1}{2}$, et ceci même si la loi de probabilité n'est pas connue à l'encodage. Ce résultat est prouvé analytiquement. Ceci est intéressant dans la mesure où cette hypothèse est largement utilisée lors du codage de canal et permet d'espérer de meilleures performances en décodage conjoint source/canal.

2 Systèmes de ré-écriture entropiques

Dans la suite les variables aléatoires sont notées en majuscules et les réalisations correspondantes en minuscules. Les ensembles sont écrits en caractères calligraphiques. Le cardinal d'un ensemble $\mathcal{X}$ est noté $|\mathcal{X}|$. Nous définissons $\mathcal{X}^+ = \bigcup_{i=1}^{\infty} \mathcal{X}^i$ et $\mathcal{X}^* = \{\varepsilon\} \cup \mathcal{X}^+$, où ε représente la séquence vide, *i.e.* l'élément neutre pour la concaténation. Ainsi $\mathcal{X}^*$ représente l'ensemble des séquences formées d'éléments de $\mathcal{X}$. Soit $\mathbf{S} \in \mathcal{A}^+$ une séquence de symboles à valeur dans l'alphabet $\mathcal{A} = \{a_1, ...a_i, ...\}$. La longueur d'une telle séquence sera notée $L(\mathbf{S})$. Soit $\mathcal{B} = \{0, 1\}$ l'alphabet binaire. Dans la suite, le train binaire émis est noté $\mathbf{E} = E_1...E_{L(\mathbf{E})} \in \mathcal{B}^*$ et la réalisation correspondante $\mathbf{e}$.

Définition 1: un Système de Ré-écriture Entropique (SRE) est un ensemble $\mathcal{R}$ de règles de production de la forme $r_{i,j}$: $a_i \overline{l}_{i,j} \rightarrow \overline{b}_{i,j}$ où $\overline{l}_{i,j} \in \mathcal{B}^*, \overline{b}_{i,j} \in \mathcal{B}^+$. La règle $r_{i,j}$ représente ici la $j^{ème}$ règle de production associée au symbole a_i et $\mathcal{R}_i$ représente l'ensemble des règles associées à un symbole donné a_i. Le SRE doit en outre vérifier les conditions suivantes :

1. $\forall i,\ |\mathcal{R}_i| \geq 1$,
2. l'ensemble des mots de code $\bigcup_{i=1}^{|\mathcal{A}|} \bigcup_{j=1}^{|\mathcal{R}_i|} \{\overline{b}_{i,j}\}$ est tel qu'aucun mot de code n'est le préfixe d'un autre mot de code, *i.e.* cet ensemble correspond à un code préfixe [4].
3. $\forall i,\ \bigcup_{i=1}^{|\mathcal{R}_i|} \{\overline{l}_{i,j}\}$ est l'ensemble $\{\varepsilon\}$ ou est un code préfixe complet. (*i.e.* tel que l'égalité de Kraft est vérifiée).
4. $\forall i\, \forall i' \neq i, \forall j, j',\ \overline{b}_{i,j} = \overline{l}_{i',j'}$ ou $\overline{b}_{i,j}$ n'est *pas* préfixe de $\overline{l}_{i',j'}$.

La définition 1 ne garantit pas qu'un tel système définisse un système de compression valide. Par exemple, une règle de production $r_{i,j}$ telle que $\overline{b}_{i,j}$ est préfixe de $\overline{l}_{i,j}$ n'est pas valide. Nous supposons dans la suite que le système considéré est valide. Ces règles de production permettent de transformer une suite de symboles $\mathbf{s}$ en une suite de bits $\mathbf{e}$ par application successive de règles de production. Ces règles sont supposées réversibles, *i.e.* inverser les directions des flèches du système de ré-écriture permet de retrouver la séquence $\mathbf{s}$ à partir de $\mathbf{e}$. Un CLV peut être vu comme un SRE. C'est le cas de $\mathcal{C}_1$ dans l'exemple ci-dessous.

Exemple :

$$\mathcal{C}_1 = \begin{cases} r_{1,1} : & a_1 & \rightarrow & 0 \\ r_{2,1} : & a_2 & \rightarrow & 10 \\ r_{3,1} : & a_3 & \rightarrow & 11 \end{cases}$$

$$\mathcal{C}_2 = \begin{cases} r_{1,1} : & a_1 0 & \rightarrow & 10 \\ r_{1,2} : & a_1 1 & \rightarrow & 01 \\ r_{2,1} : & a_2 & \rightarrow & 00 \\ r_{3,1} : & a_3 & \rightarrow & 11 \end{cases} \quad (1)$$

$$\mathcal{C}_3 = \begin{cases} r_{1,1} : & a_1 1 & \rightarrow & 0 \\ r_{1,2} : & a_1 0 & \rightarrow & 10 \\ r_{2,1} : & a_2 & \rightarrow & 110 \\ r_{3,1} : & a_3 & \rightarrow & 111 \end{cases} \quad (2)$$

Le code $\mathcal{C}_2$ est un code à contrainte de suffixe, c'est-à-dire un code tel que $\forall i, j\ \overline{l}_{i,j}$ est suffixe de $\overline{b}_{i,j}$. Ces codes ont été considérés dans [5] dans un contexte de résistance aux erreurs. Les SRE sont un ensemble de codes qui inclut et généralise celui proposé dans [5]. Ainsi, le code $\mathcal{C}_3$ n'est pas un code à contrainte de suffixe. Les SRE sont également représentés sous forme d'arbres, comme présentés sur la figure 1. La structure d'arbre correspond à celle du code préfixe défini par

$$\bigcup_{i=1}^{|\mathcal{A}|} \bigcup_{j=1}^{|\mathcal{R}_i|} \{\overline{b}_{i,j}\}. \quad (3)$$

Les feuilles sont en relation à la fois avec le symbole a_i et avec la suite de bits $\overline{l}_{i,j}$.

Rappelons qu'à l'encodage, les règles de production transforment $\mathbf{s}$ pour obtenir la séquence binaire $\mathbf{e}$. Un segment quelconque de la séquence courante (composée de symbole(s) *et* de bit(s)) peut être ré-écrit s'il existe une règle de production qui a ce segment en entrée. De par la définition des règles proposées, cette entrée est composée d'un symbole et d'un nombre variable de bits. Un SRE n'est valide que s'il n'y a pas d'ambiguïté sur la règle à appliquer. Par conséquent, les règles de productions s'arrêtent lorsqu'il n'y a plus de symbole et que la séquence courante n'est constituée que de bits.

Dans le cas général, comme c'est le cas pour les codes $\mathcal{C}_2$ et $\mathcal{C}_3$, les règles de production qui définissent le SRE ne permettent pas d'encoder la suite $\mathbf{S}$ en passe avant. Ainsi l'encodage est effectué en passe arrière. Pour initier le processus d'encodage, des règles spécifiques doivent être utilisées pour la fin de la séquence, car le dernier symbole peut ne pas être suffisant pour déclencher l'application d'une règle de production. Dans la plupart des cas, les règles à utiliser peuvent être arbitrairement définies en supposant que les bits manquant pour déclencher une règle sont égaux à zéro. Ceci est un choix valide si ces zéros ne permettent pas de déclencher une règle par eux-mêmes. En effet au décodeur un symbole de trop serait alors décodé. Ainsi, le choix du bit de terminaison 0 est valide pour les codes $\mathcal{C}_1$ et $\mathcal{C}_2$ mais ne peut pas être utilisé pour le code $\mathcal{C}_3$, parce que 0 suffit à déclencher la règle $r_{1,1}$. Si un code à contrainte de suffixe est utilisé, les bits de terminaison ne sont, par définition, pas modifiés par une règle subséquente. Il n'est donc pas requis de les transmettre, puisqu'ils sont connus au décodeur par convention. Dans le cas général, ces bits de terminaison doivent être transmis.

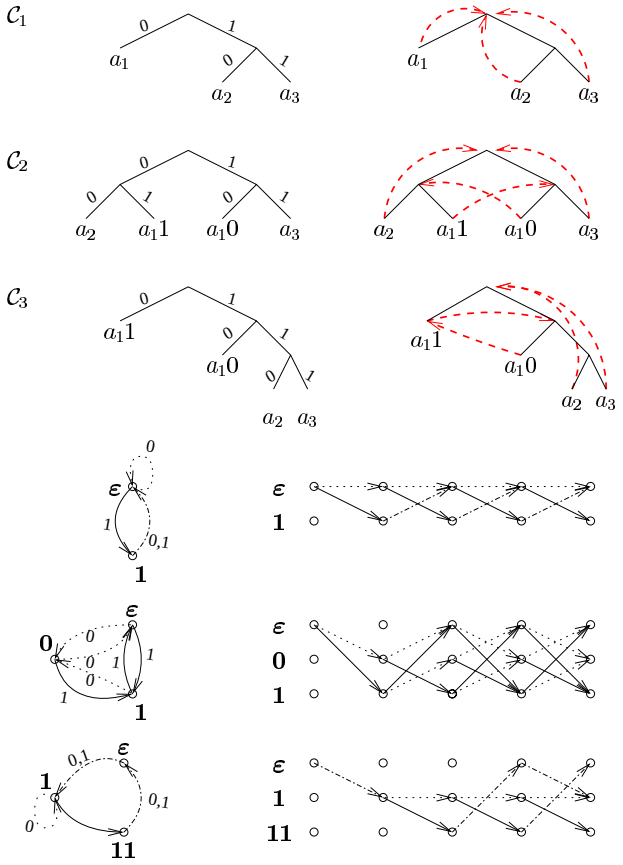


FIG. 1: *Exemples de SRE. De haut en bas puis de gauche à droite : représentation en arbre des codes, transitions générées par les règles, automates de décodage associés, treillis de décodage. Les transitions qui correspondent au bit 0 et au bit 1 sont représentées respectivement en pointillé et en train continu.*

Exemple : la séquence $\mathbf{s}_1 = a_1 a_1 a_1 a_1 a_1$ *est encodée avec le code* $\mathcal{C}_3$ *de la manière suivante (le bit de terminaison est* 1*) :*

$$\begin{array}{ll} r_{1,1} : & \mathbf{s}_1 1 = a_1 a_1 a_1 a_1 \underline{a_1 1} \\ r_{1,2} : & a_1 a_1 a_1 \underline{a_1 0} \\ r_{1,1} : & a_1 a_1 \underline{a_1 1} 0 \\ r_{1,2} : & a_1 \underline{a_1 0} 0 \\ r_{1,1} : & \underline{a_1 1} 00 \\ & \mathbf{e}_1 = 000 \end{array}$$

Remarquez que l'encodage de la séquence $\mathbf{s}_1$ produit un train binaire de longueur 3, ce qui revient à transmettre le symbole a_1 avec 0.6 bit seulement. Cet exemple illustre donc la capacité de ces codes à encoder des séquences avec moins d'un bit. Ceci est impossible dans le cas des codes de Huffman. Pour ces derniers, seule une vectorisation de la source permet d'encoder avec moins d'un bit. On parle alors de codes de Huffman généralisés. L'utilisation de codes généralisés (Huffman ou Tunstall [6]) requiert des tables dont la dimension croît géométriquement avec la longueur de la séquence, ce qui rend leur utilisation impossible.

Au décodage, les règles inverses sont utilisées. L'encodage et le décodage peuvent être facilement implantés sous forme d'automates. Ces automates sont utilisés afin de conserver la mémoire des processus d'encodage et de décodage. Aux états

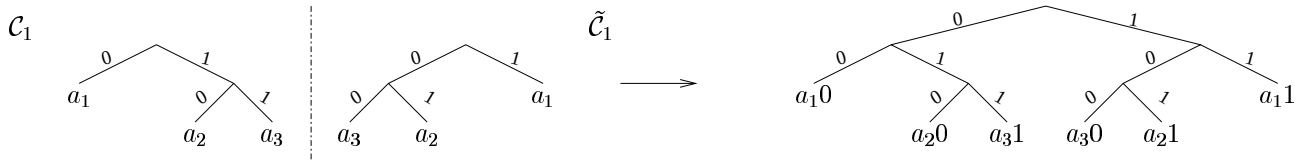


FIG. 2: *Le code primitif $\mathcal{C}_1$, son inverse $\tilde{\mathcal{C}}_1$ et le SRE résultant.*

de ces automates correspondent ainsi les segments de bits utiles pour appliquer les règles de production suivantes. Les automates sont générés à partir de l'ensemble des règles de réécriture. Les transitions sur l'automate d'encodage représentent le processus d'encodage et sont déclenchées par des symboles de $\mathbf{s}$. Les états internes de cet automate sont définis par les segments de longueur variable $\{\overline{l}_{i,j}\}$. Ainsi, l'ensemble des états de l'automate d'encodage est respectivement donné, pour les codes $\mathcal{C}_1$, $\mathcal{C}_2$ et $\mathcal{C}_3$ par $\{\varepsilon\}$, $\{0,1\}$ et $\{0,1\}$. Les états de l'automate de décodage correspondent aux segments de bit qui ont déjà été décodés, mais qui ne suffisent pas à identifier un symbole. Pour les CLV tels que les codes de Huffman, ces noeuds internes correspondent aux noeuds internes de l'arbre du code. Ainsi, pour les codes $\mathcal{C}_1$, $\mathcal{C}_2$ et $\mathcal{C}_3$, l'ensemble des noeuds internes est donné respectivement par $\{\varepsilon, 1\}$, $\{\varepsilon, 0, 1\}$ et $\{\varepsilon, 1, 11\}$.

La représentation graphique des automates de décodage ainsi que les treillis correspondants peuvent être facilement déduits de la représentation en arbre des codes, comme indiqué sur la figure 1. Bien qu'ils n'apparaissent pas sur la figure pour des raisons de clarté, notez que l'ensemble des symboles générés est également associé à chaque bit de transition. Pour les codes $\mathcal{C}_1$ et $\mathcal{C}_2$, il y a au maximum 1 symbole associé à chaque bit de transition. Ce n'est pas le cas pour le code $\mathcal{C}_3$, où la transition partant du l'état $\mathbf{1}$ de l'automate de décodage génère deux symboles a_1, permettant ainsi d'obtenir un coût d'encodage inférieur au bit pour ce SRE.

3 Efficacité en compression

La source $\mathbf{S}$ est supposée être sans mémoire et caractérisée par sa densité de probabilité sur $\mathcal{A}$: $\boldsymbol{\mu} = \{\mathbb{P}(a_1), \dots \mathbb{P}(a_i), \dots\}$. Soit

$$\delta(r_{i,j}) = L(\overline{b}_{i,j}) - L(\overline{l}_{i,j}) \tag{4}$$

le nombre de bits générés par une règle de production donnée $r_{i,j}$. Pour le cas particulier où $\forall i,\ \forall j, j'\ \delta(r_{i,j}) = \delta(r_{i,j'})$, l'espérance E_l de la longueur de description est donnée par

$$E_l = \sum_{a_i \in \mathcal{A}} \mathbb{P}(a_i)\delta_{i,1}. \tag{5}$$

Ainsi, si $\boldsymbol{\mu}_1 = \{0.7, 0.2, 0.1\}$, nous avons $E_l = 1.3$ pour $\mathcal{C}_1$ comme pour $\mathcal{C}_2$. Remarquez que l'entropie de la source est de 1.157.

Soit $R_t : S_t\overline{L}_t \to \overline{B}_t$ la variable aléatoire qui correspond à la règle utilisée pour encoder le symbole S_t. Le processus obtenu à partir de R_t en inversant l'horloge symbole forme une chaîne de Markov invariante $(Z_{t'})$. Les probabilités de cette chaîne de Markov sont obtenues à partir de $\boldsymbol{\mu}$ de la manière suivante :

$$\begin{aligned}\mathbb{P}(Z_{t'}|Z_{t'-1}) &= \mathbb{P}(R_t = r_{i,j}|R_{t+1} = r_{i',j'}) \\ &= \begin{cases} \mathbb{P}(a_i) & \text{si } \overline{l}_{i,j} \text{ est préfixe de } \overline{b}_{i',j'}, \\ 0 & \text{sinon.} \end{cases}\end{aligned} \tag{6}$$

Sous l'hypothèse que $(Z_{t'})$ est irréductible et apériodique, hypothèse dont il est possible de vérifier la validité par une analyse spectrale du noyau de transition de ce processus, la probabilité marginale du processus est obtenue comme l'unique vecteur de probabilité invariant pour le noyau de transition de l'équation 6. C'est-à-dire c'est le seul vecteur propre positif normalisé associé à la valeur propre 1. On en déduit la valeur de $\mathbb{E}(\delta(R_{t'}))$ de l'espérance du nombre de bits produits. Avec le théorème de Cesaro, cette valeur correspond également à la valeur asymptotique de E_l lorsque la taille de la séquence $\mathbf{S}$ tend vers l'infini.

Les performances obtenues en pratique sont très proches de cette performance asymptotique. Il est également possible de calculer de manière exacte de l'espérance du nombre de bits émis pour les séquences finies.

Exemple : pour le code $\mathcal{C}_3$, le calcul de la loi stationnaire mène à $\mathbb{P}(R_t = r_{i,j}) = \{0.412, 0.288, 0.2, 0.1\}$ et donc $E_l = 0.412 \times 0 + 0.288 \times 1 + 0.2 \times 3 + 0.1 \times 3 = 1.188$, contre $E_l = 1.3$ pour le code de Huffman.

4 Construction de codes avec une probabilité bit de $1/2$

Pour construire le code, on part d'un CLV initial

$$\mathcal{H} = \{\overline{b}_{1,1}, \dots \overline{b}_{|\mathcal{A}|,1}\}, \tag{7}$$

par exemple un code de Huffman. On considère alors le CLV $\tilde{\mathcal{H}}$ défini tel que chaque bit de transition de l'arbre de $\tilde{\mathcal{H}}$ est l'inverse du bit de transition correspondant de $\mathcal{H}$, comme décrit sur la figure 2. Le SRE recherché est obtenu en rassemblant ces deux codes avec un nouveau bit de transition. Remarquez que ce SRE est un code à contrainte de suffixe. La démonstration de la propriété annoncée repose sur le calcul de la quantité $f_t = \mathbb{P}(B_t^1 = 0)$:

$$f_t = \sum_{i \in [1..|\mathcal{A}|], j \in [1..2]} \mathbb{P}(R_t = r_{i,j}, B_t^1 = 0) \tag{8}$$

$$= \sum_{i \in [1..|\mathcal{A}|], j=1} \mathbb{P}(S_t = a_i, B_{t+1}^1 = l_{i,1}^{L(l_{i,1})}) \tag{9}$$

$$= \sum_{i \in [1..|\mathcal{A}|], j=1} \mathbb{P}(S_t = a_i, L_t^{L(L_t)} = 0)\, f_{t+1}$$

$$+ \sum_{i \in [1..|\mathcal{A}|], j=1} \mathbb{P}(S_t = a_i, L_t^{L(L_t)} = 1)\,(1 - f_{t+1}). \tag{10}$$

Ainsi cette probabilité vérifie la relation de récurrence

$$f_t = \alpha\, f_{t+1} + (1-\alpha)(1-f_{t+1}), \tag{11}$$

où

$$\alpha = \sum_{a_i \in \mathcal{A}} \mathbb{P}(a_i, l_{i,1}^{L(\overline{l}_{i,1})} = 0) \tag{12}$$

vérifie

$$0 < \alpha < 1, \tag{13}$$

ce qui permet d'appliquer le théorème du point fixe car la fonction $g(x) = \alpha\, x + (1-\alpha)(1-x)$ est contractante. En effet $g'(x) = 2\,\alpha - 1$ et avec l'inégalité stricte de l'équation 13 on déduit $|g'(x)| < 1$. Ce point fixe est 0.5. On en déduit alors la propriété annoncée en remarquant que deux mots de code opposés ont la même probabilité d'occurence, ce qui garantit l'équiprobabilité du 0 et du 1.

Références

[1] J. Ziv and A. Lempel, "A universal algorithm for data compression," *IEEE Trans. Inform. Theory*, vol. 23, pp. 337–343, 1977.

[2] H. Kieffer and E.-H. Yang, "Grammar-based codes: a new class of universal lossless source codes," *IEEE Trans. Inform. Theory*, vol. 46, pp. 737–754, 2000.

[3] D. Huffman, "A method for the construction of minimum redundancy codes," in *Proc. of the IRE*, vol. 40, 1952, pp. 1098–1101.

[4] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley, 1991, ch. 5, pp. 78–124.

[5] H. Jégou and C. Guillemot, "Suffix-constrained codes for progressive and robust data compression: self-multiplexed codes," in *Proc. EUSIPCO'2004*, Sept. 2004, vienna.

[6] B. Tunstall, "Synthesis of noiseless compression codes," *Ph.D Dissertation, Georgia Institute of Technology, Atlanta*, 1967.

Allocation MC-CDMA : augmentation des débits sur les lignes de transmission

Jean-Yves Baudais, Matthieu Crussière
Institut d'électronique et de télécommunications de Rennes — UMR CNRS 6164
INSA, 20 avenue des Buttes de Coësmes, CS 14315, 35043 Rennes Cedex, France
{jean-yves.baudais,matthieu.crussiere}@insa-rennes.fr

Résumé – Dans ce papier nous présentons l'application et la configuration de la technique MC-CDMA pour les liaisons filaires, obtenues à partir de l'étude théorique de la capacité du système émission-réception. L'intérêt de cette technique est de pouvoir regrouper les sous-porteuses de la modulation DMT afin d'accroître les débits de transmission.

Abstract – In this paper we present the application and the configuration of the MC-CDMA technique for wireline communications, obtained from the theoretical study of the transmission system capacity. The interest of this technique is its capability to gather the subcarriers of DMT modulation in order to increase the transmission rates.

1 Introduction

La modulation DMT (*discrete multitone*) est couramment proposée dans les instances de normalisation pour les transmissions sur ligne. Cette modulation permet de profiter pleinement de la connaissance du canal à l'émission et d'appliquer le principe du *water-filling* via les algorithmes d'allocation de l'information ou *bit loading*. L'information est alors distribuée sur les sous-porteuses du signal DMT en fonction de la valeur du rapport entre la puissance du signal et celle du bruit (RSB) propre à chaque sous-porteuse. Cependant, avec une densité spectrale de puissance (DSP) du signal émis limitée, une perte de quantification liée aux ordres discrets de modulation est inévitable. Cette perte peut être limitée en regroupant simplement les sous-porteuses à l'aide de codes d'étalement [1]. La technique MC-CDMA (*multicarrier code division multiple access*) permet ainsi de combiner les énergies des sous-porteuses sans changer la forme d'onde initiale du signal. La composante CDMA ne réalise pas un accès multiple à proprement parler car tous les codes sont utilisés pour la même liaison point-à-point [2, 3]. La perte de quantification est alors minimisée en transmettant l'information non plus sur des sous-porteuses isolées mais sur des groupes de sous-porteuses. De même, les sous-porteuses présentant un trop faible RSB ne pourront pas être utilisées par un système DMT alors qu'elles pourront l'être par un système MC-CDMA. Toute la difficulté est alors de configurer correctement le système MC-CDMA afin de maximiser les débits de transmission.

Alors que la modulation DMT bénéficie d'algorithmes d'allocation optimaux, il n'existe pas aujourd'hui d'équivalent pour les systèmes MC-CDMA. À partir d'une étude théorique de la capacité et des débits réalisables, nous développons une allocation de l'information adaptée au système MC-CDMA qui permet d'obtenir la configuration qui maximise le débit du système de communication. Nous montrons alors qu'à travers des canaux difficiles, le système MC-CDMA peut encore transmettre de l'information contrairement à la technique DMT.

2 Le système de transmission MC-CDMA

Dans un premier temps, nous nous intéressons à un système MC-CDMA classique qui est le résultat de la transformation de Fourier discrète inverse d'un signal CDMA [4]. Le vecteur de symboles complexes $X = {}^t[X_1, \cdots, X_U]$ est étalé par la matrice des codes $C = (c_{i,u})_{0<i\leq k\,,\,0<u\leq U}$. Les symboles X_u appartiennent à l'alphabet des modulations MAQ-2^n, et U codes de longueur k sont utilisés. Nous verrons que l'utilisation de codes orthogonaux issus des matrices de Hadamard de dimension $k \times k$ permet de maximiser la capacité du système de communication. Avant la modulation OFDM (*orthogonal frequency division multiplex*), les échantillons sont les éléments CX, et le système utilise k sous-porteuses.

Les caractéristiques fréquentielles du signal MC-CDMA sont les mêmes que celles d'un signal DMT. La composante OFDM du signal est supposée adaptée au canal de transmission. L'intervalle de garde, le nombre de sous-porteuses, et l'espace entre ces sous-porteuses sont correctement choisis afin d'absorber parfaitement les échos du canal, et de limiter la perte d'efficacité spectrale liée à cet intervalle de garde.

Le signal reçu subit tout d'abord une démodulation OFDM par suppression de l'intervalle de garde et transformation de Fourier discrète. Nous choisissons un système de réception simple où la correction du canal et le désétalement ne sont pas réalisés conjointement. Cela se justifie par le fait que le système optimal utilise des codes orthogonaux et est à pleine charge, comme nous le verrons. Le signal sera alors égalisé à l'aide d'un coefficient par sous-porteuse avant d'être désétalé. La mise en œuvre de la réception reste simple, et l'inversion de matrice n'est

pas nécessaire. Dans le cas général, le vecteur de symboles reçus après démodulation OFDM, désétalement et correction du canal s'écrit $Y = {}^tCGHCX + {}^tCGZ$, avec H la matrice diagonale des coefficients du canal, G la matrice diagonale de correction du canal, et Z le vecteur des échantillons du bruit blanc additif gaussien complexe.

3 Capacité de la forme d'onde MC-CDMA

Nous nous intéressons tout d'abord, non pas à la capacité d'un système de transmission MC-CDMA, mais à la capacité de la forme d'onde MC-CDMA, dans un cadre très général. Nous allons montrer que sous la contrainte d'une DSP du signal limitée, la maximisation de la capacité d'une forme d'onde MC-CDMA conduit à l'utilisation de codes orthogonaux, avec une répartition uniforme de l'énergie entre les codes, et un système à pleine charge. Pour calculer cette capacité nous partons de la définition du signal reçu après démodulation OFDM, soit $HCX+Z$. En considérant le cas de vecteurs spéciaux gaussiens, et avec une connaissance du canal à l'émission et à la réception, l'information mutuelle s'écrit

$$I(X|H,Y|H) = \frac{1}{2}\log\det\left(I + \frac{1}{N_0}HCR_XC^*H^*\right) , \quad (1)$$

avec $R_X = \mathrm{E}\left[XX^*\right]$ la matrice de covariance du vecteur X telle que $\mathrm{E}\left[X_uX_u^*\right] = E_u$. Soit la matrice A telle que $A = I + \frac{1}{N_0}HCR_XC^*H^* = [A_1|\cdots|A_k]$, il vient en utilisant l'inégalité de Hadamard

$$\begin{aligned}|\det(A)| &= \left|\left(\delta_{i,j} + \frac{1}{N_0}\sum_{u=1}^{U} h_i\overline{h}_j c_{i,u}c_{j,u}E_u\right)_{1\le i,j\le k}\right| \\ &\le \prod_{n=1}^{k} ||A_n||_2 . \end{aligned} \quad (2)$$

Une condition nécessaire et suffisante pour maximiser le déterminant est de rendre les vecteurs A_n orthogonaux deux-à-deux, soit $\forall n \neq m\ A_n^*A_m = 0$. Quel que soit $n \neq m$,

$$\begin{aligned}A_n^*A_m = \frac{2\Re(h_n\overline{h}_m)}{N_0}\underbrace{\sum_{u=1}^{U} c_{n,u}c_{m,u}E_u}_{(a)} + \frac{h_n\overline{h}_m}{N_0} \\ \times \sum_{i=1}^{k}\underbrace{\left(\sum_{u=1}^{U} c_{i,u}c_{n,u}E_u \cdot \sum_{v=1}^{U} c_{i,v}c_{m,v}E_v\right)}_{(b)}|h_i|^2 .\end{aligned} \quad (3)$$

Avec des codes orthogonaux de type Walsh-Hadamard, lorsque le système est à pleine charge, *i.e.* $U = k$, et que les énergies sont équiréparties, *i.e.* $E_u = E/k$, (a) est nul car $n \neq m$, et (b) est aussi nul car on ne peut avoir $i = n = m$. La capacité est ainsi maximale lorsque le système est à pleine charge et la répartition de l'énergie uniforme. Cette condition est suffisante mais pas nécessaire, autrement dit, il peut exister d'autres configurations qui maximisent la capacité de la forme d'onde MC-CDMA. On peut également montrer que cette configuration de la forme d'onde MC-CDMA conduit à une capacité égale à la capacité d'une forme d'onde DMT.

4 Technique de détection et correction du canal

Après avoir obtenu quelques résultats sur la capacité d'une forme d'onde MC-CDMA, nous nous intéressons aux techniques de détection linéaire et à la correction du canal, toujours sous la contrainte d'une énergie limitée à l'émission. Le canal peut être compensé soit à l'émission, à l'aide d'une technique dite de pré-égalisation, soit à la réception. Les calculs analytiques de capacité du système émission-réception montrent que la solution optimale est obtenue lorsque la correction est effectuée à la réception, et ce quel que soit le critère choisi (rapport signal sur bruit, distorsion-crête, erreur quadratique moyenne). Ceci s'explique par le fait que c'est la DSP du signal qui est limitée et non sa puissance totale. Nous nous intéressons plus particulièrement au critère de distorsion-crête (ou ZF en anglais) qui permet des débits proches de ceux obtenus avec la détection linéaire optimale (MMSE, *minimum mean square error*). Avec le récepteur ZF les symboles démodulés s'écrivent $Y = kX + {}^tCGZ$, avec G une matrice diagonale telle que $g_{i,i} = 1/h_i$, où h_i est le coefficient du canal sur la sous-porteuse i. La capacité associée à cette technique de détection s'écrit

$$\mathcal{C} = \sum_{u=1}^{U}\log_2\left(1 + \frac{k^2}{\sum_{i=1}^{k}\frac{1}{|h_i|^2}}\,\frac{E_u}{N_0}\right) . \quad (4)$$

La contrainte de DSP se traduit par $\sum_{u=1}^{U} E_u \le E$, et la capacité d'un système DMT s'exprime par

$$\mathcal{C} = \sum_{i=1}^{k}\log_2\left(1 + |h_i|^2\frac{E}{N_0}\right) . \quad (5)$$

5 Allocation MC-CDMA-ZF optimale

Une fois la technique de détection choisie, il nous reste à configurer le système MC-CDMA en sélectionnant les sous-porteuses utiles, et en déterminant la longueur des codes, le nombre de codes, les modulations par code, et la répartition des énergies. L'objectif est alors d'obtenir une allocation de l'information adaptée au système de transmission MC-CDMA et optimale.

Deux approches peuvent être envisagées : maximiser le débit, ou maximiser une marge de bruit sous la contrainte d'un débit cible. Ces deux approches sont connues sous les acronymes BRMP et MMP pour, respectivement, *bit rate maximization problem* et *margin maximization problem* [5]. Dans les deux cas, une formule analytique nous donne la répartition optimale des sous-porteuses, des codes, des

modulations, et de l'énergie, ce qui nous permet d'éviter une recherche exhaustive de la solution optimale.

Nous nous intéressons ici à la maximisation du débit. Afin de spécifier des débits réalisables, nous introduisons la marge de bruit Γ, également appelée le RSB normalisé, et qui mesure la perte introduite par les modulations d'amplitude en quadrature par rapport à la capacité théorique [6]. Le débit réalisable associé à un système MC-CDMA utilisant la détection ZF est alors

$$R = \sum_{u=1}^{U} R_u = \sum_{u=1}^{U} \log_2 \left(1 + \frac{1}{\Gamma} \frac{k^2}{\sum_{i=1}^{k} \frac{1}{|h_i|^2}} \frac{E_u}{N_0} \right) . \quad (6)$$

Cette marge de bruit Γ peut également prendre en compte de gain apporté par le codage de canal [6].

5.1 Choix des sous-porteuses

Les k sous-porteuses peuvent être sélectionnées indépendamment des autres paramètres. Pour maximiser le débit R donné équation (6), il suffit de minimiser $\sum_{i=1}^{k} \frac{1}{|h_i|^2}$, c'est-à-dire de choisir parmi les sous-porteuses disponibles les k sous-porteuses ayant les plus forts coefficients $|h_i|^2$. Une solution simple consiste à réarranger les sous-porteuses par ordre décroissant et de sélectionner les k premières sous-porteuses.

5.2 Maximisation du débit pour une longueur de code donnée

À l'aide des multiplicateurs de Lagrange, on montre simplement que la solution optimale est obtenue pour $U = k$, et $E_u = E/k$, quel que soit k. Soit

$$\mathcal{R} = k \log_2 \left(1 + \frac{1}{\Gamma} \frac{k}{\sum_{i=1}^{k} \frac{1}{|h_i|^2}} \frac{E}{N_0} \right) . \quad (7)$$

Cette solution optimale $\mathcal{R}$ peut conduire à des ordres de modulation non entiers. Nous devons alors introduire une contrainte supplémentaire : $R_u \in \mathbb{N}$.

Théorème 1 *Sous la contrainte* $\sum_{u=1}^{k} E_u \leq E$, *le débit maximal réalisable dans* $\mathbb{N}$ *est obtenue en allouant* $\lfloor \mathcal{R}/k \rfloor + 1$ *bits à* n *codes et* $\lfloor \mathcal{R}/k \rfloor$ *bits aux* $k-n$ *autres codes, avec* $n = \lfloor k\big(2^{\mathcal{R}/k - \lfloor \mathcal{R}/k \rfloor} - 1\big) \rfloor$, *et* $\mathcal{R}$ *le débit optimal dans* $\mathbb{R}$ *donné équation* (7).

Démonstration. Nous devons démontrer que : (i) pour un débit donné la répartition est optimale, c'est-à-dire qu'elle minimise l'énergie émise ; (ii) la valeur de n donnée par le théorème maximise le débit.

Pour démontrer (i), nous cherchons à minimiser l'énergie émise pour un débit cible donné R. Cette énergie s'écrit

$$E = \sum_{u=1}^{k} E_u = \sum_{u=1}^{k} (2^{R_u} - 1) \frac{\Gamma}{k^2} \sum_{i=1}^{k} \frac{N_0}{|h_i|^2} . \quad (8)$$

Il faut alors minimiser $\sum_{u=1}^{k} (2^{R_u} - 1)$ sous la contrainte $R = \sum_{u=1}^{k} R_u$. Posons $R = kq + r$, et $f(0) = (k-r)2^q + r2^{q+1}$ la répartition optimale des bits. La répartition vérifie bien $(k-r)q + r(q+1) = R$. La démonstration se fait par l'absurde en montrant qu'il n'existe pas d'autre répartition f_i différente de $f(0)$ telle que $f_i(a) < f(0)$, $a \geq 1$ et

$$\begin{cases} f_1(a) = (k-r-2)2^q + 2^{q+a} + 2^{q-a} + r2^{q+1} \\ f_2(a) = (k-r-1)2^q + 2^{q+a} + (r-1)2^{q+1} + 2^{q+1-a} \\ f_3(a) = (k-r-1)2^q + 2^{q-a} + (r-1)2^{q+1} + 2^{q+1+a} \\ f_4(a) = (k-r)2^q + (r-2)2^{q+1} + 2^{q+1+a} + 2^{q+1-a} \end{cases}$$

Par identification, $q = \lfloor R/k \rfloor$ et $r = R - \lfloor R/k \rfloor k$. La répartition obtenue avec $f(0)$ minimise l'énergie émise quel que soit le débit R. Ce qui démontre (i). Nous obtenons également le nombre de code, à savoir $U = k$ lorsque $\lfloor R/k \rfloor \neq 0$.

Pour démontrer (ii), nous devons trouver la plus grande valeur de n telle que $\sum_{u=1}^{k} E_u \leq E$; n doit alors vérifier, en posant $\alpha \frac{\Gamma}{k^2} \sum_{i=1}^{k} \frac{N_0}{|h_i|^2} = 1$,

$$E - \sum_{u=1}^{k} E_u = \frac{k}{\alpha}\Big(2^{\mathcal{R}/k} - 1\Big) - \frac{n}{\alpha}\Big(2^{\lfloor \mathcal{R}/k \rfloor + 1} - 1\Big) - \frac{k-n}{\alpha}\Big(2^{\lfloor \mathcal{R}/k \rfloor} - 1\Big) \geq 0 , \quad (9)$$

et

$$E - \sum_{u=1}^{k} E_u = \frac{k}{\alpha}\Big(2^{\mathcal{R}/k} - 1\Big) - \frac{n+1}{\alpha}\Big(2^{\lfloor \mathcal{R}/k \rfloor + 1} - 1\Big) - \frac{k-(n+1)}{\alpha}\Big(2^{\lfloor \mathcal{R}/k \rfloor} - 1\Big) < 0 . \quad (10)$$

Ce qui conduit à $n \leq k(2^{\mathcal{R}/k - \lfloor \mathcal{R}/k \rfloor} - 1) < n+1$. La valeur maximale est alors $n = \lfloor k(2^{\mathcal{R}/k - \lfloor \mathcal{R}/k \rfloor} - 1) \rfloor$. Ce qui démontre (ii) et finit la démonstration du théorème. ■

Nous nous sommes intéressés ici à la maximisation du débit (BRMP), mais le résultat intermédiaire (i) permet également de résoudre le problème de maximisation de la marge (MMP).

5.3 Longueur optimale des codes

Les résultats obtenus au paragraphe 5.2 nous permettent de maximiser le débit quel que soit la longueur des codes. Mais nous n'avons pas encore la longueur optimale des codes. Malheureusement, il n'existe pas de formule analytique pour obtenir cette longueur. En utilisant toutes les constructions possibles des matrices de Hadamard [7], les longueurs possibles des codes sont $\{1, 2, 4n | n \in \mathbb{N}\}$. Il y a donc un nombre limité de cas à comparer. De plus, la fonction $f : k \mapsto \mathcal{R}$ est concave ce qui permet de limiter encore le nombre de cas à comparer.

6 Application au cas multibloc

Le gain en débit apporté par le regroupement des sous-porteuses peut être perdu si les codes sont trop longs car

la distorsion subie par les codes et apportée par le canal devient trop importante. On est alors amené à remplacer le système MC-CDMA initial par un système MC-CDMA multibloc connu sous l'acronyme SS-MC-MA pour *spread spectrum multicarrier multiple access*. Le système en devient plus souple, avec un nombre de sous-porteuses actives qui n'est plus strictement égal à la longueur des codes. Nous nous limitons à une longueur k des codes identique pour tous les blocs. Une généralisation est possible en prenant une longueur des codes propre à chaque bloc, mais il n'existe pas *a priori* de formule analytique donnant la solution optimale .

Les formules analytiques obtenues dans le cas monobloc peuvent encore s'appliquer dans le cas multibloc. En notant K le nombre de blocs et U_s le nombre de codes par bloc, le débit réalisable s'écrit

$$R = \sum_{s=1}^{K} R_s = \sum_{s=1}^{K} \sum_{u=1}^{U_s} R_{u,s} = \sum_{s=1}^{K} \sum_{u=1}^{U_s} \log_2 \left(1 + \frac{1}{\Gamma} \frac{k^2}{\sum_{i=1}^{k} \frac{1}{|h_{i,s}|^2}} \frac{E_{u,s}}{N_0} \right) . \qquad (11)$$

Il reste cependant à répartir les sous-porteuses entre les blocs. La solution optimale consiste à choisir les sous-porteuses de façon à maximiser les débits bloc par bloc. Il suffit alors d'ordonner les sous-porteuses par ordre décroissant de l'amplitude de leur coefficient, puis d'attribuer les k premières sous-porteuses au premier bloc, les k suivantes au deuxième bloc, *etc.* Ce résultat se démontre simplement en maximisant une somme de fonctions logarithmiques.

La figure 1 donne un exemple de débits réalisables obtenus sur une ligne téléphonique standard de 0,4 mm de diamètre et supposée homogène [8], en fonction de la longueur de cette ligne. La DSP du signal émis est limitée à -39 dBm, celle du bruit de fond en réception est de -140 dBm, et le point de fonctionnement correspond à $\Gamma = 6{,}06$ dB. Le nombre de sous-porteuses utilisables est limité à 220, et la modulation maximale est la MAQ-2^{15}. La configuration du système MC-CDMA multibloc est telle que pour chaque point de mesure, la longueur optimale des codes a été choisie. Quelle que soit la longueur de la ligne, la solution MC-CDMA multibloc permet des débits de transmission plus élevés que ceux obtenus avec la solution DMT, ce qui permet également d'augmenter la portée des systèmes de communication. Ce gain en débit, et en portée, est principalement apporté par les sous-porteuses supplémentaires que le système MC-CDMA peut exploiter contrairement au système DMT. Des résultats similaires sont obtenus sur les canaux PLC (*power line communication*) [9], dans un contexte point-à-multipoint.

7 Conclusion

L'étude théorique d'un système de transmission MC-CDMA monobloc et multibloc nous a permis d'obtenir la configuration optimale de ce système et de développer une procédure d'allocation simple. Les résultats de simulation montrent tout l'intérêt de ce système par rapport à un système DMT, en permettant des débits de transmission plus élevés, et en augmentant la portée des communications.

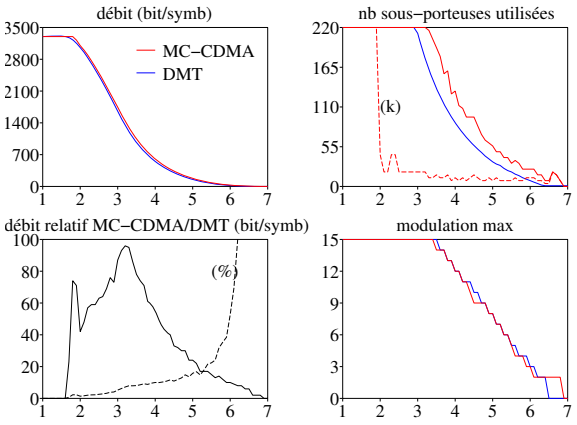


FIG. 1 – Comparaison des systèmes MC-CDMA multibloc et DMT en fonction de la longueur de la ligne ADSL en kilomètre.

References

[1] O. ISSON, J-M. BROSSIER et D. MESTDAGH, "Multi-carrier bit-rate improvement by carrier merging". *Electronics Letters*, vol. 38, n°19, pages 1134–1135, septembre 2002.

[2] S. MALLIER, F. NOUVEL, J-Y. BAUDAIS, D. GARDAN et A. ZEDDAM, "Multicarrier CDMA over lines — Comparison of performances with the ADSL system". In *IEEE International Workshop on Electronic Design, Test and Applications*, pages 450–452, janvier 2002.

[3] J-Y. BAUDAIS, "Amélioration de la robustesse du système ADSL en présence de brouilleurs : utilisation des techniques MC-CDMA". In *Colloque GRETSI, Groupe de recherche et d'étude de traitement du signal*, septembre 2003.

[4] N. YEE, J-P. LINNARTZ et G. FETTWEIS, "Multi-carrier CDMA in indoor wireless radio networks". In *IEEE Personal, Indoor and Mobile Radio Communications Symposium*, pages 109–113, septembre 1993.

[5] J. CAMPELLO, "Pratical bit loading for DMT". In *IEEE International Conference on Communications*, pages 801–805, 1999.

[6] J.M. CIOFFI, "A multicarrier primer". Rapport, *ANSI T1E1.4/91–157, Committee contribution*, 1991.

[7] A. S. HEDAYAT, N. J. A. SLOANE et John STUFKEN, *Orthogonal Arrays: Theory and Applications*, "7". Springer-Verlag, New York, 1999.

[8] F. MOULIN, *Étude de l'impact des bruits implusifs sur la qualité de transmission d'une liaison ADSL*. Thèse de Doctorat, Institut national des sciences appliquées de Rennes, France, 2001.

[9] M. CRUSSIÈRE, J-Y. BAUDAIS et J-F. HÉLARD, "Robust and high-bit rate communications over PLC channels: A Bit-loading multi-carrier spread-spectrum solution". In *International Symposium on Power-Line Communications and Its Applications*, (Vancouver, Canada), avril 2005.

Égalisation fréquentielle sous-optimale d'un système mono ou multiporteuse

Carine SIMON[1], Sylvie MAYRARGUE[2], Juan José DAÑOBEITIA CANALES[1]

[1]UTM, CSIC. Passeig Maritim de la Barceloneta, 37. 08003 Barcelona. ESPAGNE

[2]CEA-LETI. 17, rue des Martyrs. 38000 Grenoble FRANCE

csimon@utm.csic.es, sylvie.mayrargue@cea.fr, jjdanobeitia@utm.csic.es

Résumé – Une méthode itérative d'estimation de symboles peu complexe, dans le cas de systèmes mono ou multi-porteuses avec préfixe cyclique est présentée. On profite de la structure particulière du canal convolutif inconnu ainsi que l'appartenance des symboles à un alphabet fini. La technique présentée est une technique de minimisation au sens des moindres carrés puis de projection. Elle s'avère peu coûteuse et efficace dans les problèmes de détection et d'estimation, même dans le cas de signaux bruités.

Abstract – We present a low complexity iterative method for estimation of symbols in mono or multicarrier systems where cyclic prefixes are combined to data symbols. We take advantage of the special structure of unknown convolutive channels of such systems and of the fact that the symbols belong to a finite alphabet. The presented technique is a minimum mean squared technique followed by a projection. Even in a noisy environment, it gives good results for detection and estimation.

1 Introduction

Parmi les techniques de transmission, celles utilisant des préfixes cycliques ont soulevé un intérêt considérable, en particulier dans le contexte des réseaux locaux mobiles ou la diffusion vidéo. En effet, utiliser des systèmes dans lesquels les symboles de données sont combinés avec des préfixes cycliques (CP) plus longs que le canal à réponse impulsionnelle finie est un moyen efficace d'éviter les problèmes d'interférence entre symboles. Le remplissage par symboles connus (ou « known symbol padding ») est une structure équivalente où le CP est connu. C'est cette dernière qui sera utilisée dans cette étude mais pour l'utilisation d'un préfixe inconnu, il suffit de changer la méthode d'initialisation de l'estimation du canal.

Dans cette communication, on part d'un système de transmission par blocs monoporteuse avec un canal multichemin stationnaire qui est supposé convolutif et inconnu. Ce modèle s'étend directement à un système de transmission par bloc multiporteuse, de type OFDM (« orthogonal frequency decision multiplexing »).

Dans ce contexte, on étudie une technique itérative d'estimation de symboles qui utilise la structure cyclique spécifique de la transformée de Fourier rapide (FFT) du canal convolutif ainsi qu'une projection sur l'alphabet fini BPSK (« Binary Phase Shift Keying ») auquel appartiennent les données.

Des détails sur le modèle étudié sont présentés au deuxième paragraphe. Le troisième paragraphe explique la méthode proposée, partant de [2] et l'améliorant. Plusieurs manières d'utiliser au mieux la structure de notre système et ainsi d'augmenter la qualité des résultats sont proposées au quatrième paragraphe. Avant de conclure, le paragraphe 5 résume l'algorithme puis le sixième paragraphe montre un certain nombre de simulations, comparant notre technique à celle proposée par [2].

2 Modèle

Le canal de transmission convolutif est d'ordre L : $h = [h_1, \ldots, h_L]^T$ où ^{T} représente la transposée. La séquence transmise est de la forme $\mathbf{x} = \left[\mathbf{t}\,\mathbf{s}^1\,\mathbf{t}\,\ldots\,\mathbf{s}^V\,\mathbf{t}\right]$ où $\mathbf{t}$ est le préfixe cyclique de taille M et $\mathbf{s}^k$ est le $k^{\text{ème}}$bloc de données pour $k = 1, \ldots, V : \mathbf{s}^k = \left[s_{kB+1}, \ldots, s_{k(B+1)}\right] : B \times 1$. Les symboles sont de type BPSK. Pour traiter le cas multiporteuse, les symboles transmis se trouvent dans le domaine fréquentiel.

On pose $P = B + M$. Les séquences d'observation au récepteur sont de la forme : $y_i = \sum_{l=0}^{L} h_l x_{i-l} + n_i$ où n_i est un bruit considéré blanc et Gaussien. On peut réécrire le modèle de façon matricielle en définissant $\mathbf{y}^n = [y_{M+(n-1)P+1}, \ldots, y_{M+nP}]$. En omettant le premier $\mathbf{t}$, on pose encore :

$$\begin{aligned} Y &= [(\mathbf{y}^1)^T \ldots (\mathbf{y}^V)^T] : P \times V \\ X &= \begin{bmatrix} (\mathbf{s}^1)^T & \ldots & (\mathbf{s}^V)^T \\ \mathbf{t} & \ldots & \mathbf{t} \end{bmatrix} = \begin{bmatrix} S \\ t_V \end{bmatrix} : P \times V \end{aligned} \tag{1}$$

On définit la matrice du canal $P \times P$ H_c circulante Töplitz :

$$H_c = \begin{bmatrix} h_0 & 0 & \ldots & & 0 & h_L & \ldots & h_2 & h_1 \\ h_1 & h_0 & 0 & \ldots & & 0 & h_L & \ldots & h_2 \\ \vdots & & \ddots & \ddots & & & \ddots & \ddots & \vdots \\ h_L & \ldots & & h_0 & 0 & \ldots & & 0 & h_L \\ 0 & h_L & \ldots & & h_0 & 0 & \ldots & & 0 \\ \vdots & & \ddots & \ddots & & & \ddots & \ddots & \vdots \\ 0 & & \ldots & & 0 & h_L & \ldots & & h_0 \end{bmatrix} \tag{2}$$

En définissant le bruit N de la même manière que Y (équation (1)), le modèle peut être réécrit :

$$Y = H_c X + N \tag{3}$$

Comme la matrice du canal est circulante, elle devient diagonale dans la base vectorielle de Fourier : $H_f = \mathcal{F}_P H_c \mathcal{I}_P$ où

$(\mathcal{F}_P)_{kn} = P^{-1/2}e^{-\frac{2i\pi kn}{P}}$ et $(\mathcal{I}_P)_{kn} = P^{-1/2}e^{\frac{2i\pi kn}{P}}$. On a alors $Y = \mathcal{I}_P H_f \mathcal{F}_P X + N$ ou encore

$$Y_f = H_f X_f + N_f \tag{4}$$

avec $A_f = \mathcal{F}_P A$ pour $A = X, Y, N$.

3 Méthode

Lorsque le bruit additif est blanc et Gaussien la méthode de maximum de vraisemblance est optimale ; elle consiste en la minimisation conjointe au sens des moindres carrés suivante :

$$\min_{\mathbf{x},\mathbf{h}} \|\mathbf{y} - \mathbf{h}\mathbf{x}\|^2 \tag{5}$$

Cette méthode étant très coûteuse, on utilise souvent des techniques sous-optimales. La méthode des moindres carrés itérative avec projection (ILSP ou « Iterative Least Square with Projection », cf [1] par exemple) consiste à minimiser cette fonction alternativement par rapport au canal et par rapport aux symboles.

Pour notre application, on utilisera le fait que les données appartiennent à un alphabet fini en projetant les symboles estimés sur l'alphabet BPSK après leur estimation.

3.1 Estimation du canal

Comme la FFT est bijective, la minimisation par rapport à $\mathbf{h}$ est équivalente à la minimisation suivante : $\min_{H_f} \|Y_f - H_f X_f\|^2$. Et comme H_f est diagonal, on peut découpler ce problème en :

$$\min_{H_f(j,j)} \|Y_f(j,:) - H_f(j,j)X_f(j,:)\|^2 \qquad \forall j \tag{6}$$

Ici et dans ce qui suit, la norme 2 est utilisée pour les vecteurs, la norme de Fröbenius pour les matrices et $A(j,:)$ représente la j$^{\text{ème}}$ligne de toute matrice A.

[2] utilisent directement (6) et obtiennent l'estimation suivante du canal : $\hat{H_f}^{-1}(j,j) = X_f(j,:)Y_f(j,:)'[Y_f(j,:)Y_f(j,:)']^{-1}$ où $'$ dénote le transconjugué.

On peut cependant affiner cette estimation en utilisant la structure particulière de H_f. Si H_v est le vecteur des éléments diagonaux de H_f, on peut montrer que

$$H_v = C\mathbf{h} \tag{7}$$

où les éléments de C (de taille $P \times (L+1)$) sont $C_{nm} = e^{-\frac{2i\pi m(n-1)}{P}}$. [4] ont également utilisé cette propriété mais ensuite une technique d'EM (expectation maximisation) bien plus coûteuse que celle que nous allons développer ci-dessous.

On calcule la norme dans (6) :

$$\begin{array}{l} \|Y_f(j,:) - H_f(j,j)X_f(j,:)\|^2 \\ = Y_f(j,:)Y_f(j,:)' - C(j,:)\mathbf{h}X_f(j,:)Y_f(j,:)' \\ \quad -Y_f(j,:)\mathbf{h}'C(j,:)'X_f(j,:)' \\ \quad +C(j,:)\mathbf{h}X_f(j,:)X_f(j,:)'C(j,:)'\mathbf{h}' \end{array} \tag{8}$$

et sa minimisation par rapport à h_m $\forall m$,

$$\begin{array}{l} \frac{\partial}{\partial h_m'}\|Y_f(j,:) - H_f(j,j)X_f(j,:)\|^2 \\ = -Y_f(j,:)C(j,m)'X_f(j,:)' \\ \quad +C(j,:)\mathbf{h}C(j,m)'X_f(j,:)X_f(j,:)' \end{array} \tag{9}$$

En vectorisant ce résultat, on obtient :

$$\hat{\mathbf{h}} = \left(C'\text{diag}(\langle X_f, X_f\rangle)C\right)^{-1} C'\langle Y_f, X_f\rangle \tag{10}$$

où $\langle A, B\rangle = \text{diag}(AB')$ et, suivant les notations de Matlab, si V est une matrice $\text{diag}(V)$ est le vecteur de ses éléments diagonaux ; inversement, si V est un vecteur, $\text{diag}(V)$ est une matrice diagonale avec les éléments de V sur sa diagonale principale.

Comme $C'\text{diag}(\langle X_f, X_f\rangle)C$ est une matrice hermitienne Töplitz, on note qu'elle est aisément inversible.

3.2 Estimation des symboles

On pose $\mathcal{I}_{BP} = \mathcal{I}_P(1:B,:)$. En utilisant la caractère diagonal de H_f, on obtient l'estimation des symboles suivante :

$$\hat{S} = \mathcal{I}_{BP}\hat{H}_f^{-1}Y_f \tag{11}$$

La dernière étape est une projection sur l'alphabet fini BPSK.

4 Améliorations

Dans cette partie, quelques idées sont proposées afin d'améliorer cette méthode, utilisant au mieux les données dont nous disposons et estimant la variance du bruit.

4.1 Estimation de la variance du bruit

On suggère ici de remplacer l'estimation par forçage à zéro (11) par une estimation des symboles par minimisation au sens des moindres carrés. Pour cela, on suppose que le bruit blanc Gaussien est de variance constante σ_n^2. On peut montrer que l'estimée de cette variance $\hat{\sigma}_n^2 = \overline{\sigma}_y^2 - \|\hat{\mathbf{h}}\|^2$ où $\overline{\sigma}_y^2 = \sum_{k=1}^{V} \text{var}(\mathbf{y}^k)$. En effet, lors des simulations, on s'est aperçu qu'en deux itérations, on obtient une bonne estimation du canal. L'estimation de σ_n^2 est donc suffisamment précise dès la deuxième itération pour pouvoir être utilisée.

La variance du bruit reste inchangée par FFT. On modifie alors (11) de la manière suivante :

$$\hat{S} = \mathcal{I}_{BP}(\hat{H}_f'\hat{H}_f + \hat{\sigma}_n^2 I)^{-1}\hat{H}_f'Y_f \tag{12}$$

Cette idée est non seulement très attractive mais en plus très peu coûteuse : d'une part, on a besoin de calculer $\overline{\sigma}_y^2$ une seule fois et d'autre part, la matrice à inverser dans (12) reste diagonale.

4.2 Parties réelle et imaginaire

Comme les données sont réelles, on peut séparer le système d'équations (3) en ses parties réelle et imaginaire :

$$\begin{bmatrix} Y_{\mathcal{R}} \\ Y_{\mathcal{I}} \end{bmatrix} = \begin{bmatrix} H_{\mathcal{R}} \\ H_{\mathcal{I}} \end{bmatrix} X + \begin{bmatrix} N_{\mathcal{R}} \\ N_{\mathcal{I}} \end{bmatrix} \tag{13}$$

L'avantage d'utiliser ce système d'équations plutôt que (3) est qu'on a ainsi deux fois plus de degrés de liberté qu'au départ. Si on peut aisément montrer que (13) donne une estimation du canal équivalente à (3), on verra dans les simulations que l'utilisation de cette structure mène à une amélioration notable de l'estimation des symboles. Cette même idée est utilisée dans le domaine fréquentiel.

4.3 Réduction du bruit ?

Comme on connaît le CP et son estimée, il est tentant de réduire le bruit en déduisant sa partie due au CP. On va malheureusement montrer que ce n'est pas possible. De la même manière qu'on a estimé les symboles de données, on peut estimer la séquence d'entraînement (éq, 12) : $\hat{t}_V = \mathcal{I}_{MP}\hat{H}_f^{-1}Y_f$ où $\mathcal{I}_{MP}$ représente les M premières lignes de $\mathcal{I}_P$. Les définitions suivantes nous seront également nécessaires :

$$\mathcal{F}_{PB} = \mathcal{F}_P(:, 1:B) \qquad \mathcal{F}_{PM} = \mathcal{F}_P(:, B+1:P)$$
$$S_f = \mathcal{F}_{PB}S \qquad t_{Vf} = \mathcal{F}_{PM}t_V$$

En supposant que $\hat{H}_f \simeq H_f$, on a :

$$\begin{aligned} \varepsilon_t &\stackrel{\text{def}}{=} \hat{t}_V - t_V \\ &= \mathcal{I}_{MP}H_f^{-1}(Y_f - H_f X_f) \\ &= \mathcal{I}_{MP}H_f^{-1}N_f \end{aligned} \tag{14}$$

Il semble donc naturel de remplacer Y_f par sa version moins bruitée : $Z_f = Y_f - H_f\mathcal{F}_{PM}\varepsilon_t$ dans (11). Cependant,

$$\begin{aligned} Z_f &= H_f\hat{S}_f + H_f\hat{t}_{Vf} - H_f\mathcal{F}_{PM}(\hat{t}_V - t_V) \\ &= H_f(S_f + \varepsilon_{Sf}) + H_f\hat{t}_{Vf} - H_f(\hat{t}_{Vf} - t_{Vf}) \\ &= H_f X_f + H_f\varepsilon_{Sf} \end{aligned} \tag{15}$$

où ε_{S_f} est l'erreur commise lors de l'estimation de S_f. En utilisant Z_f au lieu de Y_f dans (11), on peut ainsi voir qu'il n'y aura aucune amélioration : si $\varepsilon_S = \mathcal{I}_P\varepsilon_{Sf}$,

$$\begin{aligned} \hat{S}_Z &= \mathcal{I}_{BP}H_f^{-1}Z_f \\ &= \mathcal{I}_{BP}(X_f + \varepsilon_{Sf}) \\ &= S + \varepsilon_S \\ &= \hat{S} \end{aligned} \tag{16}$$

5 Résumé de l'algorithme

1. Calcule de la variance moyenne de l'observation.
2. Initialisation : des symboles de données aléatoires sont générés.
3. Estimation des coefficients du canal avec (10).
4. Si le nombre d'itérations est > 1, estimation de la variance du bruit (cf section 4.1).
   Estimation des symboles (12) et projection sur l'alphabet BPSK.
5. Calcul de la différence entre l'estimation des symboles à cette itération et à l'itération précédente. Si elle est non nulle, retour à 3.

6 Simulations

Les simulations ont été faites avec un canal de Rayleigh. Les différentes dimensions utilisées sont les suivantes : $L = M = 5$, $V = 100$ et $P = 64$. On a fait 10000 simulations de Monté-Carlo pour un rapport signal à bruit allant de -10 à 20 dB par pas de 5dB (représenté sur l'axe x). Les figures suivantes montrent les résultats moyens relatifs à ces simulations.

Deux sortes de canaux ont été utilisés : le premier est un canal sélectif en fréquence ; le second vient d'un modèle proposé par France Télécom R&D [3]. Dans les figures, **metH = 1** quand la technique de [2] a été utilisée, **metH = 2** quand notre estimation de canal a été utilisée ; **noise** signifie qu'on a inclus $\hat{\sigma}_n$ dans l'estimation des symboles ; **RI** quand on a séparé les parties réelle et imaginaire de l'observation. Enfin **inHex** (resp. **inSex**) représente une itération de l'algorithme en initialisant le canal (resp. les symboles) à sa (resp. leur) vraie valeur. Pour **inHex**, toutes les variantes de **metH = 2** donnent la même courbe tandis que pour **inSex**, les méthodes utilisant la décomposition RI donnent une même courbe tandis que les trois autres en donnent une autre. Pour chaque type

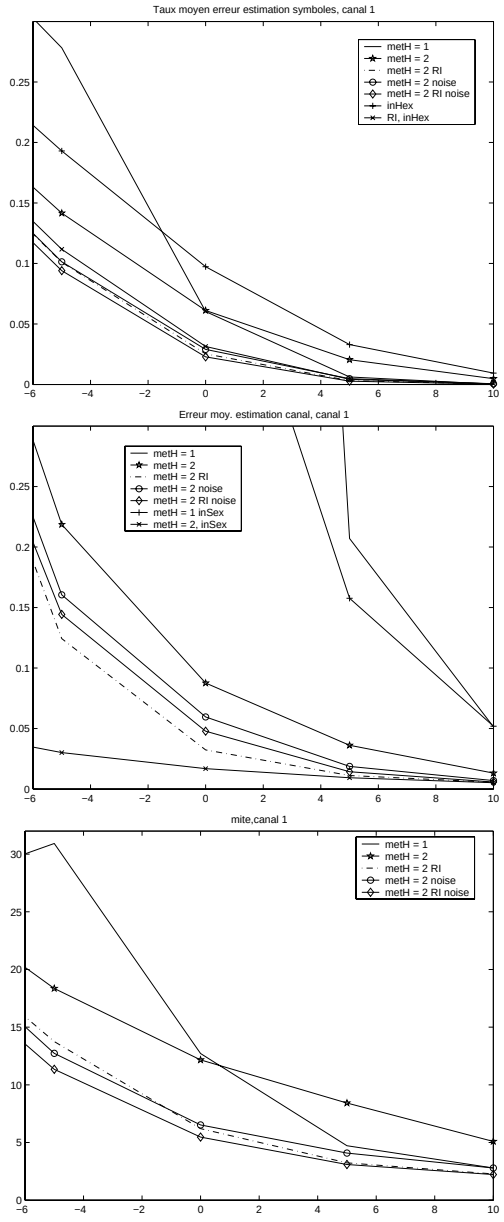


FIG. 1 – Erreurs moy. sur l'estimation des symboles, du canal et nombre moyen d'itérations, canal 1. Zoom de -5 à 10dB

de canal, sur la première figure, l'axe y représente la moyenne du taux d'erreur ; la deuxième figure donne l'erreur moyenne de l'estimation du canal normalisée : $moy\left(\|\hat{\mathbf{h}} - \mathbf{h}\|/\|\mathbf{h}\|\right)$; la dernière donne une idée sur le nombre d'itérations moyen

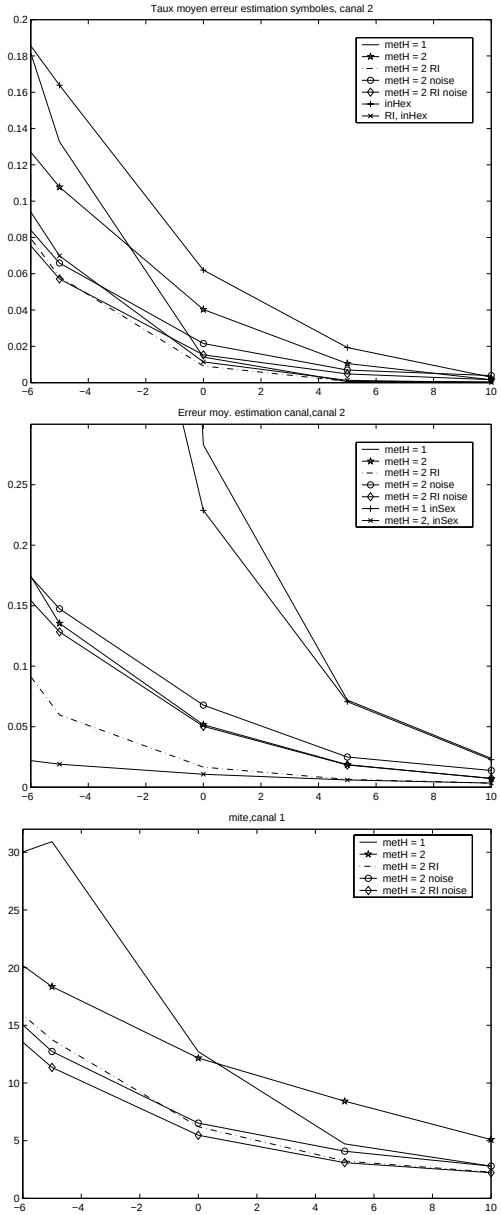


FIG. 2 – Erreurs moy. sur l'estimation des symboles, du canal et nombre moyen d'itérations, canal 2. Zoom de -5 à 10dB

nécessaire à la convergence des méthodes. Pour les trois aspects, notre méthode produit de biens meilleurs résultats que celle de [2] ; dans le cas du taux d'erreur, cette conclusion est surtout vraie à partir d'une variance de bruit de 5dB. On note également une amélioration importante quand les parties réelles et imaginaires de Y sont séparées ainsi que lorsque on ajoute l'estimation du bruit (au niveau de l'estimation du canal et celle des symboles). Enfin, des courbes appelées **inHex** et **inSex**permettent de montrer la faiblesse inhérente à la méthode [2] par rapport à la nôtre (courbes sur $\hat{\mathbf{h}}$) ainsi l'avantage de l'utilisation de la décomposition RI (courbes sur $\hat{S}$)

7 Conclusion

Ce communiqué présente une procédure de minimisation au sens des moindres carrés sous-optimale et peu coûteuse pour le problème de détection et estimation de systèmes mono ou multiporteuse avec canaux convolutifs inconnus et CP. Un algorithme ILSP a été utilisé, exploitant la structure particulière du canal : la FFT d'une matrice circulante est en effet diagonale, ce qui permet de découpler ce problème de minimisation.

On a montré que notre algorithme converge en très peu d'itérations même pour une variance du bruit de 10dB et qu'il présente d'excellents résultats tant sur l'estimation des coefficients du canal que sur la détection du bruit. Deux améliorations intéressantes ont également été proposées, l'une d'entre elles fournissant un moyen efficace d'estimer la variance du bruit, améliorant par là même l'estimation des données au prix d'un faible ajout de complexité.

8 Remerciements

Nous remercions les relecteurs pour leurs remarques constructives. Ce travail a été possible grâce au financement du projet SIGSENSUAL, ref. REN2003-08341-C03-C01-02 et CTM2004-04510-C03-02, conjointement réalisé avec l'Unité de Technologie Marine (UTM) du CSIC, l'Université Polytechnique de Catalogne (UPC) et l'Université de Navarre. Merci en particulier à Antoni Manuel pour la confiance qu'il m'a accordée.

Références

[1] S. Talwar, M. Vilberg & A. Paulraj, "Blind separation of synchronous co–channel digital signals using an antenna array–Part I : Algorithms", *IEEE Trans. on Signal Processing*, vol. 44, no 5, pp. 1184–1197.

[2] O. Rousseaux, G Leus & M. Moonen, "A suboptimal Iterative Method for Maximum–Likelihood Sequence Estimation in a Multipath Context", *EURASIP Journal on Applied Signal Processing*, pp. 1437–1447, dec. 2002.

[3] P. Pajusco, "Experimental characterization of DOA at the basse station in rural and urban area", *IEEE Vehicular Technology Conference*, Ottawa, pp. 993–997, 1998.

[4] L. Mazet, V. Buzenac-Settineri, M. de Courville & P. Duhamel, "An EM based semi-blind channel estimation algorithm designed for OFDM systems", *Asilomar conference on Signals, systems, and computers*, pp. 1642–1646, 2002.

Un nouveau système de transmission basé sur les techniques MIMO SS-MC-MA avec sauts de fréquence et estimation de canal pour la voie montante des futurs réseaux cellulaires de quatrième génération

Laurent CARIOU, Jean-François HÉLARD

Institut d'Électronique et de Télécommunications de Rennes
20 avenue des Buttes de Coësmes, 35043 Rennes Cedex, France
laurent.cariou@ens.insa-rennes.fr, jean-francois.helard@insa-rennes.fr

Résumé – Ce papier présente un nouveau système de transmission pour la voie montante des futurs réseaux cellulaires de quatrième génération. Il repose sur la combinaison de techniques STBC et SS-MC-MA avec sauts de fréquence. Une nouvelle estimation de canal simple, efficace et adaptée à ce système est par ailleurs décrite.

Abstract – This paper deals with a new transmission scheme for the uplink of the future 4G cellular networks. It's based on the combination of STBC and SS-MC-MA techniques with frequency hopping. Besides, a new simple and efficient channel estimation well suited for the proposed system is presented.

1 Introduction

La technique MC-CDMA (multi-carrier code division multiple access), combinant les avantages des modulations à porteuses multiples de type OFDM (orthogonal frequency division multiplexing) et de l'accès multiple par répartition de codes (CDMA) [1] recueille depuis quelques années beaucoup d'attention pour la voie descendante des futurs réseaux cellulaires sans fils. Les travaux menés pendant trois ans par le projet européen IST FP5 MATRICE sur ce sujet se poursuivent aujourd'hui au sein du projet IST FP6 4MORE [2] en prenant en compte la dimension spatiale à travers l'étude des techniques MIMO (multiple input multiple output) et en optimisant l'implémentation de ces techniques sur system on chip (SoC).

En voie montante cependant, le signal MC-CDMA reçu au niveau de la station de base est le résultat de la somme de N_u signaux, potentiellement asynchrones, émis par N_u terminaux mobiles et ayant chacun subi les distorsions d'amplitude et de phase de N_u différents canaux de propagation. Une estimation de canal complexe et des techniques de détection multi-utilisateur sont alors nécessaires pour compenser les effets de l'interférence d'accès multiple (MAI). Pour surmonter ce problème, la technique SS-MC-MA (spread spectrum multi-carrier multiple access) a été récemment proposée pour la voie montante [3]. En quelques mots, le système SS-MC-MA attribue à chaque utilisateur son propre sous-ensemble de sous-porteuses en respectant un schéma de multiplexage fréquentiel FDMA (frequency division multiple access). Dans ce cas, la station de base ne doit plus estimer, pour chaque sous-porteuse, qu'un seul canal au lieu de N_u dans le cas du MC-CDMA avec N_u utilisateurs. De plus, les symboles multipliés par un code d'étalement sont transmis simultanément sur un sous-ensemble de sous-porteuses et subissent les mêmes distorsions. La self-interference (SI), qui remplace dans ce cas la MAI, peut alors facilement être compensée par une détection mono-utilisateur.

Dans cet article, nous décrivons un nouveau système pour la voie montante, basé sur la combinaison des techniques SS-MC-MA et STBC avec sauts de fréquence. Pour compenser la faible diversité fréquentielle du schéma SS-MC-MA utilisant des sous-ensembles de sous-porteuses adjacentes, nous proposons d'appliquer une règle de sauts de fréquence afin que chaque utilisateur bénéficie de la diversité fréquentielle relative à la largeur de bande totale. Les performances de ce nouveau système sont évaluées sur un canal MIMO réaliste en utilisant un codage de canal turbo-convolutif. Enfin, une technique d'estimation de canal MIMO adaptée au système proposé est présentée.

2 Description du système proposé STBC SS-MC-MA

La figure 1 présente une version simplifiée du système MIMO SS-MC-MA, pour l'utilisateur j, basé sur le code d'Alamouti avec N_t = 2 antennes d'émission et N_r = 2 antennes de réception. L'utilisateur j transmet simultanément N_L couples de symboles $x^0_{j,l}$ et $x^1_{j,l}$ sur les antennes 1 et 2 respectivement à l'instant *t*, et -$x^{1*}_{j,l}$ et $x^{0*}_{j,l}$ à l'instant *t*+T_s avec $l = 1, \ldots, N_L$. Le nombre N_c de sous-porteuses est fixé à $N_u.L$ avec N_u le nombre d'utilisateurs et L la longueur d'étalement. Les données d'un utilisateur sont étalées par une simple FHT et la fonction de $chip\ mapping$ détermine le sous-ensemble de L sous-porteuses sur lequel les chips des symboles étalés de l'utilisateur j sont distribués.

En réception, après démodulation OFDM et égalisation, pour chaque antenne de réception r, les deux signaux reçus successivement sont combinés. Les signaux résultants des N_r antennes sont alors sommés pour détecter les symboles de données $x^0_{j,l}$ et $x^1_{j,l}$ et les transmettre au bloc de désétalement :

$$[\hat{x}^0_{j,l}\ \hat{x}^0_{j,l}]^T = (\mathbf{I}_2 \otimes \mathbf{c}_l^T)\mathcal{Y} = (\mathbf{I}_2 \otimes \mathbf{c}_l^T)\sum_{r=1}^{N_r} \mathcal{G}_r \mathcal{R}_r \qquad (1)$$

$$avec \quad \mathcal{G}_r = \begin{bmatrix} G_{1r} & G_{2r}^* \\ G_{2r} & -G_{1r}^* \end{bmatrix}$$

où $\mathbf{I}_2$ représente la matrice identité 2×2, $\otimes$ le produit de Kronecker, $\mathcal{Y} = [y_1^0 \dots y_k^0 \dots y_L^0 \quad y_1^1 \dots y_k^1 \dots y_L^1]^T$ le vecteur des signaux reçus, égalisés et combinés à partir des N_r antennes. G_{tr} est la matrice diagonale contenant les coefficients d'égalisation du canal entre l'antenne de transmission *t* et celle de réception *r* et R_r est le vecteur des signaux reçus avant égalisation sur l'antenne r. $\mathbf{c}_l = [c_{l,1} \dots c_{l,k} \dots c_{l,L}]^T$ est le l^{ieme} code d'étalement orthogonal de Walsh-Hadamard. Comme pour un système OFDM classique utilisant une technique de détection mono-utilisateur, l'égalisation consiste à multiplier par un coefficient complexe l'échantillon porté par la sous-porteuse. Les deux techniques utilisées ici sont le MMSE (minimum mean square error) dont les coefficients sont donnés par $g_{tr,k} = (h_{tr,k}^* \times \rho) / \left[\sum_{t=1}^{N_t=2} \sum_{r=1}^{N_r} |h_{tr,k}|^2 + \frac{1}{\gamma_{r,k}}\right]$ et le ZF (zero forcing) pour lequel le terme $\frac{1}{\gamma_{r,k}}$ doit être omis. ρ est un facteur de normalisation.

3 Méthode d'allocation de fréquences optimale avec SS-MC-MA

La technique SS-MC-MA associe à chaque utilisateur un sous-ensemble de sous-porteuses qui lui sont propres. Une première approche appelée entrelacement par bloc (BI: bloc interleaved) ou multiplexage fréquentiel, distribue les chips d'un symbole étalé d'un utilisateur sur la bande totale en maximisant la séparation fréquentielle entre deux sous-porteuses appartenant au même sous-ensemble. Elle permet ainsi de bénéficier de la diversité fréquentielle liée à la bande totale et d'atteindre les performances optimales de la technique MC-CDMA. La deuxième approche consiste à utiliser des sous-porteuses adjacentes (AS: adjacent subcarriers) pour transmettre les chips d'un symbole étalé d'un utilisateur donné. Dans ce cas, chaque utilisateur n'utilise qu'une bande limitée et l'exploitation de la diversité fréquentielle se retrouve sévèrement réduite.

Cependant, il a été démontré dans [4] que la solution AS est nettement plus robuste que la solution BI face aux écarts en fréquence des oscillateurs des différents utilisateurs. Quand de tels écarts surviennent, l'orthogonalité entre sous-porteuses

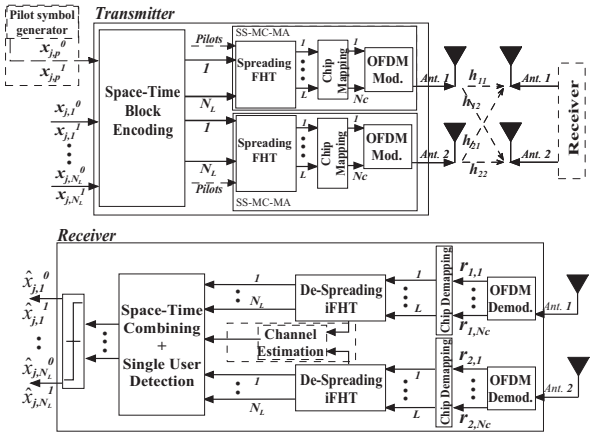


FIG. 1: *Émetteur et récepteur STBC SS-MC-MA pour un utilisateur j avec deux antennes d'émission et deux antennes de réception*

est perdue et de l'interférence entre porteuses (ICI), qui est en fait de la MAI, apparaît. Étant donnée la décroissance rapide de la fonction $sinc$, la contribution la plus critique de l'interférence sur une sous-porteuse donnée provient de ses deux sous-porteuses voisines. Ainsi, avec la solution BI, $2 \times L$ interférents doivent être considérés contre seulement deux pour la solution AS. De plus, en présence de fort écarts de fréquence, l'approche AS permet l'insertion, entre deux sous-ensembles de sous-porteuses, de sous-porteuses de garde pour absorber l'ICI, conduisant à une amélioration des performances pour une faible perte en efficacité spectrale. En ce qui concerne l'estimation de canal, la complexité est réduite avec la solution AS car une plus faible bande de fréquences doit être estimée. En outre, des techniques d'interpolation peuvent être appliquées sur la dimension fréquentielle et une réduction du nombre de pilotes est possible, conduisant à moins de pertes en puissance et en efficacité spectrale, comparée à la solution BI.

Pour palier aux imperfections de ces deux approches, nous avons récemment proposé dans [5] une nouvelle technique d'allocation de fréquence intermédiaire qui consiste à rajouter à la solution AS une composante de sauts de fréquence (FH: frequency hopping). Le sous-ensemble de sous-porteuses associé à un utilisateur change d'un symbole OFDM à l'autre en respectant une loi simple permettant à chaque utilisateur d'expérimenter chaque sous-ensemble de sous-porteuses de manière équiprobable en évitant toute collision :

$$s_{i,j} = (s_{i-1,j} + inc) \mod N_s, \tag{3}$$

$s_{i,j}$ est l'indice du sous-ensemble de sous-porteuses pour le symbole OFDM i et l'utilisateur j. N_s est le nombre de sous-ensembles de sous-porteuses au sein d'un symbole OFDM et inc l'incrément optimisé en terme de performances pour profiter au maximum de la diversité fréquentielle.

Cette solution doit permettre de tirer profit de la diversité fréquentielle liée à la bande totale, égalant les performances optimales de la solution BI, tout en conservant les avantages de la solution AS en terme de robustesse par rapport aux écarts en fréquence des utilisateurs et d'estimation de canal. De plus, avec sa composante d'accès multiple FDMA, la technique SS-MC-MA est sensible aux interférences inter-cellulaires. Certains événements catastrophiques peuvent apparaître lorsque deux utilisateurs proches de la frontière entre deux cellules transmettent sur le même sous-ensemble de sous-porteuses. Si les sauts dans les cellules adjacentes suivent des lois différentes, l'utilisation du saut de fréquence moyenne les interférences sur tous les utilisateurs.

4 Estimation de canal adaptée à la solution FH SS-MC-MA proposée

En prenant en considération les contraintes en efficacité spectrale et de puissance liées à une technique d'estimation de canal adaptée à la solution SS-MC-MA avec sous-porteuses adjacentes et sauts de fréquence, nous nous sommes intéressés à la technique du pilote étalé [6], [7]. Inspirée de la technique du pilote superposé, elle consiste à transmettre, sur chaque sous-porteuse, un pilote de faible puissance simultanément avec les données. Pour réduire les interférences croisées entre données et pilotes, le concept du pilote étalé est de rendre la séquence

pilote orthogonale avec la séquence de données en attribuant au pilote un des codes orthogonaux de Walsh Hadamard (WH) $\mathbf{c}_p = [c_{p,1} \dots c_{p,L}]^T$ sur chaque sous-ensemble de sous-porteuses. Une faible perte en efficacité spectrale doit être concédée. Les pertes en puissance peuvent être exprimées par $10log(\frac{N_L}{N_L+\eta})$, en considérant le rapport de puissance η entre les pilotes P_p et les données P_d.

Pour que l'estimation soit la plus simple possible, le désétalement est réalisé avant détection mono-utilisateur. La séquence pilote étant définie pour un seul coefficient de canal, cette solution nécessite l'invariance du canal sur le sous-ensemble de sous-porteuses sur lequel la séquence est distribuée. Connaissant les statistiques à l'ordre deux du canal, des techniques de chip mapping comme l'étalement à deux dimensions (2D) peuvent être appliquées au SS-MC-MA de la même manière qu'elles sont appliquées au MC-CDMA pour constituer ce que l'on appelle OFCDM (orthogonal frequency code divison multiplexing). Cette technique consiste tout simplement à répartir les chips d'un symbole étalé sur des sous-porteuses adjacentes (dimension fréquentielle SF_{freq}) et sur des symboles OFDM consécutifs (dimension temporelle SF_{time} avec $L = SF_{freq} \times SF_{time}$) pour obtenir un canal aussi plat que possible sur le symbole étalé.

4.1 Principe de l'estimation de canal par pilote étalé et extension au cas MIMO

Un des principaux intérêts de la technique du pilote étalé est sa facilité à s'adapter au cas MIMO. La combinaison du code d'Alamouti et de FH SS-MC-MA peut être appliquée aux pilotes comme aux données, permettant une simple estimation des différents canaux MIMO. La première approche considérée ici sera appelée STBC puisque le code d'Alamouti est appliqué sur la dimension temporelle, sur deux symboles OFDM consécutifs. Dans ce cas, $N_L + 1 \leq L$, les contraintes du code d'Alamouti se combinent à celles du pilote étalé et l'invariance du canal sur un sous-ensemble de $2 \times SF_{freq} \times SF_{time}$ sous-porteuses doit être respectée, comparé à $SF_{freq} \times SF_{time}$ dans le cas SISO.

Pour pallier à cet inconvénient, nous proposons d'appliquer le code d'Alamouti dans la dimension des codes, sur deux séquences de WH au sein du même symbole étalé. De cette manière, les techniques de codage espace-temps et espace-fréquence sont combinées pour former ce qui peut être nommé un code en bloc espace-code (SCBC). Avec ce nouveau schéma, les contraintes sur l'invariance du canal sont réduites à un sous-ensemble de $SF_{freq} \times SF_{time}$ sous-porteuses puisque les deux codes expérimentent exactement les mêmes canaux, respectant les contraintes imposées par le code d'Alamouti. Par contre, N_L doit maintenant être pair et respecter $N_L + 2 \leq L$. Deux pilotes étalés sont utilisés pour chaque sous-ensemble de sous-porteuses et les pertes de puissance sont égales à $10log(\frac{N_L}{N_L+2\eta})$. Cette deuxième approche est appelée SCBC.

Les composantes liées à l'estimation de canal sont présentées en pointillés sur la figure 1. A l'émission, un symbole pilote pour STBC et deux pour SCBC, x_{p_1} et x_{p_2} sont insérés dans le flux de données, tous les N_L symboles de données. Les codes de WH $\mathbf{c}_{p_a}$ et $\mathbf{c}_{p_b}$ sont associés aux symboles pilotes pour SCBC, nous les appelons $\mathbf{c}_{p_t}$ et $\mathbf{c}_{p_{t+T}}$ pour STBC. Dans ce qui suit, nous approfondissons plus en détail la technique SCBC. A la réception, sur chaque antenne r, le désétalement sépare les données et les pilotes par corrélation pour obtenir

$$(c_{p_a}, r) = x_{p_1}.\bar{h_{1r}} + x_{p_2}.\bar{h_{2r}} + SI_{p_a} + \sum_{k=0}^{L-1} c_{p_a,k} n_k \quad (4)$$

$$(c_{p_b}, r) = -x^*_{p_2}.\bar{h_{1r}} + x^*_{p_1}.\bar{h_{2r}} + SI_{p_b} + \sum_{k=0}^{L-1} c_{p_b,k} n_k \quad (5)$$

La détection du code d'Alamouti permet de générer les estimées au sens des moindres carrés de la moyenne des coefficients du canal pour chaque antenne de réception r

$$\hat{h}_{1r} = \frac{(c_{p_a}, r)\, x^*_{p_1} - (c_{p_b}, r)\, x_{p_2}}{(|x_{p_1}|^2 + |x_{p_2}|^2)} \quad (6)$$

$$\hat{h}_{2r} = \frac{(c_{p_b}, r)\, x_{p_1} - (c_{p_a}, r)\, x^*_{p_1}}{(|x_{p_1}|^2 + |x_{p_2}|^2)} \quad (7)$$

De façon similaire au cas SISO [7], le terme de SI est composé de tous les symboles de données et pilotes (autres que le symbole utile) transmis par le sous-ensemble de sous-porteuses ($i \in \mathbf{E}$) et multipliés par l'écart par rapport à la moyenne des coefficients du canal $\epsilon_{h_{mr,k}}$.

$$SI_a = \sum_{m=1}^{N_t=2} \sum_{\substack{i \in \mathbf{E} \\ i \neq p_a}} x_{i,m} \sum_{k=0}^{L-1} \epsilon_{h_{mr,k}} c_{j,k} c_{p_a,k} \quad (8)$$

La variance de la SI σ^2_{SI} pour tous les symboles de données est proportionnelle aux variances des canaux sur le sous-ensemble de sous-porteuses σ^2_h. Elle peut être exprimée par:

$$\sigma^2_{SI} = \frac{(2 \times (N_L - 1)P_d + 2 * P_p)}{L} \frac{\sigma^2_h}{2} \quad (9)$$

5 Performances du système proposé STBC SS-MC-MA sur canaux réalistes

5.1 Paramètres du système

Les paramètres du système ont été choisis en fonction du temps et de la bande de cohérence du canal pour remplir les contraintes de la 4G des futurs systèmes cellulaires sans fils. Les fréquences porteuses et d'échantillonnage ont été fixées respectivement à 5 GHz et 57.6 MHz. La taille de la FFT est de 1024 et le nombre de sous-porteuses modulées 736. L'intervalle de garde atteint 3.75 μs et la durée totale d'un symbole 21.52 μs. Le codage de canal est réalisé par le turbo-code UMTS de rendement $R = 1/2$ après poinçonnage. La profondeur d'entrelacement a été ajustée à la trame composée de 32 symboles OFDM. Avec une modulation QPSK, le système offre une efficacité spectrale asymptotique de 1 bit/s/Hz. Nous utilisons un modèle de canal MIMO développé spécifiquement au sein du projet européen IST MATRICE et basé sur la proposition 3GPP/3GPP2 pour des canaux MIMO large bande exploitant les caractéristiques angulaires des multi-trajets. Adapté à 5 GHz, le modèle utilise le profil puissances-retards BRAN E qui correspond à une propagation multi-trajet de type urbain outdoor. La bande de cohérence est de 1.5 MHz à -3dB. La corrélation spatiale est inférieure à 0.1 pour une espacement entre antennes de 10 longueurs d'ondes (λ) à la station de base et proche de 0.3 pour 1 λ au niveau du terminal. Enfin, même pour une vitesse de mobile de 72 km/h, le temps de cohérence reste proche de la durée de la trame.

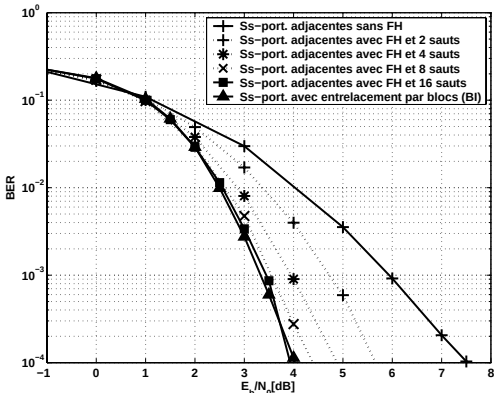


FIG. 2: *Gain offert par le saut de fréquence pour un nombre variable de sauts; système Alamouti SS-MC-MA à pleine charge, détection MMSE, sous-porteuses adjacentes, QPSK, turbo-code UMTS avec R=1/2*

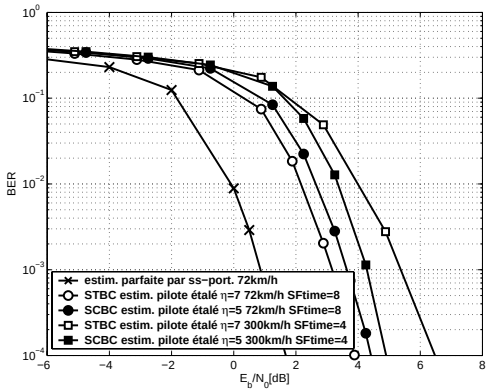


FIG. 3: *Performances de la technique du pilote étalé pour deux vitesses du mobile; système Alamouti FH SS-MC-MA à pleine charge; détection ZF, meilleure configuration d'étalement 2D, QPSK, turbo-code UMTS avec R=1/2*

5.2 Résultats de simulation

Tous les résultats sont donnés en fonction du rapport E_b/N_0, où E_b représente l'énergie par bit utile. La puissance totale transmise est normalisée et égale à P dans les cas SISO et MIMO. Le gain de puissance généré par l'utilisation de deux antennes à la réception est pris en compte. La figure 2 présente les performances en terme de taux d'erreur binaire (TEB) du système turbo-codé Alamouti STBC22 SS-MC-MA avec sous-porteuses adjacentes pour différents nombres de sauts de fréquences par trame, une estimation de canal parfaite et une détection MMSE. Les performances de la solution d'entrelacement par bloc (BI) sont également données en référence. Comme il est avéré que la dernière solution bénéficie de la diversité fréquentielle liée à la bande totale, elle représente l'objectif à atteindre en terme de performances pour notre solution AS FH. Nous pouvons observer que pour 16 sauts de fréquence par trame, *c.a.d.* un saut tous les deux symboles OFDM, les performances de notre solution tendent vers celles de la solution BI. Nous pouvons donc clairement en conclure que la solution proposée SS-MC-MA avec sous-porteuses adjacentes et sauts de fréquence tire profit de la diversité fréquentielle liée à la bande totale, tout en conservant les avantages de la solution AS introduits dans la section 3.

La figure 3 présente, pour 2 antennes d'émission et 2 de réception, les performances en terme de TEB du système Alamouti STBC22 SS-MC-MA avec sous-porteuses adjacentes et sauts de fréquence, des configurations d'étalement 2D optimale pour 72 et 300 km/h et une détection ZF qui offre des performances très proches d'une détection MMSE. Les techniques d'estimation de canal par pilote étalé STBC et SCBC sont comparées à l'estimation parfaite par sous-porteuse en prenant en compte les pertes de puissance liées à l'insertion de pilotes. L'optimisation de puissance des pilotes permet un gain de 2 dB par rapport au cas où les pilotes et les données ont même puissance η=1, réduisant l'écart par rapport à une estimation parfaite à 2.2 dB. L'optimisation de puissance des pilotes est plus efficace pour STBC que pour SCBC à 72 km/h, car dans le second cas, deux pilotes étalés sont transmis par chaque sous-ensemble de sous-porteuses, conduisant à de plus fortes pertes de puissance que dans le cas STBC. Le rapport $\eta_{opt_{SCBC}}$=5 optimal est ainsi plus faible que pour STBC $\eta_{opt_{STBC}}$=7. La solution SCBC est, quant à elle, plus robuste en présence d'un canal fortement sélectif en temps mais aussi en fréquence, les contraintes sur la stationnaire du canal étant moins fortes. Lorsque la vitesse du récepteur augmente de 72 à 300 km/h, la perte de performances pour un TEB de 10^{-4} reste inférieure à 0.5 dB pour la solution SCBC alors qu'elle atteint 2,5 dB avec la solution STBC. Les performances de la solution SCBC s'avèrent donc peu dépendantes des paramètres du canal de propagation.

6 Conclusion

Ce papier décrit une nouvelle solution pour la voie montante des futurs réseaux cellulaires sans fils. Elle repose sur la combinaison de techniques SS-MC-MA et MIMO. L'utilisation de sous-porteuses adjacentes avec sauts de fréquence permet au système SS-MC-MA de profiter de la diversité fréquentielle liée à la bande totale, tout en conservant de nombreux avantages, notamment en terme de robustesse vis-à-vis des écarts en fréquence des oscillateurs des différents utilisateurs. En outre, une nouvelle technique d'estimation de canal MIMO adaptée au système proposé et utilisant des pilotes étalés sur les deux dimensions temps-fréquence est présentée. Dans le cas d'un schéma Alamouti, deux différentes approches, STBC et SCBC, sont étudiées. Conduisant à une perte de seulement 2 dB comparé à une estimation parfaite, cette technique d'estimation s'avère simple et efficace même en présence de canaux fortement sélectifs.

Remerciements : Les auteurs remercient France Télécom R&D/RESA/BWA qui a soutenu ce travail dans le cadre du contrat 461 28 187.

Références

[1] S. Hara and R. Prasad, "Overview of multicarrier cdma," *IEEE communication magazine*, vol. 35, no. 12, pp. 126–133, Dec. 1997.

[2] IST 4MORE project, "http://www.ist-4more.org," .

[3] S. Kaiser and W.A. Krzymien, "Performance effects of the uplink asynchronism in a spread spectrum multi-carrier multiple access system," *Europeen Transactions on Telecomunications*, vol. 10, no. 4, July/August 1999.

[4] A. Arkhipov and M. Schnell, "The influence of user frequency offset on the up-link performance of ss-mc-ma," *Proc. of European Conference on Wireless Technology (ECWT'03)*, pp. 451–454, 2003.

[5] L. Cariou and J.F. Hélard, "Mimo frequency hopping ofdm-cdma: a novel uplink system for b3g cellular networks," *Proc. IEEE International Conference on Networking*, April 2005.

[6] N. Chapalain, D. Mottier, and D. Castelain, "An ofcdm uplink transmission system with channel estimation based on spread pilots," *Proc. of IST Mobile Summit*, 2004.

[7] L. Cariou and J-F. Hélard, "A simple and efficient channel estimation for mimo ofdm code division multiplexing uplink systems," *Proc. IEEE International Workshop on Signal Processing Advances for Wireless Communications (SPAWC'05)*, 2005.

Throughput evaluation of HARQ Schemes with Packet and Code Combining over multipath fading channels for DS-SSS

M. Wissem El Bahri, Hatem Boujemaa, Mohamed Siala

Ecole Supérieure des Communications de Tunis, MEDIATRON Laboratory
Route de Raoued Km 3.5, 2083 El Ghazela, Ariana, Tunisia
wissem_bahri@yahoo.fr, boujemaa.hatem@supcom.rnu.tn , mohamed.siala@supcom.rnu.tn

Résumé − Dans cet article, nous évaluons les performances des schémas ARQ et HARQ I, utilisant la stratégie Packet Combining, et HARQ II, utilisant la stratégie Code Combining, en présence d'un canal multi-trajets à évanouissements de Rayleigh. L'analyse est effectuée dans deux scénorios extrèmes où le canal est constant et indépendent durant les différentes retransmissions. Les résultats théoriques obtenus sont ensuite validés par des simulations.

Abstract − In this paper, the performance of both ARQ and HARQ I, with Packet Combining (PC), and HARQ II, with Code Combining (CC), are evaluated in terms of throughput efficiency over multipath fading channels. The analysis is done in two extreme scenarios where the channel is assumed to be constant and independent during the different transmissions. The theoretical results are then compared to those obtained by computer simulations and confirm the great advantage of using a PC strategy.

1 Introduction

The High Speed Downlink Packet Access (HSDPA) is a new mechanism proposed by 3GPP [1] to support higher data transmission rate for mobile users, and to provide streaming, interactive and background services with a good quality of service. This is accomplished by using different techniques such as Hybrid Automatic Repeat reQuest (HARQ) scheme. HARQ consists in combining ARQ and FEC schemes [2]. Two types of Hybrid ARQ schemes [2] have been identified in which error correction followed by error detection are applied at every received packet. In HARQ I schemes, the correction of errors is first attempted at each received packet. If this fails, the entire packet is discarded and its retransmission is requested. Each subsequent retransmission is decoded independently of the prior retransmissions. The main disadvantage of ARQ and HARQ I schemes is their low throughput at low signal to noise ratios. In fact, when a packet is detected in error, it must be retransmitted and, therefore, for low or mean signal to noise ratios, the number of transmissions needed before receiving a packet correctly is high and then the transmission rate is reduced to an unacceptable level. In HARQ II schemes, the receiver stores the erroneous packets in a buffer so that they can be reused after subsequent retransmissions. Storing the previously received packets allows the technique of Code Combining [3] to be exploited which consists in alternately sending one of the outputs of a convolutional encoder. In this strategy, the first transmission occurs without coding and successive repetition of the convolutional encoder outputs yields a family of repetition codes of decreasing rates equal to $1/j$ where j is the number of transmission tentative.

Significant improvement of the performance of ARQ and HARQ I schemes can be obtained by using a Packet Combining strategy. In this technique, all received packets are averaged in a soft manner at the input of the channel decoder. This strategy is considered in the HSDPA mechanism since it is less complex than HARQ II. Packet combining was initially introduced by Sindhu [5]. It uses hard decisions and tries to correct errors based on the differences between received packets. Benelli [6], Metzner and Chang [7] generalised Sindhu's idea by using soft decisions. The performance of ARQ and HARQ I with packet combining have been analyzed in [5] for a gaussian and binary symmetric channel. The analysis carried in [4] assumes a single block fading path where the channel is assumed to be independent between different transmissions. In this paper, we evaluate the performance of HARQ schemes with Packet Combining and Code Combining strategy over multipath fading channels for Direct Sequence Spread Spectrum System (DS-SSS). The analysis is done into extreme scenarios in which fading is assumed to be constant and independent during the different transmissions.

The paper is organized as follows. Section II gives the system model. Sections III, IV and V give respectively the performance of ARQ with PC, HARQ I with PC and HARQ II with CC, over multipath fading channels. Section VI compares simulation and theoretical results of the different schemes in terms of throughput efficiency (Thr). Finally, section VII draws some conclusions.

2 System model

In this section, we describe the channel decoder input for DS-SSS using a Rake receiver. The Rake receiver is assumed to have a perfect knowledge of complex path gain modules $\{\alpha_l\}_{l=1}^{L}$, where L is the number of paths. If path

delays are well separated and in the absence of Inter Symbol Interference and Multi-User Interference, the Rake receiver output for symbol s_n can be written as [9]

$$\hat{s}_n = \sum_{l=1}^{L} \alpha_l^2 s_n + w_n \tag{1}$$

where s_n is the n-th transmitted symbols which is dropped from a BPSK constellation and w_n is a complex Gaussian noise with variance

$$Var(w_n) = \sum_{l=1}^{L} \alpha_l^2 N_0 \tag{2}$$

where N_0 is the Power Spectral Density of the complex channel noise.
The wireless link implements the Selective Repeat (SR) protocol for retransmission of erroneous packets with perfect code detection and suitably large buffers at the transmitter and the receiver. Furthermore, we assume an error free feedback channel over which positive or negative acknowledgements can be sent.

3 ARQ with Packet Combining

In the conventional ARQ scheme, the transmitter sends a packet consisting of k information bits and n_p parity bits for error detection. At the receiver, packets declared in error are discarded and replaced by another copy. The performance of conventional ARQ schemes can be improved if the erroneous packets are not discarded but averaged with the new packet to reduce the effect of the channel noise and therefore the average number of transmissions. Let Tr be the average number of transmission attempts that must be made before a packet is accepted by the receiver. The average number of transmission Tr is given by

$$Tr = 1 + \sum_{j=1}^{+\infty} P(R_d^j) \tag{3}$$

where R_d^j is the event "received sequence obtained by combining the first j packets contains detected errors".
Assuming that the probability of undetected errors is negligible, the probability of detected errors at the j-th retransmission can be evaluated numerically as follows

$$P(R_d^j) = 1 - \int [1 - p(\boldsymbol{\beta}^j)]^{k+n_p} f(\boldsymbol{\beta}^j) d\boldsymbol{\beta}^j \tag{4}$$

where $\boldsymbol{\beta}^j = (\boldsymbol{\alpha}^1, ..., \boldsymbol{\alpha}^j)$, $\boldsymbol{\alpha}^j = (\alpha_1^j, ..., \alpha_L^j)$, α_l^j is the l -th path gain during the j-th transmission, $p(\boldsymbol{\beta^j})$ is the Conditioned Bit Error Rate (CBER) at the channel decoder input and $f(\boldsymbol{\beta^j})$ is the joint probability density function of path gains.
If the channel is independent during the different transmissions, the CBER is given by

$$p(\boldsymbol{\beta^j}) = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{E_s}{N_0}\sum_{i=1}^{j}\sum_{l=1}^{L}(\alpha_l^i)^2}\right) \tag{5}$$

where E_s is the average transmitted energy per symbol.
When the channel is constant during the different transmissions ($\boldsymbol{\alpha}^1 = ... = \boldsymbol{\alpha}^j = \boldsymbol{\alpha} = (\alpha_1, ..., \alpha_L)$), the CBER is given by

$$p(\boldsymbol{\beta^j}) = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{E_s}{N_0} j \sum_{l=1}^{L}(\alpha_l)^2}\right). \tag{6}$$

The throughput efficiency Thr is given by

$$Thr = \frac{1}{Tr}\frac{k}{(k+n_p)} \tag{7}$$

where the factor $k/(k+n_p)$ is the loss in throughput due to the added parity check bits for error detection.

4 HARQ I with Packet Combining

In HARQ I with packet combining, erroneous received packets are averaged with the new packet before decoding. In this paper, we assume that a convolutional encoder of rate $R_c = u/v$ and memory m is used for error correction. At the receiver, a soft input Viterbi decoder is used.
If the channel is independent during the different transmissions, the packet error probability at the $j-th$ transmission $P(D_d^j)$ can be upper-bounded as

$$P(D_d^j) \leq 1 - \int [1 - P_E(\boldsymbol{\beta^j})]^{\frac{k+n_p}{u}} f(\boldsymbol{\beta^j}) d\boldsymbol{\beta^j} \tag{8}$$

where D_d^j is the event "decoded sequence obtained by combining the first j packets contains detected errors", $P_E(\boldsymbol{\beta^j})$ is the Error Event Probability (EEP) upper-bounded by

$$P_E(\boldsymbol{\beta^j}) \leq \sum_{d=d_f}^{+\infty} a_d P_d(\boldsymbol{\beta^j}) \tag{9}$$

where d_f is the free distance, a_d is the number of incorrect paths at distance d and

$$P_d(\boldsymbol{\beta^j}) = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{dE_s}{N_0}\sum_{i=1}^{j}\sum_{l=1}^{L}(\alpha_l^i)^2}\right) \tag{10}$$

is the conditioned probability that a wrong path at distance d is selected.
If the channel is constant during the different transmissions, equation (8) remains valid and equation (10) reduces to

$$P_d(\boldsymbol{\beta^j}) = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{dE_s}{N_0} j \sum_{l=1}^{L}(\alpha_l)^2}\right). \tag{11}$$

The average number of transmissions is given by

$$Tr = 1 + \sum_{j=1}^{+\infty} P(D_d^j) \tag{12}$$

The throughput efficiency of HARQ I with PC is given by

$$Thr = \frac{R_c}{Tr}\frac{k}{(k+n_p+um)}, \tag{13}$$

where the factor $k/(k+n_p+um)$ is the loss in throughput due to the added parity bits for error detection and to the tail of um known bits appended to each transmitted sequence.

5 HARQ II with Code Combining

HARQ II with Code Combining [3] belongs to the class of incremental redundancy schemes in which incremental bits are progressively transmitted in order to optimize

the throughput. In HARQ II with Code Combining, the transmitter alternatively sends one of the outputs of a convolutional encoder. At the reception, the receiver has the possibility to extract the information from the last uncoded received packet or from the combined version of all previously received packets. By using this strategy, the first transmission occurs only with code detection and successive transmissions yield a family of repetition codes of decreasing rates equal at $1/j$ where $j \geq 2$. The average number of transmissions is upper bounded by [3]

$$Tr \leq 1 + P(R_d) + \sum_{j=1}^{+\infty} P(D_d^j). \tag{14}$$

where R_d is the event "last received sequence contains detected error". Assuming that the probability of undetected errors is negligible, the probability that a received sequence is detected in error is given by

$$P(R_d) = 1 - \int (1 - p(\boldsymbol{\alpha}))^{k+n_p} f(\boldsymbol{\alpha}) d\boldsymbol{\alpha} \tag{15}$$

where $p(\boldsymbol{\alpha})$ is the CBER given by

$$p(\boldsymbol{\alpha}) = \frac{1}{2}\mathrm{erf_c}\left(\sqrt{\frac{E_s}{N_0}\sum_{l=1}^{L}(\alpha_l)^2}\right). \tag{16}$$

The packet error probability $P(D_d^j)$ is also given by (8). If the channel is constant during the different transmissions, $\boldsymbol{\beta}^j$ has to be replaced by $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_L)$ and the EEP is upper-bounded by

$$P_E(\boldsymbol{\alpha}) \leq \sum_{d=d_f^{(j)}}^{+\infty} a^{(j)}(d) P_d(\boldsymbol{\alpha}) \tag{17}$$

where

$$P_d(\boldsymbol{\alpha}) = \frac{1}{2}\mathrm{erf_c}\left(\sqrt{\frac{dE_s}{N_0}\sum_{l=1}^{L}(\alpha_l)^2}\right), \tag{18}$$

$a^{(j)}(d)$ and $d_f^{(j)}$ are respectively the distance spectra and the free distance of rate $1/(1+j)$ code.
If the channel is independent during the different transmissions, the total distance between two paths has to be divided into component distances, $d = d_1 + ... + d_j$, where d_l gives the distance between the error path and the zero path during the $l-th$ transmission. The EEP is upper bounded by

$$P_E(\boldsymbol{\beta^j}) \leq \sum_{d_1=d_{1,f}} ... \sum_{d_j=d_{j,f}} a(d_1, ..., d_j) P_d(\boldsymbol{\beta^j}) \tag{19}$$

where

$$P_d(\boldsymbol{\beta^j}) = \frac{1}{2}\mathrm{erf_c}\left(\sqrt{\frac{E_s}{N_0}\sum_{p=1}^{j} d_p \sum_{l=1}^{L}(\alpha_l^p)^2}\right), \tag{20}$$

$a(d_1, ..., d_j)$ is the number of error events with distance vector $\boldsymbol{d} = (d_1, ..., d_j)$. The lower summation limits, $\boldsymbol{d_f} = (d_{1f}, ..., d_{jf})$, are the free component distances associated with each subcode. To compute this distance spectra, we can use the generalized transfer function [10] of the code, or the modified version of the fast Cederval algorithm searching the code tree [8]. The throughput efficiency is given by

$$Thr = \frac{1}{Tr}\frac{k}{k + n_p + um} \tag{21}$$

where the factor $k/(k+n_p+um)$ is the loss in throughput due to the added parity bits for error detection and to the tail of um known bits appended to each transmitted sequence.

6 Numerical and simulation results

In this section, we compare theoretical and simulation results in terms of throughput efficiency evolution with respect to E_s/N_0 for a two paths Rayleigh fading channel ($L = 2$). The code used for HARQ I and HARQ II with Code Combining is a half rate convolutional encoder of constraint length $m + 1 = 7$ and generators polynomials (133,171). The free distance d_f and the distance spectra a_{d_f} of repetition codes of HARQ II with Code Combining are listed in table VI of [3]. In order to compare the performance of the different schemes, the same length of the transmitted packets must be maintained. In the simulations, the transmitted packet length was equal to 960 bits. The parameters of the different schemes are listed in Table 1. Figure 1 compares the performance of ARQ and

TAB. 1: Parameters for different schemes used for numerical and simulation results

Schemes	k (bits)	n_p (bits)
ARQ with and without PC	940	20
HARQ I with and without PC	454	20
HARQ II with CC	940	14

ARQ with PC in the constant and the independent channel scenarios. We can observe that at high E_s/N_0, the PC strategy does not yields a throughput improvement. However, as the channel degrades, the throughput without PC drops rapidly to zero, whereas with PC a high throughput is still achieved even at very low E_s/N_0. We also notice that performance improvement brought by packet combining is greater when the channel is constant.

Figure 2 shows the performance of HARQ I and HARQ I with PC in the same context. We show that PC strategy improves also the performance of HARQ I.

In figures 3 and 4, we have compared the performance of packet combining to Code Combining schemes respectively in the constant and the independent channel scenarios. We see that at low E_s/N_0 the performance of ARQ and HARQ I with PC are close to those of HARQ II with Code Combining. We also notice that ARQ with PC offers close performance to HARQ II with Code Combining at high E_s/N_0. Finally, we verify that simulation results are in a good agreement with the theoretical ones.

7 Conclusion

In this paper, we have derived the performance of ARQ and HARQ I with PC in multipath Rayleigh fading channels. The analysis is done in two extreme scenarios where the channel is assumed to be constant and independent during the different transmissions. We have shown that

the performance improvement brought by PC is greater when the channel is constant during the different transmissions.

References

[1] 3GPP TR 25.848 V4.0.0, Physical layer aspects of UTRA high speed downlink packet access (release 4), 2001.

[2] M. Wissem El Bahri, Hatem Boujemâa and Mohamed Siala "Performance Comparison of Type I, II and III Hybrid ARQ Schemes over AWGN channels," *IEEE ICIT*, Hammamet, Tunisia, Dec. 2004.

[3] S. Kallel, "Analyis of a type II hybrid ARQ schemes with code combining", *IEEE Trans. Commun.*, Vol. 38, pp. 1133-1137, Aug. 1990.

[4] K Akhavan, S. Farahvash and M. Kavehrad , "Performance of Hybrid ARQ Schemes with ML Code Combining over a Block-Fading Rayleigh Channel", *Inter. Jour. of Wirel. Infor. Netw.*, Vol. 7, No. 1, 2000.

[5] Y. M. Wang and S. Lin "A modified selective type II hybrid ARQ systems and its performance analysis", *IEEE Trans. Commun.*, Vol. COM-31 pp. 593-608, May 1983.

[6] G. Benelli, "An ARQ scheme with memory for mobile communications", *Euro. Trans. on Telecom.* 3(5) pp. 11-15, 1992.

[7] J. J. Metzner, and D. Chang "Efficiency selective repeat ARQ strategies for very noisy and fluctuating channels", IEEE Trans. Commun., Vol. COM-33 pp. 409-416, 1985.

[8] M. Cedervall and R. Johannesson "A fast algorithm for computing distance spectrum of convolutional codes ", *IEEE Trans. on Infor. Theory*, Vol. 35 pp. 1146-1159, Nov. 1989.

[9] H. Boujemâa, M. Siala, "On the rake receiver performance," *IEEE Veh. Tech. Conf.*, Vol. 3, pp. 1483-1488, Sept. 2000.

[10] Y. S. Leung, S. G. Wilson, J. W. Ketchung, Multifrequency treillis coding with low delay for fading channels", *IEEE Trans. on Commun.*, Vol. 41, no. 10, pp. 1450-1459, Oct. 1993.

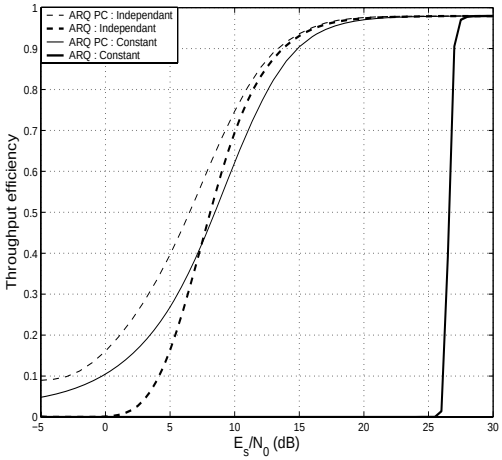


FIG. 1: Performance improvement of ARQ using PC strategy: $k = 940, n_p = 20$ and $L = 2$.

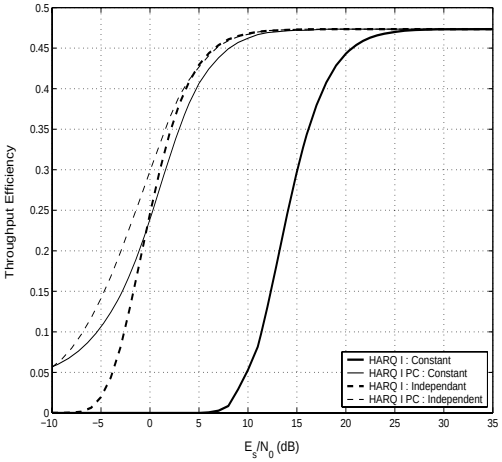


FIG. 2: Performance improvement of HARQ I using PC strategy: $k = 454, n_p = 20$ and $L = 2$.

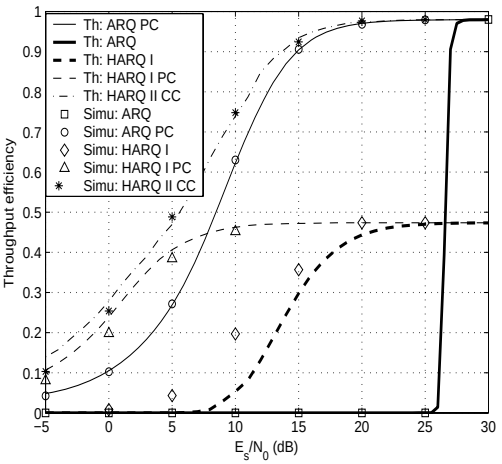


FIG. 3: Performance comparison of ARQ and HARQ I with PC to HARQ II with Code Combining for a constant channel ($L = 2$).

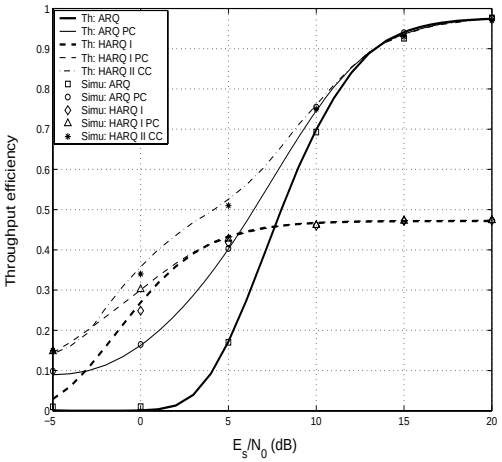


FIG. 4: Performance comparison of ARQ and HARQ I with PC to HARQ II with Code Combining for an independent channel ($L = 2$).

Propriétés du second ordre de l'effet de grenaille

Bernard PICINBONO

Laboratoire des Signaux et Systèmes
Supélec, Plateau de Moulon, 91190 Gif-sur-Yvette, France
bernard.picinbono@lss.supelec.fr

Résumé – L'action sur un filtre linéaire de chocs aléatoires apparaissant à des instants aléatoires engendre un signal dénommé effet de grenaille. Le spectre d'entrée de ce filtre a été analysé par de nombreux auteurs et le but de cet exposé est d'abord de montrer que les calculs connus se simplifient beaucoup si l'on utilise une fonction dite de coïncidence. Il consiste ensuite à présenter des calculs explicites avec représentations graphiques de certaines densités spectrales permettant mieux qu'une formule mathématique d'en comprendre les propriétés physiques.

Abstract – The shot noise is the signal observed at the output of a linear filter when the input is a sequence of random pulses arriving at random times instants. The input spectrum of this filter has been analyzed in various papers and the first aim of this paper is to show that known expressions can be obtained very simply by using a coincidence function. In the second part we present examples of spectra and explicit calculations with graphical representations which yield a better understanding of the phenomena than mathematical expressions only.

1 Introduction et historique

L'effet de grenaille est le signal de sortie d'un filtre linéaire dont l'entrée est formée de chocs se produisant à des instants aléatoires $\{t_i\}$ constituant un processus ponctuel (PP). Il s'agit d'un phénomène macroscopique très courant dont la première mise en évidence expérimentale a été présentée dans [1]. Du point de vue mathématique c'est aussi l'un des plus anciens modèles de signal aléatoire étudié indépendamment dans [2, 3]. Dans tous ces travaux on supposait que les chocs aléatoires étaient de même amplitude et se produisaient aux instants d'un processus de Poisson stationnaire défini par sa densité λ. Ainsi le seul élément aléatoire dans ce phénomène était les instants $\{t_i\}$ du PP, ce qui conduisait à l'expression d'*effet de grenaille pur*. Dans ce cas on peut obtenir des expressions plus ou moins complexes de l'ensemble de la loi temporelle du signal.

Diverses généralisations ont été présentées. La plus simple consiste à introduire des amplitudes aléatoires de sorte que le signal effet de grenaille s'écrit

$$x(t) = \sum_i a_i h(t - t_i). \tag{1}$$

Le caractère aléatoire du signal provient alors des instants $\{t_i\}$ et des amplitudes $\{a_i\}$ souvent dénommées signal de *marque*, ou tout simplement marque. L'exemple le plus connu de signal marqué apparaît dans la détection optique des signaux faibles par un photomultiplicateur. À un photon absorbé à un instant t_i le système de multiplication associe un nombre aléatoire a_i de photoélectrons qui forment le signal de grenaille à la sortie du filtre de réponse percussionnelle $h(t)$. Lorsque les amplitudes a_i sont des variables aléatoires (VA) IID il est encore possible de calculer la loi temporelle du signal de sortie [4]. D'autres extensions ont été faites dans le cas de processus de Poisson composés largement utilisés en optique statistique pour décrire la détection de photons dans un champ électromagnétique [4]. En principe on peut associer un effet de grenaille à tout processus ponctuel (PP), mais même pour les processus de renouvellement qui sont les extensions les plus naturelles des processus de Poisson il est très difficile, voire impossible, d'obtenir sous forme explicite la loi temporelle du signal.

Toutefois on peut se contenter d'une vision plus simple en se limitant aux *propriétés du second ordre*. On peut introduire une fonction de corrélation ou un spectre d'entrée dont on déduit par les formules traditionnelles du filtrage linéaire la fonction de corrélation ou le spectre de $x(t)$. Dans le cas d'un processus de Poisson non marqué ($a_i = 1$) le spectre d'entrée est constant (bruit blanc) et égal à la densité λ du processus de Poisson. C'est ce qui apparaît dans [1, 2, 3]. Ce spectre blanc provient du fait que le processus de Poisson est sans mémoire. Cependant il y a bien d'autres types de PP ayant des effets de mémoire et la question s'est posée de savoir comment définir dans ces cas le spectre d'entrée, comment le calculer et en tirer des conséquences. Présentons un bref historique pour ensuite situer notre approche.

La première tentative de définir le spectre d'entrée d'un PP apparaît dans [5]. Cette approche est faite essentiellement à partir des propriétés de comptage du PP et l'auteur ne donne qu'un petit nombre d'exemples. La question a été reprise pour les PP de renouvellement dans [6]. Malgré des calculs laborieux, cet article indique déjà la relation entre le spectre d'entrée et les temps de vie du processus, ce qui sera analysé dans la suite. Enfin le cas des processus marqués a été étudié dans [7] et la formule générale qui s'y trouve est la plus complète à ce jour.

Notre approche se situe en amont et en aval de cette formule. En amont parce que l'usage des fonctions de coïncidences introduites au paragraphe suivant permet la démonstration de loin la plus simple de cette formule. En aval car tous les articles cités se contentent de formules algébriques dont la complexité rend difficile l'interprétation. Au contraire les représentations graphiques de quelques spectres présentées ci-dessous permettent d'en visualiser l'aspect, ce qui pour le praticien est la question la plus importante.

2 Fonctions de coïncidence

2.1 Processus non marqués

Soit N_1, N_2 et N_3 les nombres de points du PP apparaissant respectivement dans les intervalles $[\theta, \theta + d\theta[$, $[\theta', \theta' + d\theta'[$ et $[\theta + d\theta, \theta'[$, $\theta < \theta'$. Si le PP est régulier les VA N_1 et N_2 ne prennent que les valeurs 0 ou 1 et l'on a $P[N_1 = 1] = \lambda d\theta$, où λ est la densité du PP. Cette densité est constante dans le cas d'un PP stationnaire, ce qu'on admet dans la suite.

La fonction de coïncidence $c_N(t)$ est définie par la relation

$$\mathsf{E}(N_1 N_2) = c_N(\theta' - \theta) d\theta d\theta'. \quad (2)$$

Cette fonction est dénommée "densité de covariance" dans [5]. Ce terme ne nous paraît pas opportun car pour le lecteur rapide cela peut laisser entendre que $c_N(t)$ est une fonction de corrélation, comme on le trouve écrit dans divers articles d'optique statistique. Or ce n'est pas le cas et, comme on le verra dans la suite, sa transformée de Fourier peut très bien prendre des valeurs négatives.

Le calcul de cette fonction revient donc à celui de l'espérance mathématique $\mathsf{E}(N_1 N_2)$. Mais comme les VA N_1 et N_2 ne prennent que les valeurs 0 ou 1 il en est de même pour leur produit. On a donc $\mathsf{E}(N_1 N_2) = P(C)$, où C est l'événement coïncidence défini par $C = (N_1 = 1) \times (N_2 = 1)$. Le mot coïncidence est choisi car C s'obtient expérimentalement par un dispositif à coïncidence courant en physique des particules ou en optique statistique.

On a évidemment $P(C) = P[N_1 = 1] P[(N_2 = 1) | (N_1 = 1)]$. Cette probabilité conditionnelle peut s'exprimer au moyen des densités de probabilité (DDP) des *temps de vie successifs* du PP. On appelle temps de vie d'ordre k d'un PP la distance entre un point de ce PP et le $k$ième qui lui succède. Le temps de vie d'ordre 1 est évidemment la distance entre deux points successifs. Il résulte de cette définition que $f_k(\theta' - \theta) d\theta'$ est la probabilité conditionnelle $P[(N_2 = 1)(N_3 = n - 1) | (N_1 = 1)]$. La probabilité conditionnelle apparaissant dans la coïncidence est donc la somme de ces quantités puisque k n'est pas fixé. Notant finalement que la fonction de coïncidence est par définition paire, alors que les fonctions $f_n(t)$ sont nulles pour $t < 0$, on en déduit que

$$c_N(t) = \lambda \sum_{k=1}^{\infty} f_k(|t|). \quad (3)$$

On peut montrer que cette fonction est en général bornée et tend vers λ^2 quand $t \to \infty$.

2.2 Processus marqués

Soit $M(t)$ le signal aléatoire en escalier variant par sauts d'amplitudes a_i aux points t_i du processus ponctuel de départ. Il en résulte que les accroissements infinitésimaux $dM(\theta) = dM(\theta + d\theta) - dM(\theta)$ ne prennent que les valeurs 0 ou a_i. Ce dernier événement se réalise si $dN(\theta) = 1$, où $N(t)$ est la fonction utilisée précédemment, et sa probabilité vaut donc $\lambda d\theta$. Soit M_1 et M_2 les VA respectivement égales à $dM(\theta)$ et $dM(\theta')$.

Supposons d'abord que le signal a_i soit déterministe (non aléatoire). La VA M_1 ne prend que les valeurs 0 ou a_j, et cette dernière avec la probabilité $\lambda d\theta$. D'après la définition du temps de vie on a

$$P[(M_1 = a_j)(M_2 = a_{j+k})] = \lambda f_k(\theta' - \theta) d\theta d\theta'. \quad (4)$$

Il en résulte que

$$\mathsf{E}[M_1 M_2] = \lambda \sum_{k=1}^{\infty} a_j a_{j+k} f_k(\theta' - \theta) d\theta d\theta' \quad (5)$$

où les $\{a_i\}$ ne sont pas aléatoires.

En prenant maintenant la moyenne sur les a_i quand ils deviennent aléatoires et en utilisant la même formule que (2), on déduit que

$$c_M(t) = \lambda \sum_{k=1}^{\infty} \mathsf{E}[a_j a_{j+k}] f_k(t). \quad (6)$$

Lorsque le signal a_i est stationnaire on a

$$\mathsf{E}[a_j a_{j+k}] = \gamma_a[k] + m_a^2, \quad (7)$$

où m_a et $\gamma_a[k]$ sont respectivement la moyenne et la fonction de corrélation (FC) des a_i. Reprenant le même raisonnement que ci-dessus concernant la parité de $c(t)$ on déduit

$$c_M(t) = \lambda \sum_{k=1}^{\infty} (\gamma_a[k] + m_a^2) f_k(|t|). \quad (8)$$

Cette formule englobe évidement (3) puisque dans le cas de l'effet de grenaille pur les a_i sont tous égaux à un. Ceci donne $\gamma_a[k] = 0$ et $m_a^2 = 1$. On est alors conduit à comparer la fonction de coïncidence du processus marqué et celle du processus non marqué. On déduit de (3) et (8) la relation

$$c_M(t) = m_a^2 c_N(t) + s(t) \quad (9)$$

avec

$$s(t) = \lambda \sum_{k=1}^{\infty} \gamma_a[k] f_k(|t|). \quad (10)$$

Dans le cas d'un signal de marque centré on a $m_a = 0$ d'où $c_M(t) = s(t)$.

3 Fonction de corrélation et spectre d'entrée

La relation (1) peut s'écrire

$$x(t) = \int h(t - \theta) dM(\theta) \quad (11)$$

puisque l'accroissement $dM(\theta)$ est nul, sauf aux instants t_i où il vaut a_i. On a évidemment $\mathsf{E}[dM(\theta)] = \lambda m_a d\theta$, ce qui permet de calculer la valeur moyenne $m_x = \mathsf{E}[x(t)]$. Le calcul de la moyenne du second ordre $r_x(\tau) = \mathsf{E}[x(t)x(t - \tau)]$ est un peu plus compliqué car il nécessite la connaissance complète du moment du second ordre $\mathsf{E}[dM(\theta) dM(\theta')]$. Pour $\theta \neq \theta'$ il s'écrit à l'aide de la fonction de coïncidence définie par (8). Mais il faut une formule englobant le cas $\theta = \theta'$. Or dans ce cas $\mathsf{E}[dM^2(\theta)]$ vaut $\lambda \mathsf{E}(a^2) d\theta$, ce qui donne, $\forall\ \theta$ et θ'

$$\mathsf{E}[dM(\theta) dM(\theta')] = [c_M(\theta - \theta') + \delta(\theta - \theta') \lambda \mathsf{E}(a^2)] d\theta d\theta' \quad (12)$$

où $\delta(.)$ est la distribution de Dirac. La FC de $x(t)$ vaut par définition $\gamma_x(t) = r_x(t) - m_x^2$. Par un calcul immédiat utilisant les résultats précédents on trouve que $\gamma_x(t)$ se déduit par

les relations classiques de transformation par filtrage linéaire d'une FC d'entrée valant

$$\gamma_M(t) = \lambda \mathsf{E}(a^2)\delta(t) + c_M(t) - (\lambda m_a)^2 \qquad (13)$$

où $c_M(t)$ est la fonction de coïncidence définie par (8). Utilisant (9) et (10) on obtient également

$$\gamma_M(t) = \lambda \mathsf{E}(a^2)\delta(t) + m_a^2 \hat{c}_N(t) + s(t) \qquad (14)$$

où $\hat{c}_N(t) = c_N(t) - \lambda^2$ tend vers 0 quand $t \to \infty$.

On peut faire deux remarques sur ces formules. Tout d'abord l'existence du terme $\delta(t)$ rend infinie la FC pour $t = 0$. Ceci signifie que le signal d'entrée générant $x(t)$ par filtrage linéaire n'est pas du second ordre, ayant une variance infinie. Par transformation de Fourier ce terme introduit dans le spectre une composante constante $\lambda \mathsf{E}(a^2)$. On l'appellera dans la suite la *composante blanche*. En particulier dans le cas de l'effet de grenaille pur $\gamma_M(t)$ se réduit à $\lambda\delta(t)$, et l'on retrouve le spectre constant et égal à λ noté au début de ce texte. Dans ce cas le signal d'entrée est un bruit blanc et cela provient du fait qu'un PP de Poisson est sans mémoire.

Par ailleurs il est facile de voir que $\gamma_M(t)$ tend vers 0 quand $t \to \infty$, et ceci signifie que dans le domaine spectral il n'y a pas en général de raie à la fréquence nulle.

Passons maintenant à l'étude du spectre (ou densité spectrale de puissance) de $x(t)$ qui selon (11) et (14) vaut $\Gamma_x(\nu) = |G(\nu)|^2 \Gamma_M(\nu)$ où $G(\nu)$ est la réponse en fréquence du filtre et

$$\Gamma_M(\nu) = \lambda \mathsf{E}(a^2) + m_a^2 \hat{C}_N(\nu) + S(\nu) = \lambda\sigma_a^2 + m_a^2 \Gamma_N(\nu) + S(\nu), \qquad (15)$$

où $S(\nu)$ est la TF de la fonction $s(t)$ définie par (10) et σ_a^2 la variance de a. Lorsque les a_i sont IID le dernier terme disparaît, et l'on trouve une expression constituant le théorème 1 de [8] démontré de manière beaucoup plus complexe.

4 Applications aux processus de renouvellement

Ce sont des PP définis par le fait que les intervalles entre points successifs sont des VA IID. Ainsi le PP est entièrement défini par la DDP du temps de vie d'ordre 1. Lorsque cette DDP est une exponentielle monolatérale, le PP de renouvellement devient un processus de Poisson. L'indépendance entre les intervalles entraîne que la DDP $f_k(t)$ apparaissant dans les expressions précédentes est une puissance de convolution d'ordre k de la DDP du temps de vie d'ordre 1. On en déduit que $c(0^+) = 0$ et que si la DDP $f(t)$ est bornée il en est de même pour $f_k(t)$.

Ceci conduit à utiliser les transformées de Laplace (TL) monolatérales. Plus précisément soit $F_L(s)$ et $C_L(s)$ les TL respectivement de $f(t)$ et $c(t)$. Il résulte de (8) que

$$C_L(s) = \lambda \sum_{k=1}^{\infty} (\gamma_a[k] + m_a^2) F_L^k(s). \qquad (16)$$

On vérifie simplement que cette série et convergente dans le demi-plan de droite.

Nous avons noté que la fonction de coïncidence était paire. Il en résulte que sa TF peut se déduire de la TL monolatérale par la relation

$$C_F(\omega) = 2\mathrm{Re}[C_L(j\omega)]. \qquad (17)$$

Lorsque l'on prend la partie réelle de la TL le dernier terme de (13) est éliminé car sa TL est en $1/s$. Les deux autres termes donnent alors le spectre $\Gamma_M(\omega)$, TF de $\gamma_M(t)$, soit

$$\Gamma_M(\omega) = \lambda \mathsf{E}(a^2) + 2\mathrm{Re}[C_L(j\omega)]. \qquad (18)$$

Il résulte de ceci que le spectre d'entrée $\Gamma_M(\omega)$ de $x(t)$ se déduit directement de la fonction de coïncidence, ce qui en montre l'intérêt. Donnons maintenant quelques exemples.

Si le processus n'est pas marqué ($\gamma_a = 0$, $m_a = 1$) on a

$$C_L(s) = \lambda \, \frac{F_L(s)}{1 - F_L(s)}. \qquad (19)$$

On en déduit aisément que $c_N(t)$ tend vers λ^2 quand $t \to \infty$.

Si le processus est marqué le calcul devient plus difficile à cause du terme $\gamma_a[k]$. Toutefois il se simplifie pour toutes les corrélations en somme d'exponentielles correspondant à des marques de type ARMA. Le cas le plus élémentaire se produit quand la marque a_i est un signal AR (1) centré et de variance unité, ce qui donne $\gamma[k] = r^{|k|}$ où r est le coefficient de régression. Dans ce cas on a

$$C_L(s) = \lambda \, \frac{rF_L(s)}{1 - rF_L(s)}. \qquad (20)$$

Pour poursuivre le calcul il faut donc se donner la fonction $F_L(s)$, TL de la DDP du temps de vie. Dans cet exposé on se limite à l'exemple où cette DDP est uniforme, soit $f(t) = 0$, sauf dans l'intervalle $[m-b, m+b]$ où $f(t) = 1/2b$. Il est clair que le paramètre b est en relation biunivoque avec la variance du temps de vie. Cette DDP permet de décrire un PP avec une incertitude uniforme sur la distance entre points (phénomène de gigue). Quand $b \to 0$ on arrive à une distribution périodique de points. La TL de cette DDP est aisée à calculer et l'on trouve $C_L(j\omega) = (1/m)N/D$ avec

$$N(\omega) = \sin(b\omega)\exp(jm\omega) \qquad (21)$$

et

$$D(\omega) = b\omega - \sin(b\omega)\exp(jm\omega). \qquad (22)$$

Il suffit maintenant d'appliquer (18) soit avec (19) soit avec (20). Les résultats sont calculés avec la fréquence ordinaire $\nu = \omega/2\pi$ et l'on suppose dans tous les cas, pour simplifier, que $\lambda = 1$. Les résultats pour d'autres valeurs de la densité s'en déduisent immédiatement.

Sur la **figure 1** on considère le cas d'un PP non marqué. Supposant $m = 1$ on étudie l'influence du paramètre b définissant la DDP. Les courbes 1, 2 et 3 correspondent respectivement aux valeurs 1, 0,5 et 0,003 de b. La courbe 3 s'explique aisément. En effet si dans (1) on a $a_i = 1$ et $t_i = iT$ le signal $x(t)$ devient un peigne de Dirac dont la TF est également un tel peigne. Pour $m = 1$ les fréquences de ce peigne sont les entiers, ce l'on voit clairement sur la courbe. L'influence du paramètre b est particulièrement nette dans le domaine des basses fréquences. À mesure que b décroît les oscillations de $\Gamma(\nu)$ autour de la fréquence 1 diminuent pour tendre vers une simple résonance devenant d'ailleurs infinie quand $b = 0$. De plus, comme $m_a = \lambda = 1$, la composante blanche du spectre vaut 1 et elle est atteinte d'autant plus vite que b est plus grand. Enfin il résulte de (21) que le spectre vaut 1 pour les fréquences annulant $N(\omega)$, soit $k/2b$, et l'on peut montrer que $\Gamma(0) = b^2/3$. Ceci se voit clairement sur la figure.

Sur la **figure 2** on montre l'influence de la marque caractérisée par la régression r. Les résultats sont présentés pour les fréquences comprises entre 0 et 2. On a choisi, comme dans la figure 1, $m = 1$ et $b = 0,003$ et on étudie l'influence de la régression r. La courbe 1 correspond à une régression nulle. Dans ce cas les VA a_i sont IID et de moyenne nulle. Le spectre se réduit à sa composante blanche $\lambda \mathsf{E}(a^2)$ qui vaut 1 pour les valeurs choisies des paramètres. La courbe 4 correspond à $r = 1$ ce qui signifie l'absence de marque et l'on retrouve la situation de la figure 1. Les courbes 2 et 3 correspondent respectivement aux valeurs 0,2 et 0,8 de r. On voit que plus la régression diminue, plus la résonance aux fréquences entières est amortie pour disparaître complètement quand $r = 0$.

La **figure 3** présente l'influence de la moyenne m_a d'un processus marqué dont la composante blanche vaut 1. Le calcul est fait avec les conditions $b = 0,5$ et $r = 0,3$. Les valeurs de m_a sont 1, 0,8, 0,695 et 0 respectivement pour les courbes 1, 2, 3 et 4. Les spectres doivent prendre la valeur 1 aux fréquences entières et au delà de la fréquence 1 ils sont très semblables. Au contraire l'influence de la moyenne de la marque est très sensible pour les fréquences inférieures à 1. Pour $m_a = 1$ on retrouve la courbe 2 de la figure 1, ce qui signifie que l'on a un PP non marqé et la courbe 1 ne dépend pas de r. Par contre pour $m_a = 0$ on retrouve la situation de la figure 2. Enfin pour la valeur $m_a = 0,695$ le spectre prend pour la fréquence nulle sa valeur asymptotique 1. D'autres résultats seront présentés dans l'exposé oral.

Ces résultats peuvent être généralisés au cas où le modèle AR(1) est remplacé par un modèle ARMA avec un nombre fini de pôles. Dans l'exposé oral on présentera des courbes correspondant au cas de deux pôles complexes conjugués et on étudiera plus particulièrement l'influence du module et de la phase sur le spectre d'entrée.

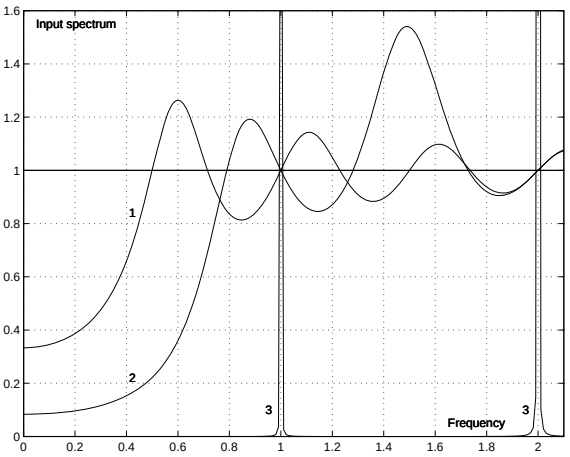


FIG. 1: *Influence du paramètre b.*

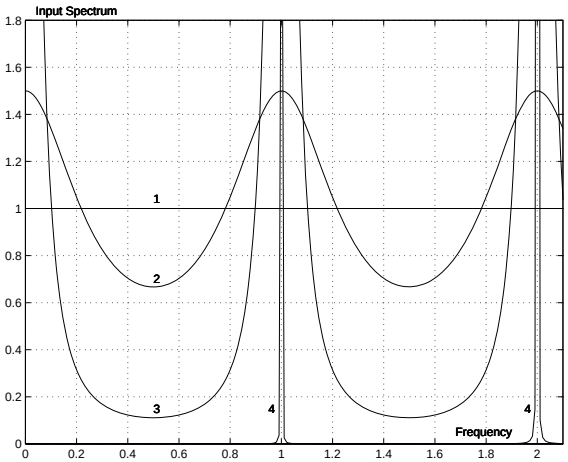


FIG. 2: *Influence de la régression de la marque.*

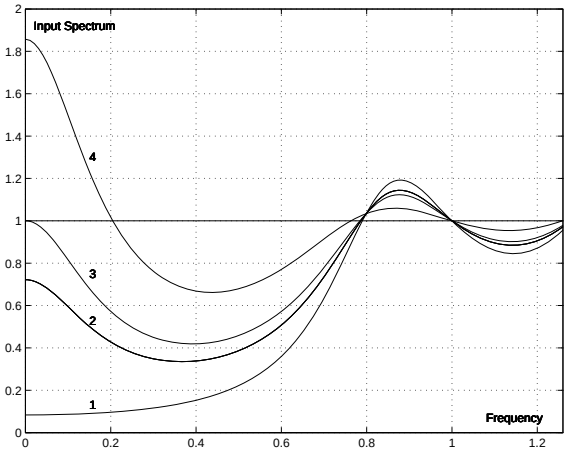


FIG. 3: *Influence de la moyenne m_a.*

Références

[1] Blanc-Lapierre A., *Étude expérimentale de l'effet de grenaille pur, application á l'amplification des courants photoélectriques très faibles,* Paris : Masson, 1944.

[2] Blanc-Lapierre A., *Sur certaines fonctions aléatoires stationnaires; applications à l'étude des fluctuations dues à la structure électronique de l'electricité,* Paris : Masson, 1945.

[3] Rice S. O., "Mathematical analysis of random noise," *Bell Systems Technical Journal*, 24, pp. 46-156, 1945.

[4] Picinbono B., *Signaux aléatoires, tome 2,* Paris : Dunod université, 1993.

[5] Bartlett M. S., "The spectral analysis of point processes," *J. Roy. Statist. Soc. B*, 25, 264-296, 1962.

[6] Kryukov V. L., "Calculation of the correlation function and the spectral power density of random sampling," *Radio Eng. and Electr. Physics*, 12, 169-176, 1967.

[7] Beutler J. and Leneman O. "The spectral analysis of impulse processes," *Information and Control*, vol. 12, pp. 236-258, 1968.

[8] Brémaud P. and Massoulié L., "Power spectra of general shot noises and Hawkes point processes with a random excitation," Adv. Appl. Probab. , 34, 205-222, 2002.

Sur la Décomposition Modale Empirique des signaux échantillonnés

Gabriel Rilling, Patrick Flandrin

Laboratoire de Physique de l'ENS–Lyon
46 allée d'Italie, 69364 Lyon Cedex 07, France
Gabriel.Rilling@ens-lyon.fr,Flandrin@ens-lyon.fr

Résumé – On cherche ici à mieux comprendre le fonctionnement d'une technique récente de décomposition de signaux appelée Décomposition Modale Empirique (ou EMD pour "Empirical Mode Decomposition") et plus particulièrement l'influence de l'échantillonnage sur son comportement. Pour ce faire, on lui soumet des signaux sinusoïdaux échantillonnés en faisant varier les paramètres de l'échantillonnage. Les effets dus à l'échantillonnage observés sont tout d'abord caractérisés et quantifiés avant de faire l'objet d'une justification théorique.

Abstract – This paper is intended to give new insights into a recent technique of signal decomposition, referred to as "Empirical Mode Decomposition" (EMD) and more precisely into the effect of sampling on its behavior. Therefore we analyze the EMD of sampled sinusoidal signals while varying the sampling parameters. The observed effects related to sampling are first characterized and quantified and then justified theoretically.

1 Introduction

N.E. Huang et ses collaborateurs ont récemment introduit une technique de décomposition de signaux non stationnaires, appelée *Décomposition Modale Empirique* (ou EMD, pour "Empirical Mode Decomposition") [1]. Si celle-ci possède un caractère très intuitif, elle est restée jusqu'à présent rêtive à une véritable évaluation théorique dans la mesure où sa définition n'est en fait donnée que par l'algorithme de décomposition proposé. L'EMD a néanmoins été appliquée à un certain nombre de situations réelles et a motivé des études systématiques de la méthode dans des conditions contrôlées afin de mieux cerner ses possibilités et limitations (cf., par exemple, [2, 3]).

2 Le problème de l'échantillonnage

L'EMD est un algorithme récursif dans lequel on itère une opération qui permet de séparer tout signal $x(t)$ en deux composantes, l'une oscillant à tout instant plus rapidement que l'autre. Ces deux composantes, qu'on appellera *détail local* et *moyenne locale*, sont usuellement obtenues par la procédure [1] :

1. extraire les extrema du signal $x(t)$
2. en déduire une enveloppe supérieure (resp. inférieure) par interpolation des maxima (resp. minima)
3. définir une *moyenne locale* $m(t)$ comme étant la demi-somme des enveloppes.
4. en déduire un *détail local* $d(t) = x(t) - m(t)$

En pratique, la procédure précédente, dite de « sifting », est généralement itérée sur le *détail local* pour affiner la décomposition. On notera cependant que le résultat obtenu après une seule itération constitue habituellement une bonne approximation du résultat final, les itérations suivantes apportant des contributions moins importantes. On pourra trouver davantage de détails sur l'EMD dans [1, 4] et des codes `MATLAB` la mettant en œuvre sur internet [5].

Dans l'exposition qu'on vient de faire du principe de l'EMD, on a implicitement considéré les signaux comme étant à temps continu. Pourtant, en pratique, l'EMD ne peut être mise en œuvre que pour des signaux à temps discret. Or lorsqu'un signal est échantillonné, la position et l'amplitude de chacun de ses extrema est susceptible de varier. Compte tenu du rôle clé que jouent ces points dans l'EMD, ces variations peuvent avoir des répercussions sensibles sur la décomposition. C'est pourquoi il est généralement recommandé de n'appliquer l'EMD qu'à des signaux largement suréchantillonnés de manière à pouvoir négliger ces effets.

3 Quantification de l'influence de l'échantillonnage

3.1 Procédure d'évaluation

À l'heure actuelle, il n'existe aucun argument théorique permettant de distinguer un échantillonnage suffisant ou insuffisant. Dans cette optique et dans la continuité de [4], nous nous sommes intéressés au comportement de l'EMD lorsqu'on lui soumet un signal sinusoïdal échantillonné

$$x_{f,\varphi}[n] = \cos(2\pi f n + \varphi), \text{ avec } 1 \leq n \leq N,$$

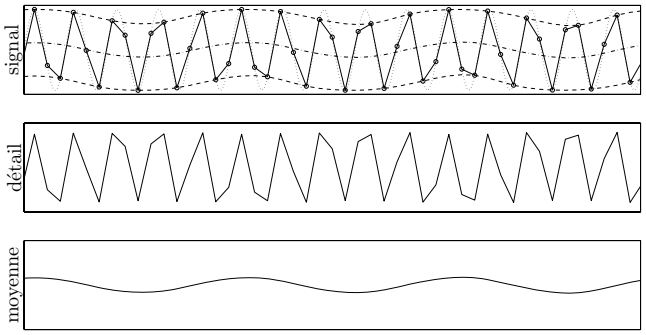


FIG. 1 – Illustration de l'influence de l'échantillonnage — *En haut : le signal sinusoïdal continu (pointillés), échantillonné (trait plein), ses enveloppes (tirets) et leur demi-somme (tiret-point). Au milieu et en bas : le détail local et la moyenne locale correspondants.*

en fonction de sa fréquence réduite f et de la phase φ. Sachant que l'équivalent à temps continu d'un tel signal est considéré comme une unique composante par l'EMD (sa *moyenne locale* étant nulle), on pourrait imaginer qu'il en soit de même avec la version échantillonnée. Cependant, l'échantillonnage modifiant les amplitudes des extrema, l'EMD le sépare en deux composantes (cf. figure 1). Parmi celles-ci, la composante *détail local* est plutôt proche du signal analysé tandis que la *moyenne locale*, apparaît comme une trace des effets de l'échantillonnage. C'est pourquoi, on choisit pour quantifier cette influence de s'intéresser à l'énergie de la *moyenne locale*. Plus précisément, on regarde l'évolution de la quantité

$$e(f,\varphi) = \left(\frac{\sum_{n=1}^{N} m_{f,\varphi}^2[n]}{\sum_{n=1}^{N} x_{f,\varphi}^2[n]} \right)^{1/2}$$

où $m_{f,\varphi}$ est la *moyenne locale* de $x_{f,\varphi}$.Dans la plupart des cas, la valeur de $e(f,\varphi)$ est pratiquement indépendante de φ. C'est pourquoi on s'intéressera également à la quantité moyennée par rapport à φ, notée $\bar{e}(f)$.

Pour mesurer ces quantités, nous avons effectué une série de simulations avec des échantillons de taille $N = 2048$. Les résultats sont présentés figure 2.

3.2 Interprétation — Modélisation

L'analyse de ces mesures met en évidence essentiellement deux caractéristiques. Premièrement, on constate que $e(f,\varphi)$ s'annule pour toutes les fréquences de la forme $f = 1/2k$ avec k entier. Deuxièmement, on voit qu'il est possible de trouver une majoration de $e(f,\varphi)$ de la forme $e(f,\varphi) \leq \lambda f^2$.

Pour ce qui est de l'interprétation, on peut aisément vérifier que lorsque $f = 1/2k$, le signal échantillonné a des enveloppes supérieure et inférieure constantes et symétriques. Il s'ensuit que sa *moyenne locale* est nécessairement nulle et donc $e(f,\varphi)$ aussi.

En ce qui concerne la majoration, on démontrera dans le paragraphe suivant que

$$e(f,\varphi) \leq \frac{1-\cos\pi f}{\sqrt{2}} \leq \frac{\pi^2 f^2}{2\sqrt{2}}. \qquad (1)$$

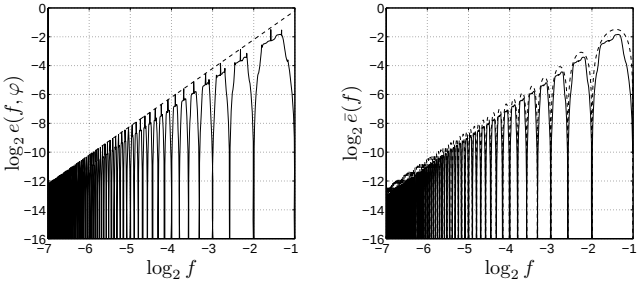


FIG. 2 – Influence de l'échantillonnage en fonction de la fréquence — *À gauche, $e(f,\varphi)$ en fonction de la fréquence f pour $\varphi = 0$. À droite, $\bar{e}(f) :=$ « $e(f,\varphi)$ moyenné sur toutes les valeurs de φ possibles ». À gauche, on met en évidence une majoration par une droite de pente 2 sur le diagramme log-log correspondant à une majoration de la forme λf^2. À droite, on a tracé en tirets le modèle correspondant à la section 3.2.2*

On donnera également un modèle plus perfectionné fournissant un bon ordre de grandeur de $\bar{e}(f)$ permettant de rendre compte à la fois des deux caractéristiques mentionnées précédemment.

3.2.1 Majoration

La preuve du résultat (1) proposée ici est obtenue à l'aide de trois hypothèses simplificatrices :

1. L'enveloppe supérieure (resp. inférieure) interpolant les maxima (resp. minima) est toujours comprise entre la valeur du plus petit maximum (resp. plus grand minimum) et 1 (resp. -1).
2. Une unique itération de sifting suffit à séparer les deux composantes.
3. L'énergie de $x_{f,\varphi}$ vaut $N/2$.

Compte tenu de l'hypothèse 2, on peut exprimer la moyenne locale sous la forme

$$|m_{f,\varphi}| = \left| \frac{e_{min} + e_{max}}{2} \right| = \frac{|e_{max}| - |e_{min}|}{2},$$

où on note e_{max} et e_{min} les enveloppes supérieure et inférieure de $x_{f,\varphi}$. L'hypothèse 1 fournit alors la majoration

$$|m_{f,\varphi}| \leq \frac{1-\alpha}{2},$$

où α est la valeur absolue du plus grand des minima de $x_{f,\varphi}$. Sachant qu'on peut montrer que $\alpha \geq \cos\pi f$, l'hypothèse 3 permet d'obtenir la première partie du résultat (1). La seconde partie découle alors directement du résultat classique « $\cos u \geq 1 - u^2/2$ ».

On pourra noter de plus que la majoration correspondant à la première partie du résultat (1) est atteinte dans le cas $\varphi = 0$ pour toutes les fréquences de la forme $f = 1/(2k+1)$ avec k entier. Par conséquent, la majoration obtenue est optimale.

3.2.2 Ordre de grandeur

S'il rend bien compte de la majoration en f^2, le modèle qu'on vient de présenter ne permet pas de justifier que

l'influence de l'échantillonnage s'annule pour toutes les fréquences de la forme $f = 1/2k$. Pour ce faire, il faut analyser plus finement le comportement de la *moyenne locale*. Si on reprend les mêmes hypothèses que pour le modèle de majoration, cette dernière est simplement la demi-somme des enveloppes supérieure et inférieure. Le problème se ramène donc à l'évaluation de l'ordre de grandeur de chacune des enveloppes. L'idée consiste alors à considérer que l'enveloppe supérieure (resp. inférieure) est localement égale à la valeur du maximum (resp. minimum) le plus proche. De là, on obtient

$$m^2_{f,\varphi}[n] \approx \frac{[\cos(2\pi f n_{max} + \varphi) + \cos(2\pi f n_{min} + \varphi)]^2}{4},$$

où n_{max} et n_{min} correspondent aux positions des maximum et minimum les plus proches. Pour tout maximum situé à une position n_{max}, il existe un entier k tel que $\phi_{max} = 2\pi f n_{max} + \varphi - 2k\pi \in [-\pi f, \pi f]$. Dans la très grande majorité des cas, les différentes valeurs de ϕ_{max} correspondant aux différents maxima sont bien réparties dans $[-\pi f, \pi f]$. Par conséquent, on peut donner un ordre de grandeur de l'énergie de la *moyenne locale* de la façon suivante :

$$\sum_{n=1}^{N} m^2_{f,\varphi}[n] \approx N\mathbb{E}_\psi \frac{[\cos(\psi) + \cos(2\pi f n_{min}(\psi) + \psi)]^2}{4}$$

où on prend ψ uniformément réparti dans $[-\pi f, \pi f]$ et $n_{min}(\psi)$ correspond à la position du premier minimum de $\cos(2\pi f n + \psi)$, qu'on peut calculer en fonction de ψ. Finalement, le calcul aboutit à une expression de $\bar{e}(f)$ qui se simplifie sous la forme suivante

$$\bar{e}(f) \approx \frac{2\pi^2 f^2}{\sqrt{3}}(K + 1 - \frac{1}{2f})(\frac{1}{2f} - K)),$$

où K est la partie entière de $1/2f$. Cette dernière expression permet de retrouver à la fois le comportement en f^2, mis en évidence dans le préfacteur, et le fait que $\bar{e}(f)$ s'annule pour les fréquences de la forme $f = 1/2k$ (cf figure 2 à droite).

4 Analyse spectrale de l'influence de l'échantillonnage

On s'intéresse maintenant à la manière dont se manifestent les effets de l'échantillonnage quantifiés dans la partie précédente. Par construction la composante *moyenne locale* correspond à des oscillations plus lentes que le *détail local* qui oscille à la fréquence f comme le signal analysé. Par conséquent, l'EMD d'un signal sinusoïdal échantillonné fait apparaître de nouvelles fréquences inférieures à f. Pour tenter de caractériser ces nouvelles fréquences, on s'est intéressé au spectre de la composante *moyenne locale* en fonction de f. Ce spectre étant constitué de raies, on a représenté figure 3 les fréquences correspondant aux principaux pics du spectre de $m_{f,\varphi}$ en fonction de f.

L'analyse des spectres en fonction de f met en évidence l'existence de deux cas suivant la régularité du spectre. Celui-ci est constitué soit d'une série de pics régulièrement

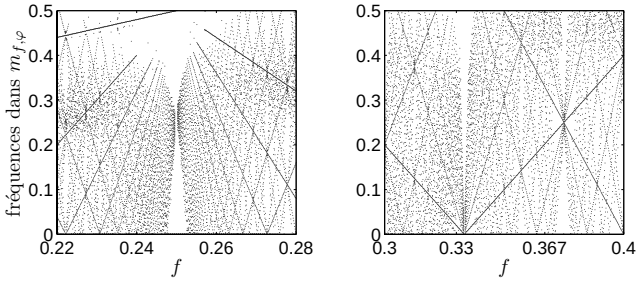


FIG. 3 – Répartition spectrale de l'influence de l'échantillonnage — *Pour chaque valeur de f, en abscisse, on représente, en ordonnée, une série de points correspondant aux pics apparaissant dans le spectre de la* moyenne locale *et du* détail local *de $x_{f,\varphi}$. À gauche, on observe ce qui se passe autour de $f = 1/4$. À droite, autour de $f = 1/3$. Plus généralement, on retrouve les mêmes allures autour de toutes les fréquences de la forme $f = 1/2k$ et $1/(2k+1)$ avec k entier.*

espacés soit de pics répartis de manière dense sans régularité apparente. Dans le premier cas, l'analyse des diagrammes de la figure 3 permet de mettre en évidence que les fréquences f' qui apparaissent dans les spectres sont de la forme $f' = (2k+1)f + k'$ avec k, k' entiers, ou encore $f' \in f + 2f\mathbb{Z} + \mathbb{Z}$.

4.1 Interprétation

L'interprétation des spectres obtenus est liée à la nature non-linéaire de l'EMD. Si une grandeur x_1 dépend de manière non-linéaire d'une autre grandeur x_2 dont le spectre ne contient que deux fréquences f_1 et f_2, on sait que le spectre de x_1 ne peut contenir des fréquences que dans l'ensemble $f_1\mathbb{Z} + f_2\mathbb{Z}$. Plus généralement, le spectre de x_1 est susceptible de contenir toutes les fréquences correspondant à des combinaisons linéaires à coefficients entiers des fréquences présentes dans le spectre de x_2. Dans le cas qui nous intéresse, le spectre de $x_{f,\varphi}$, le signal échantillonné, contient les fréquences $\pm f + k$ avec k entier. Par conséquent, le spectre de $m_{f,\varphi}$ ne peut contenir que des fréquences appartenant à $f\mathbb{Z} + \mathbb{Z}$.

De là, deux cas sont possibles. Si la fréquence f est irrationnelle, le spectre est alors constitué de pics répartis de manière dense. Ce cas correspond bien à un des deux types de spectre observés.

Supposons maintenant $f = p/q$ rationnel. On sait alors que l'ensemble $f\mathbb{Z} + \mathbb{Z}$ est égal à l'ensemble $1/q\mathbb{Z}$. Dans le cas q impair, on ne peut rien dire de plus étant donné que $f\mathbb{Z} + \mathbb{Z} = 1/q\mathbb{Z} = f + 2f\mathbb{Z} + \mathbb{Z}$. Dans le cas où $q = 2q'$ est pair, le signal $x_{f,\varphi}$ est q-périodique avec la propriété suivante :

$$\begin{aligned} \forall n, \quad x_{f,\varphi}[n+q'] &= \cos\left(2\pi\frac{p}{2q'}(n+q') + \varphi\right) \\ &= -x_{f,\varphi}[n]. \end{aligned}$$

Sachant que dans le cas de sinus échantillonnés, tous les maxima sont strictement positifs et tous les minima strictement négatifs, on en déduit que les positions des minima

peuvent se déduire de celles des maxima par un décalage de q'. Par conséquent, si on ignore les effets de bords dus à la taille limitée des signaux, l'enveloppe inférieure e_{min} de $x_{f,\varphi}$ peut se déduire de son enveloppe supérieure e_{max} par

$$\forall n, \quad e_{min}[n+q'] = -e_{max}[n].$$

À partir de là, on obtient que la *moyenne locale* m_1 obtenue après une seule itération de sifting aura également la propriété

$$\forall n, \quad m_1[n+q'] = -m_1[n].$$

Si on effectue plusieurs itérations de sifting cette propriété se transmet tout au long du processus et finalement, les deux composantes *moyenne locale* et *détail local* définitives la vérifient également. Si on regarde alors la composante de Fourier correspondant à la fréquence k/q dans le spectre de la *moyenne locale* $m_{f,\varphi}$, on peut écrire

$$\begin{aligned}\sum_n m_{f,\varphi}[n] e^{-2i\pi\frac{k}{2q'}n} &= -\sum_n m_{f,\varphi}[n+q'] e^{-2i\pi\frac{k}{2q'}n} \\ &= -\sum_n m_{f,\varphi}[n] e^{-2i\pi\frac{k}{2q'}(n-q')} \\ &= (-1)^{k+1}\sum_n m_{f,\varphi}[n] e^{-2i\pi\frac{k}{2q'}n}\end{aligned}$$

ce qui implique que, si k est pair, la composante de Fourier à la fréquence k/q est nulle. Finalement, le spectre ne peut contenir que les fréquences de $1/q+2/q\mathbb{Z}$ qui est identique à $f+2f\mathbb{Z}+\mathbb{Z}$, ce qui correspond bien aux observations.

Dans l'interprétation qu'on vient de donner, on s'est appuyé sur la distinction rationnel/irrationnel. En pratique, cette distinction est un peu différente de sa définition mathématique. En effet, comme on réalise des estimations spectrales avec des résolutions limitées, le spectre de la *moyenne locale* obtenu avec une fréquence f irrationnelle peut être très semblable à celui obtenu avec une fréquence f' rationnelle proche de f. Par conséquent, il est tout à fait possible d'obtenir un spectre estimé dense avec une fréquence f rationnelle et un spectre estimé constitué de pics espacés régulièrement avec une fréquence f irrationnelle.

Plus précisément, si M est la longueur sur laquelle on calcule la transformée de Fourier pour l'estimation spectrale, le spectre de la *moyenne locale* correspondant à une fréquence f proche d'un rationnel p/q avec $q \lesssim M$ est très semblable à celui obtenu avec la fréquence p/q. En revanche si f n'est pas particulièrement proche d'un tel rationnel, alors le spectre de la *moyenne locale* apparaît comme un ensemble de pics dense comme dans le cas d'une fréquence irrationnelle. Si l'intervalle sur lequel on peut considérer que f est proche d'un rationnel p/q reste relativement confus, on peut tout de même signaler que sa taille est d'autant plus grande que q et M sont petits.

5 Conclusion

La décomposition modale empirique (ou « EMD ») est une méthode d'analyse de signaux qui a déjà motivé un certain nombre d'études. Étant donné qu'on n'a pour le moment qu'une compréhension limitée de la méthode, une grande partie de ces travaux se concentrent sur l'analyse des performances de l'EMD appliquée à des signaux réels. Cependant, que l'approche proposée soit plutôt pratique ou plutôt théorique, la question de l'échantillonnage est généralement simplement ignorée. Or dans tous les cas étudiés, les données analysées sont implicitement largement suréchantillonnées. De fait, il apparaît que ce choix se justifie dans la mesure où les effets de l'échantillonnage sur l'EMD sont du même ordre de grandeur que les déplacements des extrema occasionnés et donc que les effets d'échantillonnage sur le signal analysé.

Dans un souci à la fois d'explorer le comportement de l'EMD et d'évaluer plus précisément la confiance qu'on peut y accorder en fonction de l'échantillonnage, nous nous sommes intéressés au cas de signaux sinusoïdaux. Cet exemple simple nous a permis de développer une première approche théorique au problème de l'échantillonnage. Parmi les résultats obtenus, on a montré en 3.2.1 que ces effets pouvaient être majorés par un terme proportionnel au carré de la fréquence réduite, soit inversement proportionnel au carré de la fréquence d'échantillonnage f_e. Or il apparaît que les arguments déployés pour justifier cette majoration restent en grande partie valables tant que le signal analysé posséde une dérivée continue. La confrontation avec la pratique, en revanche, montre que cette majoration n'est en fait valable que dans la mesure où l'amplitude des oscillations varie peu à l'échelle d'une oscillation. Ceci est lié au fait que pour établir cette majoration, nous n'avons tenu compte que des variations d'amplitude des extrema. Si l'on tient compte en plus des variations de position, on peut montrer qu'un nouveau terme s'ajoute au précédent pour finalement obtenir une majoration de la forme $\lambda/f_e^2 + \mu/f_e$. Enfin, si l'on ajoute à cela qu'on peut réaliser des estimations des coefficients λ et μ directement à partir du signal analysé, il est alors possible d'évaluer pour tout signal, une majoration des effets dus à l'échantillonnage. Ceci fera l'objet d'une prochaine publication.

Références

[1] N.E. Huang, Z. Shen, S.R. Long, M.L. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Tung et H.H. Liu, "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary time series analysis," *Proc. Roy. Soc. London A*, Vol. 454, pp. 903–995, 1998.

[2] P. Flandrin et P. Gonçalvès, "Sur la décomposition modale empirique," GRETSI-03, pp. 149–152, Paris.

[3] P. Flandrin, G. Rilling et P. Gonçalvès, "Empirical Mode Decomposition as a filterbank," *IEEE Sig. Proc. Lett.*, Vol. 11, pp. 112–114, 2004.

[4] G. Rilling, P. Flandrin et P. Gonçalvès, "On Empirical Mode Decomposition and its Algorithms", IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing NSIP-03, Grado (I), 2003.

[5] `http ://www.perso.ens-lyon.fr/patrick.flandrin/emd.html`.

Vers une nouvelle décomposition de matrice

Jean-Jacques Fuchs
IRISA/Université de Rennes I
Campus de Beaulieu - 35042 Rennes Cedex
fuchs@irisa.fr

Résumé – Pour approximer une matrice par une matrice de rang plus faible on utilise en général la décomposition en valeurs et vecteurs singuliers, ce qui est optimal pour toutes les normes unitairement invariantes comme la norme de Frobenius par exemple. Cela se justifie d'un point de vue statistique, si on suppose que chaque composante a été perturbée par un bruit additif gaussien et que l'on recherche une estimée de la matrice exacte sous-jacente au maximum de vraisemblance. Sous d'autres hypothèses il faudrait minimiser une autre norme. Nous considérons le cas où le bruit additif est supposé suivre une loi de Laplace.

Abstract – Low rank approximations of a matrix are mostly performed using the singular value decomposition which is optimal for all unitarily invariant matrix norms, such as the Frobenius norm. From a statistical point of view this approach is justified when the components are perturbed by independent and identically distributed zero mean gaussian noise. If this assumption is not valid other norms and thus approximations should be considered. The norm we consider below corresponds to noise samples that would be assumed to follow the Laplace or double-exponential distribution.

1 Introduction

En traitement du signal ou de l'image on a souvent besoin d'approximer une matrice par une matrice de rang inférieur. Si les composantes de la matrice réelle A de dimension (m,n) et de rang r sont perturbées par des variables aléatoires elle va devenir de rang plein et estimer la matrice initiale est un problème difficile qui peut nécessiter une estimation préalable du rang. Si les perturbations aléatoires sont indépendantes, centrées et gaussiennes il faut utiliser la norme de Frobénius que nous allons noter $|||A|||_2$:

$$|||A|||_2 = (\sum_{i,j} a_{i,j}^2)^{1/2} = \mathrm{trace}(A^T A)^{1/2} \qquad (1)$$

et si le rang r est connu, il faut résoudre

$$\min_{\Delta} \; |||\Delta|||_2 \quad \text{s.t.} \quad \mathrm{rang}(A-\Delta) = r \qquad (2)$$

La solution est bien connue et se déduit de la décomposition en valeurs et vecteurs singuliers (SVD) de la matrice A. Si on suppose que les perturbations suivent une loi de Laplace (exponentielle double) $f_X(x) = \frac{\beta}{2}\, e^{-\beta|x|}$ la norme à utiliser à la place de (1) est

$$|||A|||_1 = \sum_{i,j} |a_{i,j}| \qquad (3)$$

En effet comme les perturbations sont indépendantes, la loi jointe est égale au produit des lois et maximiser la log-vraisemblance conduit dans le cas gaussien à minimiser la norme de Frobenius et dans le cas Laplace à minimiser cette norme 'ℓ_1' (3).

Si on considère les matrices de dimension (m,n) comme des vecteurs dans un espace vectoriel de dimension $n \times m$, la norme vectorielle euclidienne ou ℓ_2, $\|x\|_2 = (\sum_i x_i^2)^{1/2}$ et la norme vectorielle ℓ_1 $\|x\|_1 = \sum_i |x_i|$ induisent les normes matricielles (1) et (3) respectivement.

A notre connaissance le problème de l'approximation

$$\min_{\Delta} \; |||\Delta|||_1 \quad \text{sous} \quad \mathrm{rang}(A-\Delta) = r \qquad (4)$$

associée à la norme (3) n'a pas été considéré à ce jour et nous en présentons quelques aspects.

Dans le paragraphe qui suit nous rappelons quelques résultats concernant la décomposition en valeurs et vecteurs singuliers (SVD), puis nous présentons les liens entre la norme de Frobenius, la SVD et certains problèmes d'approximation de matrices. Nous introduisons ensuite la nouvelle décomposition et un algorithme (coûteux) pour l'obtenir. Nous terminons par un exemple avant de conclure.

2 La décomposition en valeurs et vecteurs singuliers

La décomposition en valeurs et vecteurs singuliers d'une matrice réelle A de dimension (m,n) et de rang r est de la forme $A = U\Sigma V^T$ avec U et V des matrices unitaires

$$U^T A V = \begin{bmatrix} \bar{\Sigma} & 0 \\ 0 & 0 \end{bmatrix} \qquad \text{où} \quad \bar{\Sigma} = \mathrm{diag}(\sigma_1, \ \ldots \ , \sigma_r)$$

$$\text{avec} \quad \sigma_1 \geq \sigma_2 \ \ldots \ \geq \sigma_r > 0$$

On peut décomposer U en $[\ U_1\ U_2\ ]$ avec U_1 ayant r colonnes et décomposer V de façon analogue pour avoir

$$A = U_1 \bar{\Sigma} V_1^T = \sum_1^r \sigma_k u_k v_k^T \qquad (5)$$

qui est parfois appelé la factorisation en valeurs et vecteurs singuliers Les colonnes de U_1 forment une base orthogonale de $\Re(A)$, l'image de A et les colonnes de V_1 une base

orthogonale de $\Re(A^T)$. Les colonnes de U_2 et V_2 sont alors des bases arbitraires des espaces complémentaires.

La factorisation (5) permet d'écrire simplement la pseudo-inverse A^+ de A: $A^+ = V_1\bar{\Sigma}^{-1}U_1^T$, une inverse généralisée spécifique de A. On rappelle qu'une inverse généralisée de A est une matrice A^- vérifiant $AA^-A = A$.

3 Norme de Frobenius et SVD

Nous commençons par le problème suivant. Soit A une matrice réelle de dimension (m,n) et de rang r, on cherche une perturbation Δ de norme minimale qui fait chuter le rang de A de un:

$$\min_{\Delta} \; |||\Delta|||_2 \quad \text{s.t.} \quad \text{rang}(A-\Delta) = \text{rang}(A) - 1 \qquad (6)$$

Nous allons supposer que Δ est de rang un. Ceci est vrai pour la norme de Frobenius mais ne l'est pas nécessairement pour d'autres normes.

Bien que la solution soit bien connue et découle facilement de résultats plus généraux, nous allons l'établir car la démonstration utilisée est similaire à celle que nous allons utiliser plus loin pour trouver la solution dans le cas de la norme 'ℓ_1' (3). Nous allons utiliser le lemme suivant et son corollaire.

Lemme 1: Pour une matrice A donnée, toute matrice Δ de rang un qui est telle que rang(A-Δ) = rang(A) -1, est de la forme

$$\Delta = \frac{Axy^TA}{y^TAx} \qquad \text{avec} \qquad y^TAx \neq 0 \qquad \square$$

Corollaire 1: La matrice Δ de rang un peut aussi s'écrire

$$\Delta = \frac{ab^T}{b^TA^-a} \qquad \text{avec} \qquad b^TA^-a \neq 0 \qquad .$$

$a \in \Re(A)$, $b \in \Re(A^T)$ et A^- une inverse généralisée de A.$\square$

Comme on suppose que l'optimum de (6) est atteint en une matrice de rang un, en utilisant le corollaire et en prenant $A^- = A^+$, (6) se réécrit

$$\min_{a,b} \; |||\frac{ab^T}{b^TA^+a}|||_2 \;\; \text{où } b^TA^+a \neq 0,\; a \in \Re(A),\; b \in \Re(A^T)$$

comme $|||ab^T|||_2 = \|a\|_2\,\|b\|_2$, son optimum se déduit de

$$\max_{a,b} |b^TA^+a| \text{ avec } \|a\|_2 = \|b\|_2 = 1,\; a \in \Re(A),\; b \in \Re(A^T)$$

Nous montrons maintenant que l'optimum de ce dernier problème est atteint en $a = u_r$ and $b = v_r$ ce qui conduit à $\Delta = \sigma_r u_r v_r^T$:

$$\begin{aligned}
|b^TA^+a| &= |b^TV_1\bar{\Sigma}^{-1}U_1^Ta| \\
&= |y^T\bar{\Sigma}^{-1}z| \quad \text{avec} \quad y = V_1^Tb, \quad z = U_1^Ta \\
&= |\sigma_1^{-1}y_1z_1 + \;..\; + \sigma_r^{-1}y_rz_r| \\
&\leq \sigma_r^{-1}(|y_1z_1| + + \;..\; + |y_rz_r|) \\
&\leq \sigma_r^{-1}\|y\|_2\,\|z\|_2 \quad \text{(inégalité de Schwarz)} \\
&= \sigma_r^{-1}
\end{aligned}$$

le maximum est donc inférieur à σ_r^{-1} mais comme par ailleurs $|v_r^TA^+u_r| = \sigma_r^{-1}$ cette valeur est atteinte pour $a = u_r$ et $b = v_r$.

Nous n'avons fait que vérifier que Δ avait la forme attendue, mais en utilisant le lagrangien il est facile d'établir ce résultat et en tout cas de montrer que l'optimum s'obtient à partir de la SVD de A. Grâce au lemme 1 le problème se réécrit

$$\max_{x,y} \; |y^TAx| \quad \text{with} \quad \|Ax\|_2 \leq 1, \quad \|A^Ty\|_2 \leq 1$$

où nous avons remplacé les contraintes égalités par des inégalités car cela ne modifie pas le problème. Le lagrangien s'écrit :

$\ell(x,y) = -y^TAx + \frac{\lambda}{2}(\|Ax\|_2 - 1) + \frac{\mu}{2}(\|A^tY\|_2 - 1)$

et ses points stationnaires, qui vérifient les conditions nécessaires du premier ordre, doivent satisfaire:

$-A^Ty + \lambda A^TAx = 0, \quad \lambda \geq 0, \quad \|Ax\|_2 = 1$

$-Ax + \mu AA^Ty = 0, \quad \mu \geq 0, \quad \|A^Ty\|_2 = 1$

On observe qu'en un point stationnaire $\lambda = \mu = |y^TAx|$ et en posant $u = Ax$, $v = A^Ty$ et $\sigma = 1/\lambda = 1/\mu$ ces conditions deviennent

$$\begin{aligned}
A^Tu = \sigma v, \qquad \|u\|_2 = 1, \qquad \sigma > 0 \qquad (7) \\
Av = \sigma u, \qquad \|v\|_2 = 1
\end{aligned}$$

On reconnait le couple d'équations que doivent satisfaire les valeurs et vecteurs singuliers $\{\ \sigma, u, v\ \}$ aussi connues sous le nom d'équations de Schmidt. Comme $\sigma = 1/|y^TAx|$ doit être minimisé on retient le triplet $\{\ \sigma_r, u_r, v_r\ \}$ et il est facile de vérifier que les conditions suffisantes du second ordre sont satisfaites par u_r et v_r si la valeur singulière minimale σ_r est simple.

Nous avons donc montré que l'optimum de

$$\min_{\text{rang}(\Delta)=1} \; |||\Delta|||_2 \;\; \text{sous} \quad \text{rang}(A-\Delta) = \text{rang}(A) - 1$$

est atteint en $\Delta = \sigma_r u_r v_r^T$ et a pour valeur σ_r la valeur singulière minimale de A. Si on remplace maintenant A par la matrice $A - \sigma_r u_r v_r^T$ et que l'on se repose le même problème de minimisation, on trouve pour solution $\Delta = \sigma_{r-1}u_{r-1}v_{r-1}^T$. La décomposition en valeurs et vecteurs singuliers peut ainsi être obtenue de proche en proche en résolvant une suite de problèmes d'optimisation.

C'est ce que nous nous proposons de faire en utilisant la norme 'ℓ_1' (3) à la place de la norme de Frobenius (1).

4 La décomposition 'ℓ_1'

4.1 La brique de base

Soit une matrice A de dimension (m,n) et de rang r, on cherche la perturbation Δ de norme 'ℓ_1' minimale qui fait chuter son rang de un. On cherche la solution de

$$\min_{\Delta} \; |||\Delta|||_1 \quad \text{sous} \quad \text{rang}(A-\Delta) = \text{rang}(A) - 1 \qquad (8)$$

On suppose que l'optimum est atteint en une matrice Δ de rang un. Nous n'avons pas réussi à établir ce résultat sauf si A est de rang plein. En utilisant le corollaire 1 et le fait que $|||ab^T|||_1 = \|a\|_1\,\|b\|_1$, le problème (8) devient

$$\max_{a,b} \; |b^TA^-a|, \;\; \|a\|_1 \leq 1, \;\; \|b\|_1 \leq 1, \;\; a \in \Re(A), \;\; b \in \Re(A^T)$$

où, sans changer l'optimum on a remplacé les contraintes égalités par des inégalités. On réécrit maintenant les conditions $a \in \Re(A)$ et $b \in \Re(A^T)$ en utilisant les matrices U_2 et V_2 introduites plus haut (5), leurs colonnes sont des bases orthogonales arbitraires des espaces complémentaires de $\Re(A)$ et $\Re(A^T)$ respectivement. $a \in \Re(A)$ devient $U_2^T a = 0$ et $b \in \Re(A^T)$ devient $V_2^T a = 0$ et le problème (8) est alors équivalent à

$$\max_{a,b} \ |b^T A^- a| \ , \ \|a\|_1 \leq 1, \ \|b\|_1 \leq 1, \ U_2^T a = 0, \ V_2^T b = 0 \tag{9}$$

dont le lagrangien est

$$\ell(x, y, \lambda_1, \lambda_2, \Lambda_1, \Lambda_2) = -|b^T A^- a| + \lambda_1(\|a\|_1 - 1) + \lambda_2(\|b\|_1 - 1) + \Lambda_1^T U_2^T a + \Lambda_2^T V_2^T b$$

Pour écrire les conditions à satisfaire par ses points stationnaires on introduit la sous-différentielle, une extension de la notion de gradient en des points où il n'est pas défini

$$\begin{aligned} \partial \ \|z\|_1 \ &= \{v | v^T z = \|z\|_1, \ \|v\|_\infty \leq 1\} \\ &= \{v | v_i = \pm 1 \ \text{ si } \ z_i \neq 0 \ \text{ and } \ |v_i| \leq 1 \ \text{ sinon}\} \end{aligned}$$

Avec ces notations les points stationnaires satisfont:

$$\begin{aligned} &-\text{sign}(b^T A^- a) A^{-T} b + \lambda_1 \partial \ \|a\|_1 + U_2 \Lambda_1 = 0, \\ &-\text{sign}(b^T A^- a) A^- a + \lambda_2 \partial \ \|b\|_1 + V_2 \Lambda_2 = 0, \\ &\|a\|_1 = 1, \ \|b\|_1 = 1, \lambda_1 \geq 0, \ \lambda_2 \geq 0, \ U_2^T a = 0, \ V_2^T b = 0 \end{aligned}$$

Comme $a^T \partial(a) = \|a\|_1$ et $\text{sign}(b^T A^- a)(b^T A^- a) = |b^T A^- a|$, en pré-multipliant la première relation par a^T, et la seconde par b^T on a $\lambda_1 = \lambda_2 = |b^T A^- a|$. Les deux premières relations deviennent alors

$$\begin{aligned} &-A^{-T} b + \lambda_1 \partial \ \|a\|_1 + U_2 \Lambda_1 = 0, \ \lambda_1 > 0 \\ &-A^- a + \lambda_1 \partial \ \|b\|_1 + V_2 \Lambda_2 = 0, \end{aligned}$$

en posant $\sigma = 1/\lambda_1 = 1/\lambda_2$ et en observant que $AA^- a = a$ puisque $a \in \Re(A)$ et $AV_2 = 0$ et les propriétés analogues pour b et V_2, on obtient les conditions suivantes en pré-multipliant ces deux relations par A^T et A respectivement

$$\begin{aligned} A^T \partial \|a\|_1 &= \sigma b, \quad \|a\|_1 = 1, \quad \sigma \geq 0 \\ A \ \ \partial \ \|b\|_1 &= \sigma a, \quad \|b\|_1 = 1 \end{aligned} \tag{10}$$

Le couple a, b vérifiant ces conditions et associé au σ le plus petit est génériquement unique et la matrice de rang un $\sigma a b^T$ est alors une candidate à l'optimum de (8). Nous n'avons pas trouver les conditions suffisantes que ce point doit satisfaire.

Ces relations (10) sont le pendant des équations de Schmidt (7) qui caractérisent les triplets singuliers de la matrice A. Alors qu'il est facile de combiner les relations (7) pour faire apparaître les équations caractérisant les valeurs et vecteurs propres des matrices AA^T and $A^T A$, rien de tel ne semble faisable ici et nous avons seulement été capable de résoudre (10) à l'aide d'un algorithme de recherche exhaustive. Mais avant d'en arriver à un tel algorithme il nous faut établir le résultat suivant:

Lemme 2: La matrice optimal de rang un $\Delta = \frac{ab^T}{b^T A^- a} = \sigma a b^T$ déduite de la solution du problème (9) est le produit d'un vecteur colonne par un vecteur ligne ayant chacun au moins $r-1$ composantes nulles. □

Démonstration: Nous montrons que l'optimum (a_o, b_o) de (9) est tel que a_o a au plus m-$(r-1)$ composantes non nulles et b_o au plus n-$(r-1)$ composantes non nulles. Pour b_o fixé, le vecteur a est solution de

$$\max b_o^T A^- a \text{ sous } \|a\|_1 = 1, \ U_2^T a = 0$$

dans lequel on reconnaît un programme linéaire en a pour lequel l'optimum estgénériquement atteint en une solution de base admissible. Dans notre cas cela signifie que l'optimum a au plus $m - r + 1$ composantes non nulles puisque l'on dénombre $m - r + 1$ contraintes linéaires égalités. Un raisonnement analogue permet d'établir le résultat annoncé pour b. □

4.2 L'algorithme de recherche exhaustive

Nous proposons un algorithme qui donne toutes les solutions (σ, a, b) qui vérifient (10):

$$\begin{aligned} A^T \partial \|a\|_1 &= \sigma b, \qquad \|a\|_1 = 1 \qquad \sigma \geq 0 \\ A \partial \ \|b\|_1 &= \sigma a, \qquad \|b\|_1 = 1, \end{aligned}$$

pour A une matrice (m,n) de rang r. A (σ, a, b) on associe la matrice de rang un $\Delta = \sigma a b^T$.

On considère toutes les matrices carrées d'ordre $(r-1)$ que l'on peut extraire de A. En échangeant les lignes d'une part et les colonnes d'autre part, on amène chacune de ces sous-matrices dans le coin supérieure gauche et on la note $A_{1,1}$. On résout alors (10) pour chacune de ces permutations. Comme le lemme précédent nous a appris que a et b ont au moins r-1 composantes nulles, on suppose que ce sont celles de a_1 et b_1, les sous-vecteurs associés à $A_{1,1}$. Les composantes restantes dans a_2 et b_2 sont alors génériquement non nulles. On décompose de la même façon $\partial\|a\|_1$ en deux sous-vecteurs ∂a_1 et $\text{sign}(a_2)$ et de même pour $\partial\|b\|_1$. Comme $a_1 = 0$, les composantes de $\partial\|a\|_1$ sont dans l'intervalle $(-1, 1)$. Comme les composantes de a_2 sont supposées différentes de zéro, les composantes correspondantes dans $\partial\|a\|_1$ sont égales à ± 1 d'où la notation $\text{sign}(a_2)$ pour désigner la fonction signe usuelle;

Pour chaque permutation, on décompose donc la matrice A en 4 blocs $A_{i,j}$ et la seconde équation de (10) $A\partial\|b\|_1 = \sigma a$ se transforme en deux sous-relations qui s'écrivent, avec les notations que nous venons d'introduire

$$\begin{aligned} A_{1,1} \partial b_1 + A_{1,2} \ \text{sign}(b_2) &= \sigma a_1 = 0 \\ A_{2,1} \partial b_1 + A_{2,2} \ \text{sign}(b_2) &= \sigma a_2 \end{aligned}$$

Si $A_{1,1}$ n'est pas inversible, on passe à la permutation suivante, sinon la première de ces deux relations permet de calculer ∂b_1 pour chacune des $2^{(n-r)}$ valeurs possibles de $\text{sign}(b_2) \in [\pm 1 \ \ \pm 1 \ \ .. \pm 1 \ ]^T$. Pour chaque $\text{sign}(b_2)$ on vérifie si $\|\partial b_1\|_\infty \leq 1$ et si c'est la cas on a obtenu un couple cohérent ∂b_1, $\text{sign}(b_2)$. Pour chacun de ces couples on utilise la seconde relation pour en déduire $\sigma > 0$ et a_2 avec $\|a_2\|_1 = 1$. On passe alors à la première équation de (10) $A^T \partial \|a\|_1 = \sigma b$ qui donnent également deux relations supplémentaires

$$\begin{aligned} \partial a_1^T A_{1,1} + \ \text{sign}(a_2)^T A_{2,1} &= \sigma b_1^T = 0 \\ \partial a_1^T A_{1,2} \ + \ \text{sign}(a_2)^T A_{2,2} &= \sigma b_2^T \end{aligned}$$

Comme $\text{sign}(a_2)$ est maintenant connu, on vérifie si $\|\partial a_1\|_\infty \leq 1$ dans la première relation. Si tel est le cas, on déduit $\sigma' > 0$, b_2 de la seconde. Il reste alors à voir si $\sigma' = \sigma$ et si les signes des composantes de b_2 sont les

mêmes que ceux de $\text{sign}(b_2) \in [\pm 1 \quad \pm 1 \quad .. \pm 1\ ]^T$ retenu plus haut dans la première relation issue de la seconde équation de (10). Si toutes ces conditions sont satisfaites, on a trouvé un point stationnaire du Lagrangien.

Il y a en général plusieurs points stationnaires mais leur nombre n'est ni fixe ni prévisible contrairement à ce qui est le cas pour (7). On ne retient que la solution associée au σ le plus petit. Elle semble génériquement unique.

On peut développer un autre algorithme de recherche exhaustive qui donne, bien entendu, le même résultat. Dans le développement précédent, on cherche pour chaque permutation, la perturbation Δ dont seul le bloc $\Delta_{2,2}$ est non nul. On recherche donc la sous-matrice $\Delta_{2,2}$ qui fait chuter le rang de A de un quand elle est soustraite de $A_{2,2}$.

Si $A_{1,1}$ est inversible une telle matrice Δ existe toujours. La sous-matrice $\Delta_{2,2}$ associée vaut $A_{2,2} - A_{2,1}A_{1,1}^{-1}A_{1,2}$ et peut être vue comme le complément de Schur de $A_{1,1}$ dans A. On montre facilement qu e $\Delta_{2,2}$ est de rang un qui peut donc se mettre sous la forme $\sigma a_2 b_2^T$ avec $\|a_2\|_1 = 1$ et $\|b_2\|_1 = 1$. Mais rien ne garantit que le triplet σ, a_2, b_2 vérifie les conditions imposées par (10). Les rares cas où ces conditions sont satisfaites, mènent aux mêmes matrices que l'algorithme précédent. Cela semble notamment toujours être le cas pour le $\Delta_{2,2}$ de norme 'l_1' minimale

5 Un exemple

Pour illustrer la décomposition proposée nous l'appliquons à une matrice carrée d'ordre 4 inversible. On peut remarquer que le lemme 2 appliqué plus généralement à une matrice inversible d'ordre n nous dit que la matrice $\Delta = \sigma a b^T$ de rang un et de norme 'l_1' minimale qui fait chuter son rang de un est une matrice ayant juste une composante non nulle. La solution à ce problème spécifique s'obtient facilement directement. Après avoir soustrait cette matrice de A, la nouvelle matrice est d'ordre n et de rang $n-1$ et le lemme 2 nous apprend que la matrice $\Delta = \sigma a b^T$ de norme 'l_1' minimale qui fait passer son rang à $n-2$ est le produit de deux vecteurs σa et b ayant au plus deux composantes non nulles chacun. En continuant ainsi on en déduit que la k-ième matrice dans la décomposition est le produit de deux vecteurs ayant chacun au plus k composantes non nulles.

Pour la matrice de Vandermonde d'ordre 4

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ .2 & .4 & .6 & .8 \\ .04 & .16 & .36 & .64 \\ .008 & .064 & .216 & .512 \end{bmatrix}$$

la décomposition que l'on obtient et de la forme $A = GDH^T$ où dans les colonnes de G de norme ℓ_1 unité se trouvent les vecteurs a_i, dans les colonnes de H de norme ℓ_1 unité se trouvent les vecteurs b_i et les σ_i sont sur la diagonale de D, $A = \sum_i \sigma_i a_i b_i^T$

$$G = \begin{bmatrix} 0 & 0 & -0.5938 & -0.4615 \\ 0 & 0.5682 & 0 & -0.2592 \\ -1 & 0 & 0.1781 & -0.1661 \\ 0 & -0.4318 & 0.2280 & -0.1131 \end{bmatrix}$$

$$D = \begin{bmatrix} 0.0100 & 0 & 0 & 0 \\ 0 & 0.2247 & 0 & 0 \\ 0 & 0 & 2.0157 & 0 \\ 0 & 0 & 0 & 7.9121 \end{bmatrix}$$

$$H = \begin{bmatrix} 0 & -0.7000 & -0.4051 & -0.1411 \\ 1 & 0 & -0.2405 & -0.1950 \\ 0 & 0.3000 & 0 & -0.2739 \\ 0 & 0 & 0.3544 & -0.3900 \end{bmatrix}$$

6 Conclusions

Les résultats que nous avons proposés, peuvent aussi être utilisés pour résoudre ce que l'on pourrait appeler le problème des moindres déviations totales par analogie avec le problème des moindres carrés totaux. Si le problème des moindres déviations consistent à résoudre

$$\min_x \ \|Ax - b\|_1$$

avec A une matrice (m,n) de rang n et x et b des vecteurs de dimensions adéquates, qui s'écrit aussi $\min_{r,x} \ \|r\|_1$ sous $r = Ax - b$, celui des des moindres déviations totales correspond à résoudre

$$\min_{x,\ r,\ R} \ \|\!| [R\ r] |\!\|_1 \quad \text{sous} \quad (A+R)x = b + r$$

et a déjà été considère dans [9].

Références

[1] R.A. Horn and C.R. Johnson *Matrix Analysis*. Cambridge University Press., 1985.

[2] R. Fletcher, *Practical Methods of Optimization*. John Wiley and sons., 1987.

[3] R.B. Bapat, *Linear Algebra and Linear Models*. Springer, Universitext, 1999, p. 114.

[4] G.H. Golub and C.F. Van Loan. *Matrix Computations*. John Hopkins University Press, 1983.

[5] P. Dewilde and E.F. Deprettere. "Singular Value Decomposition : an introduction" *SVD and Signal Processing* pp. 3-41, North-Holland, 1988.

[6] D. G. Luenberger. *Introduction to Linear and Nonlinear Programming*. Addison Wesley, 1973.

[7] G.W. Stewart and J.G. Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[8] D.P. Bertsekas. *Convex Analysis and Optimization*. Athena Scientific, 2003.

[9] M.R. Osborne and G.A. Watson. An analysis of the total approximation problem in separable norms and an algorithm for the total ℓ_1 problem. *SIAM J. Sci. Stat. Comput.*, vol. 6, 2, 410–424, april 1985.

Estimation de signaux par noyaux d'ondelettes

Vincent Guigue Alain Rakotomamonjy Stéphane Canu

Lab. Perception, Systèmes, Information - CNRS - FRE 2645
Avenue de l'Université, 76801 St Étienne du Rouvray
Tél : 02 32 95 97 03 Fax : 02 32 95 97 08
Vincent.Guigue@insa-rouen.fr

Résumé – Cet article présente une méthode de régression pour les signaux non uniformément échantillonnés. Notre méthode propose une approche de type apprentissage supervisé, qui combine une approximation $\mathcal{L}_2$ par noyaux d'ondelettes et une régularisation $\mathcal{L}_1$ multi-échelles. L'utilisation de l'algorithme *Least Angle Regression* pour la résolution du problème permet d'introduire de nouvelles manières de régler le compromis biais-variance.

Abstract – This paper addresses the problem of regression in the case of non-uniform sampled signals. Our method is based on supervised learning theory, we propose to use $\mathcal{L}_2$ estimation with wavelet kernel combined with $\mathcal{L}_1$ multiscale regularization. The use of Least Angle Regression as solver enable us to propose new solutions to set the regularization parameter.

1 Introduction

Le problème de l'approximation de signaux à partir de données non uniformément échantillonnées apparaît dans plusieurs contextes tels que les systèmes présentant une fréquence d'échantillonnage fluctuante voire aléatoire ou la reconstruction de signaux qui comportent des données manquantes. En traitement d'image, il est également fréquent d'avoir à reconstruire des images dont les échantillons ne sont pas uniformément distribués *i.e* données géophysiques, tomographie, etc ...

Dans cette étude, nous proposons une méthode d'approximation de signaux bruités en se plaçant dans le contexte des espaces de Hilbert à noyaux reproduisants et en utilisant comme *a priori* sur la régularité du signal la parcimonie de l'estimateur. Ici, les espaces d'approximation considérés sont des sous-espaces de $\mathcal{L}_2(\mathbb{R})$ générés par un ensemble fini d'ondelettes. Les approches basées sur le *Matching Pursuit* [Mal97] mènent à des résultats sous-optimaux, tandis que le *Basis Pursuit* [CDS98] fait appel à des méthodes *backward* très coûteuses en temps de calcul. Le *wavelet shrinkage* [DJ94] était initialement limité aux signaux uniformément échantillonnés. L'extension aux signaux quelconques nécessite en réalité de ré-échantillonner les signaux [KS00]. Les méthodes existantes utilisant les noyaux d'ondelettes [AAP04] combine une résolution de type EM (coûteuse) avec l'utilisation de plusieurs paramètres de régularisation. Notre but est double : d'une part, nous souhaitons diminuer la complexité temporelle de la résolution de ce problème, d'autre part nous proposons des solutions pour régler automatiquement les paramètres de la méthode d'approximation.

Nous avons utilisé la formulation du LASSO (*Least Absolute Shrinkage and Selection Operator*) [Tib96], basée sur une régularisation $\mathcal{L}_1$, et l'algorithme du LARS (*Stepwise Least Angle Regression*) [EHJT04] pour la résolution. La régularisation $\mathcal{L}_1$ est un gage de parcimonie tandis que le calcul du chemin complet de régularisation, via le LARS, permet de définir de nouveaux critères pour trouver le compromis biais-variance optimal. Cette démarche permet d'améliorer les performances et le temps de calcul pour le LASSO tout en proposant des critères auto adaptatifs pour le réglage du compromis biais-variance, afin de rendre la méthode complètement non paramétrique.

2 Méthode

Nous nous plaçons dans le contexte d'une estimation de la fonction de régression classique i.e, nous avons un ensemble de données $\{x_i, y_i\}_{i=1...n}$ avec : $y_i = f(x_i)+b_i$ où les b_i sont des variables aléatoire et $x_i \in \mathbb{R}$. Notre objectif est d'obtenir une estimation de f à partir des échantillons bruités y_i. Les x_i ne sont pas équidistants. Nous cherchons la fonction f de l'espace des hypothèses $\mathcal{H}$ qui minimise le risque empirique régularisé :

$$R_{reg}[f] = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda\Omega(f) \qquad (1)$$

où $\Omega(f)$ mesure la régularité de la solution et où λ est le compromis biais-variance. L'approximation régularisée d'une fonction non uniformément échantillonnée rentre dans le cadre du théorème du représentant de Kimerdolf et Wahba [KW71] et la forme de la solution est donc :

$$f(x) = \sum_{i=1}^{n}\beta_i K(x_i, x) \qquad (2)$$

où K est le noyau reproduisant de l'espace de Hilbert $\mathcal{H}$ et où les β_i sont les coefficients de régression.

L'utilisation de noyaux d'ondelettes pour la régression, combinée à une régularisation $\Omega(f)$ a déjà fait l'objet d'études [AAP04, RC05]. La méthode présentée dans cet

article apporte plus de rapidité, ainsi que la possibilité de régler automatiquement tous les paramètres.

2.1 Noyaux d'ondelettes

L'application du théorème du représentant suppose que $\mathcal{H}$ soit un espace de Hilbert à noyaux reproduisants. Les conditions suffisantes d'auto-reproduction [RC05] sont données ci-dessous :

- Tout ensemble fini $\{\psi_n\}_{i=1\ldots n}$ de vecteurs d'un espace de Hilbert $\mathcal{H}$ est une structure oblique (ou frame) du sous-espace qu'elle génère.
- si par ailleurs, l'ensemble des ψ_n est borné et de norme finie dans $\mathcal{H}$ alors, l'espace engendré est à noyau reproduisant.

Dans ce cas, le noyau de cette espace s'écrit :

$$k(x,y) = \sum_{i=1}^{n} \bar{\psi}_i(x)\psi_i(y)$$

où $\bar{\psi}$ est la frame duale de ψ. Ainsi, on peut en conclure qu'une famille finie (orthogonale ou non) d'ondelettes génère un espace à noyau reproduisant. Cette espace peut donc être utilisé dans le cadre du théorème de Kimerdolf-Wahba pour résoudre un problème d'approximation.

Si l'espace d'hypothèses $\mathcal{H}$ est décrit par un ensemble d'ondelettes orthogonales ψ, alors cet espace se décompose en N sous espaces : $\mathcal{H} = \mathcal{H}_1 \oplus \ldots \oplus \mathcal{H}_N$ où chaque espace de Hilbert à noyau reproduisant $\mathcal{H}_i$ est l'espace engendré par une famille d'ondelettes de même échelle i.

Estimer la fonction sur $\mathcal{H}$ nécessite l'utilisation des noyaux des $\mathcal{H}_i$, pour modéliser chacune des échelles. En reprenant la non linéarisation de Vincent et Bengio [VB02], nous considérons chaque $K_i(x,.)$ comme une source d'information. Nous construisons donc un noyau multi-échelles comme la concaténation de différents noyaux élémentaires : $K = \left[\ K_1 \ldots K_i \ldots K_N\ \right]$. Chaque source d'information $K_i(x_j,\cdot)$ est caractérisée par un point x_j (référence temporelle) et un facteur d'échelle i. La fonction de régression s'écrit alors :

$$f(x) = \sum_{i=1}^{N}\sum_{j=1}^{n} \beta_{ij} K_i(x_j, x) \qquad (3)$$

2.2 Régularisation multi-échelles et LARS

La régularisation est souvent associée à la parcimonie de la solution. La meilleure manière de l'améliorer est de minimiser $||\beta||_{\ell_0}$, c'est à dire de prendre $\Omega(f) = ||\beta||_{\ell_0}$. Cependant, ce problème est NP-difficile et deux solutions classiques consistent à utiliser soit la norme ℓ_1 [Tib96], soit la norme ℓ_2 en approximant parcimonieusement la solution [VB02].

Grandvalet a montré que l'*Adaptive Ridge Regression* (ARR, $\Omega(f) = \sum_{ij} \lambda_i \beta_{ij}^2$) est équivalente au LASSO ($\Omega(f) = \sum_{ij} |\beta_{ij}|$) [Gra98]. Dans le cas présent, l'ARR utilise une régularisation multi-échelles, λ_i étant fixé pour chaque K_i. La formulation du LASSO permet donc d'obtenir simplement une régularisation multi-échelles en s'affranchissant d'un grand nombre d'hyper-paramètres.

La plupart des problèmes d'apprentissage sont régularisés en ℓ_2 et il existe beaucoup d'outils pour les traiter. A l'inverse, la formulation ℓ_1 demandait jusqu'à maintenant des résolutions coûteuses, basées sur la programmation linéaire [CDS98] ou sur des versions améliorées de EM [Gra98]. L'algorithme *Stepwise Least Angle Regression* (LARS) [EHJT04] offre de nouvelles perspectives pour ce genre de problème. La rapidité du LARS vient de la combinaison entre parcimonie et méthode pas à pas, en partant d'un ensemble vide. Les premières itérations sont peu coûteuses du fait de la petite taille de l'ensemble actif des sources d'information. Le coût augmente au fil des itérations, mais la régularisation ℓ_1, gage de parcimonie, limite le nombre d'itérations nécessaires. A l'inverse, les méthodes *backward* [Gra98, CDS98] nécessitent la résolution d'un système linéaire de grande dimension pour réduire les coefficients inutiles à zéro. Étant donné qu'un seul point est ajouté (ou retiré) de l'ensemble actif à chaque itération, il est possible de mettre à jour la matrice inversée au lieu de la recalculer entièrement à la manière de la formulation simple-SVM [LCV$^+$04].

Finalement, le problème du LARS s'écrit :

$$\min_{\beta} \sum_{i=1}^{n} \|y_i - \sum_{j=1}^{N}\sum_{k=1}^{n} \beta_{jk} K_j(x_k, x_i)\|^2 \quad \sum_{j,k} |\beta_{jk}| \leq t \qquad (4)$$

Cette formulation est équivalente à (1) avec $\Omega(f) = ||\beta||_{\ell_1}$.

2.3 Modification du LARS

Afin d'améliorer la parcimonie et les résultats en classification de la méthode, nous avons introduit une modification dans l'algorithme du LARS. Lors de la dernière itération de la méthode, le pas est calculé suivant la méthode des moindres carrés classiques. Cela revient en réalité à utiliser le LARS comme une méthode de séléction de sources d'information puis à faire une projection sur cette famille de fonctions.

2.4 Compromis biais-variance

Le compromis biais-variance permet de stopper la phase d'apprentissage avant d'intégrer tous les β dans la construction de la solution. L'intérêt est double : d'une part, il s'agit d'éviter les phénomènes de sur-apprentissage qui surviennent lorsque la solution est trop complexe, d'autre part, cela nous permet de réduire le temps de calcul nécessaire pour obtenir la solution optimale. Le réglage du paramètre de compromis t (4) constitue donc un enjeu essentiel.

Pour faire face à ce problème nous utilisons une propriété du LARS : le calcul le chemin complet de régularisation. Chacune des itérations du LARS aboutit à une solution optimale au sens d'un paramètre t fixé. Il est donc possible de régler le compromis biais-variance dynamiquement [BTJ04].

Le critère le plus classique pour régulariser le LARS consiste à fixer une borne sur la somme des $|\beta_i|$. Cepen-

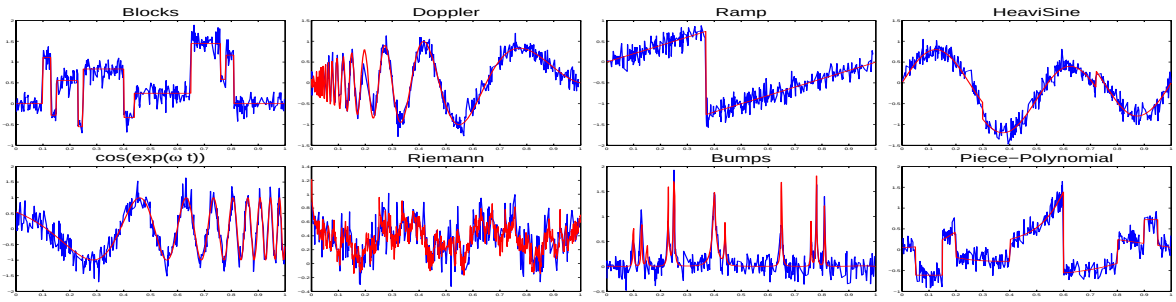


FIG. 1 – Signaux à estimer et données d'apprentissage bruitées.

dant, ce critère est particulièrement abstrait et difficile à régler.

2.4.1 Autres définitions de seuil

Nous proposons d'utiliser d'autres définitions de seuil pour stopper la phase d'apprentissage :

- le ν-LARS, où ν est un seuil sur le pourcentage de points pouvant être sélectionnés comme supports ($\beta_i \neq 0$) de la solution,
- le critère de Ljung [Lju87], où un seuil est fixé sur l'auto-correlation du résidu qui mesure sa ressemblance à un bruit blanc.

2.4.2 Critères auto-adaptatifs

L'idée est d'ajouter des sources d'information piégées, que nous ne souhaitons pas utiliser pour construire la solution. Lorsque ces sources sont sélectionnées par le LARS, nous arrêtons la procédure d'apprentissage. Nous proposons deux heuristiques pour construire ces sources pièges :

- LARS-RV. Nous introduisons des vecteurs de variables aléatoires gaussiennes parmi les sources d'informations, à la manière de [BBE$^+$03]. Lorsque ces sources seront sélectionnées, cela signifie qu'elles sont plus corrélées au résidu que n'importe quelle autre source d'information. Le résidu peut alors être assimilé à du bruit et l'apprentissage stoppé.
- LARS-HF. Nous utilisons alors les échelles hautes fréquences comme des échelles pièges. En effet, dans le noyau correspondant aux hautes fréquences (HF), les points d'apprentissage n'ont d'influence que dans une très petite sphère, c'est à dire essentiellement sur eux mêmes. La sélection d'une source d'information HF peut donc être interprétée comme un début de sur-apprentissage.

Cette dernière famille de solution est particulièrement intéressante puisqu'elle permet d'avoir un algorithme complètement non-paramétrique.

3 Résultats

Nous avons comparé le LARS, avec plusieurs critères d'arrêt, par rapport aux ϵ-SVM et au *back-fitting* [HT86] pour construire une estimation des signaux artificiels présentés figure 1. il s'agit de signaux classiques, utilisés par Donoho et Johnstone [DJ94]. L'échantillonnage est aléatoire sur l'intervalle [0, 1] et les données sont bruitées. Nous avons utilisé un signal de 400 points pour l'apprentissage et une base de test (non bruitée) de 1000 points. Le critère d'évaluation est l'erreur moyenne au sens des moindres carrés. Les bornes du LARS ($\sum_i |\beta_i|$, ν) sont optimisées par validation croisée. Les paramètres des SVM (ϵ, C et le choix du noyau) sont optimisés par triple validation croisée sur l'ensemble d'apprentissage, chaque échelle est donc traitée séparément. L'algorithme du *back-fitting* nécessite le réglage d'autant de paramètres qu'il y a de noyaux. Ici encore, nous avons utilisé la validation croisée pour estimer les compromis optimaux. Nous avons utilisé 10 échelles d'ondelettes de type *Symmlet*, correspondant aux paramètres de dilatation des ondelettes de 0 à 9.

Les résultats sont présentés dans le tableau 1.

Les méthodes LARS, basées sur les noyaux d'ondelettes multirésolution donnent les meilleurs résultats sur les bases de données faisant intervenir un aspect multirésolution (cos(exp(ωt)),Doppler,...) , mais aussi sur des bases plus classique (Ramp, Bumps...). A l'inverse, les SVM se montrent très efficaces sur les bases de données régulières (HeavySine). L'ϵ-tube autour de la solution est bien adapté à ce genre de problèmes. La parcimonie de la solution est systématiquement à l'avantage du LARS, du fait de la régularisation ℓ_1. Les résultats des méthodes LARS-RV et LARS-HF sont très intéressants : ils sont proches des meilleurs résultats dans la plupart des cas, sans faire intervenir de paramètres.

4 Conclusions

Les résultats obtenus sont très encourageants : ils montrent la performance de la méthode pour différents types de signaux artificiels, tout en améliorant la rapidité de traitement et en diminuant le nombre de paramètres nécessaires par rapport aux SVM. L'écart de performance est plus important sur les problèmes nécessitant une approche multirésolution.

Il reste néanmoins à valider cette méthode sur des cas réels. Une autre perspective de ce travail concerne l'optimisation de la forme de l'ondelette mère utilisée. Enfin, notons que l'estimation des signaux par cette méthode peut servir de description pour les problèmes de classification de signaux. Contrairement à [DJ94] ou [KS00], la description ainsi obtenue est covariante en translation.

Algorithme	Back-fitting	ϵ-SVM	LARS- $\sum_i \lvert\beta_i\rvert$	ν-LARS	LARS-RV	LARS-HF
Blocks	0.024 ± 0.0041 4000 0	0.026 ± 0.0045 107.7 0	0.021 ± 0.0033 50.3 17	$\mathbf{0.020 \pm 0.0034}$ 54 21	0.023 ± 0.0039 60.8 5	0.022 ± 0.0036 49.2 8
Doppler	0.019 ± 0.0041 4000 0	0.025 ± 0.0056 110.7 0	0.014 ± 0.0042 44.1 11	$\mathbf{0.013 \pm 0.0037}$ 42 24	0.014 ± 0.0045 47.1 9	0.013 ± 0.0038 42.5 21
Ramp	0.012 ± 0.0036 4000 0	0.014 ± 0.0037 57.6 2	$\mathbf{0.008 \pm 0.0035}$ 19.2 19	$\mathbf{0.008 \pm 0.0035}$ 17 20	0.009 ± 0.0037 21.4 13	0.010 ± 0.0035 15.8 8
HeavySine	0.0033 ± 0.0008 4000 0	$\mathbf{0.0028 \pm 0.0005}$ 29 28	0.0034 ± 0.0008 17.4 0	0.0033 ± 0.0008 18 2	0.0033 ± 0.0007 19.6 0	0.0034 ± 0.0008 20.7 2
$\cos(\exp(\omega t))$	0.025 ± 0.0064 4000 0	0.029 ± 0.0056 111.1 0	$\mathbf{0.020 \pm 0.0047}$ 42.3 24	0.020 ± 0.0048 42 19	0.021 ± 0.0049 48.7 16	$\mathbf{0.020 \pm 0.0047}$ 43.4 20
Riemann	0.019 ± 0.0016 4000 5	$\mathbf{0.018 \pm 0.0017}$ 57.9 15	0.019 ± 0.0020 51.4 7	0.019 ± 0.0019 50 8	0.019 ± 0.0019 52.1 6	0.019 ± 0.0019 57.5 6
Bumps	0.024 ± 0.0071 4000 0	0.025 ± 0.0073 52.7 0	$\mathbf{0.020 \pm 0.0050}$ 42.2 23	0.020 ± 0.0052 46 19	0.020 ± 0.0052 45.3 19	0.020 ± 0.0051 41.7 18
Piece-Polynomial	0.018 ± 0.0050 4000 0	0.019 ± 0.0051 68.8 0	$\mathbf{0.015 \pm 0.0049}$ 53.1 18	0.015 ± 0.0053 52 16	0.015 ± 0.0052 55.4 14	0.015 ± 0.0053 59.1 15

TAB. 1 – Tableau de résultats. Moyenne et écart-type de l'erreur au sens des moindres carrés sur 30 itérations. Moyenne des vecteurs supports nécessaires pour construire la solution. Nombre de meilleures performances sur les 30 itérations (le total dépasse parfois 30, plusieurs solutions LARS étant identiques).

Remerciements

Ce travail est financé en partie par le programme IST de la communauté européenne, avec le réseau d'excellence PASCAL, IST-2002-506778. Cette publication reflète uniquement le point de vue des auteurs.

Références

[AAP04] U. Amato, A. Antoniadis, and M. Pensky. Wavelet kernel penalized estimation for non-equispaced design regression. Technical report, Istituto per le Applicazioni del Calcolo Mauro Picone, 2004.

[BBE+03] J. Bi, K. Bennett, M. Embrechts, C. Breneman, and M. Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3 :1229–1243, 2003.

[BTJ04] F.R. Bach, R. Thibaux, and M.I. Jordan. Computing regularization paths for learning multiple kernels. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

[CDS98] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1) :33–61, 1998.

[DJ94] D. Donoho and I. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81 :425–455, 1994.

[EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of statistics*, 32(2) :407–499, 2004.

[Gra98] Y. Grandvalet. Least absolute shrinkage is equivalent to quadratic penalization. In *ICANN*, pages 201–206, 1998.

[HT86] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1 :297–318, 1986.

[KS00] A. Kovac and B.W. Silverman. Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of the American Statistical Association*, 95 :172–183, 2000.

[KW71] G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *J. Math. Anal. Applic.*, 33 :82–95, 1971.

[LCV+04] G. Loosli, S. Canu, S.V.N. Vishwanathan, A. J. Smola, and M. Chattopadhyay. Une boîte à outils rapide et simple pour les svm. In *CAp*, 2004.

[Lju87] L. Ljung. *System Identification - Theory for the User*. 1987.

[Mal97] S. Mallat. *A Wavelet Tour Of Signal Processing*. Academic Press, 1997.

[RC05] A. Rakotomamonjy and S. Canu. Frame, reproducing kernel, regularization and learning. Technical report, Lab. Perception, Systemes, information, 2005.

[Tib96] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist.*, 58(1) :267–288, 1996.

[VB02] P. Vincent and Y. Bengio. Kernel matching pursuit. *Machine Learning Journal*, 48(1) :165–187, 2002.

Une procédure générale d'estimation efficace de paramètres multiples avec application en localisation

P. FORSTER et G. GINOLHAC

GEA, Univ. Paris X, PST Ville d'Avray, 1 ch. Desvallières 92410 Ville d'Avray

philippe.forster,guillaume.ginolhac@u-paris10.fr

Résumé – *Nous présentons dans ce papier un cadre théorique pour affiner une estimation efficace de plusieurs paramètres en tenant compte des liens entre eux. La méthode proposée s'appuie sur un théorème généralisant l'"EXtended Invariance Principle" (EXIP). Nous montrons que l'approche développée permet de résoudre analytiquement certains problèmes non linéaires. Nous appliquons enfin ce cadre théorique à un problème de localisation par un radar multistatique.*

Abstract – *This article deals with a general procedure for efficient estimation of multiple parameters. The proposed technique is based on an original theorem which is established in the paper. The resulting method allows in some non linear cases an efficient estimation without iterative procedures. This new approach is applied to localize a target and estimate its speed in multistatic radar.*

1. Introduction

En traitement du signal, le problème d'estimation suivant est très courant. Une observation bruitée $\mathbf{m} \in \mathrm{IR}^N$ est reliée à un vecteur de paramètres inconnus $\boldsymbol{\theta} \in \mathrm{IR}^{P_\theta}$ par la relation :

$$\mathbf{m} = \mathbf{s}(\boldsymbol{\theta}_0) + \mathbf{n},$$

où $P_\theta \leq N$, $\mathbf{s}(\boldsymbol{\theta})$ est la fonction décrivant la dépendance entre les mesures non bruitées et le vecteur des paramètres, $\mathbf{n}$ est un bruit additif supposé gaussien de moyenne nulle et de matrice de covariance connue $\mathbf{Q} = \sigma^2 \mathbf{Q}_n$. Dans ce contexte, l'estimateur du Maximum de Vraisemblance (MV) est souvent retenu pour ses bonnes propriétés statistiques. En effet, cet estimateur est asymptotiquement efficace et Gaussien lorsque la puissance du bruit tend vers 0 :

$$\begin{cases} \dfrac{1}{\sigma}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{a.d.} N(\mathbf{0}, \mathbf{F}_\theta^{-1}) \\ \sigma \longrightarrow 0 \end{cases},$$

où $N(\mathbf{0}, \mathbf{V})$ désigne la distribution gaussienne de moyenne nulle et de covariance $\mathbf{V}$, et $\mathbf{F}_\theta$ est la Matrice d'Information de Fisher (MIF). L'estimateur du MV $\hat{\boldsymbol{\theta}}_{ML}$ de $\boldsymbol{\theta}_0$ s'obtient en minimisant la fonction suivante :

$$f(\boldsymbol{\theta}) = (\mathbf{m} - \mathbf{s}(\boldsymbol{\theta}))^T \ \mathbf{Q}_n^{-1} \ (\mathbf{m} - \mathbf{s}(\boldsymbol{\theta})).$$

Excepté le cas où $\mathbf{s}(\boldsymbol{\theta})$ est une fonction linéaire, il n'existe pas de solution explicite à la minimisation de $f(\boldsymbol{\theta})$ et l'on a souvent recours à un algorithme itératif (e.g. Gauss-Newton) nécessitant une initialisation *ad-hoc*. C'est pourquoi il serait intéressant de dériver des estimateurs efficaces qui contournent la procédure itérative pour éviter les difficultés de calcul associées au MV tout en gardant ses bonnes propriétés statistiques. Une première contribution de ce papier est un théorème fournissant un cadre général pour trouver de tels estimateurs, et une seconde contribution est d'appliquer ce théorème à un problème de localisation en radar multistatique. La méthode théorique proposée peut être vue comme une généralisation du *EXtended Invariance Principle* (EXIP) [1].

2. Résultat théorique

Le théorème suivant utilise deux paramétrisations. La première, dite *non contrainte*, est composée de plusieurs paramètres pouvant avoir des relations entre eux mais qui ne sont pas utilisées. La seconde paramétrisation, dite *contrainte*, est composée uniquement des paramètres n'ayant pas de lien entre eux. Le but du théorème est d'exploiter les résultats de l'estimation avec la paramétrisation *non contrainte* pour obtenir facilement une estimation efficace avec la seconde paramétrisation. Comme celle-ci utilise les liens des paramètres entre eux, le résultat de l'estimation sera bien entendu plus précis. Plus précisément, soit $\mathbf{m}$ un vecteur Gaussien de matrice de covariance connue $\mathbf{Q} = \sigma^2 \mathbf{Q}_n$ et de moyenne $\mathbf{s}$ dépendant d'un ensemble de paramètres inconnus. Soit $\mathbf{u} = \begin{pmatrix} u_1 & \cdots & u_{P_u} \end{pmatrix}^T \in \mathrm{IR}^{P_u}$ la première paramétrisation (dite *non contrainte*), notée Π_u, de la moyenne avec une valeur exacte notée $\mathbf{u}_0$, et $\mathbf{v} = \begin{pmatrix} v_1 & \cdots & v_{P_v} \end{pmatrix}^T \in \mathrm{IR}^{P_v}$ une seconde (dite *contrainte*) notée Π_v (valeur exacte $\mathbf{v}_0$) qui est plus restrictive au sens où $P_v < P_u$. Soit $\mathbf{u}(\mathbf{v})$ les valeurs des paramètres de la première paramétrisation atteints par la seconde. Un calcul simple montre que les MIF correspondantes $\mathbf{F}_u$ et $\mathbf{F}_v$ pour $\sigma = 1$ ont pour expressions :

$$\mathbf{F}_u = (\partial_{\mathbf{u}} \mathbf{s})_{\mathbf{u}_0}^T \ \mathbf{Q}_n^{-1} \ (\partial_{\mathbf{u}} \mathbf{s})_{\mathbf{u}_0}, \quad (1)$$

$$\mathbf{F}_v = (\partial_{\mathbf{v}} \mathbf{u})_{\mathbf{v}_0}^T \ \mathbf{F}_u \ (\partial_{\mathbf{v}} \mathbf{u})_{\mathbf{v}_0}, \quad (2)$$

où $(\partial_{\mathbf{u}} \mathbf{s})$ et $(\partial_{\mathbf{v}} \mathbf{u})$ sont les matrices des dérivées partielles $N \times P_u$ et $P_u \times P_v$ par rapport à $\mathbf{u}$ et $\mathbf{v}$. Soit alors $\hat{\mathbf{u}}$ un estimateur efficace et asymptotiquement Gaussien non contraint de $\mathbf{u}_0$:

$$\begin{cases} \dfrac{1}{\sigma}(\hat{\mathbf{u}} - \mathbf{u}_0) \xrightarrow{a.d.} N(\mathbf{0}, \mathbf{F}_u^{-1}) \\ \sigma \longrightarrow 0 \end{cases}, \quad (3)$$

et notons $\hat{\mathbf{F}}_u$ un estimateur consistent de la MIF (e.g. la MIF calculée avec les valeurs des paramètres estimés $\hat{\mathbf{u}}$ au

lieu de $\mathbf{u}_0$). Considérons maintenant une fonction $\mathbf{g}(\mathbf{u},\mathbf{v})$ de $\mathrm{IR}^{P_u} \times \mathrm{IR}^{P_v}$ vers IR^{P_u}, et notons :

$$\partial_{\mathbf{u}}\mathbf{g} = \left[\partial_{u_1}\mathbf{g}(\mathbf{u},\mathbf{v}) \quad \cdots \quad \partial_{u_{P_u}}\mathbf{g}(\mathbf{u},\mathbf{v})\right]. \qquad (4)$$

Théorème : La minimisation du critère $\tilde{C}(\mathbf{v}) = \left\|\mathbf{g}(\hat{\mathbf{u}},\mathbf{v}) - \mathbf{g}(\mathbf{u}(\mathbf{v}),\mathbf{v})\right\|^2$ conduit à un estimateur *consistant* $\tilde{\mathbf{v}}$ de $\mathbf{v}_0$. La minimisation du critère

$$C(\mathbf{v}) = (\mathbf{g}(\hat{\mathbf{u}},\mathbf{v}) - \mathbf{g}(\mathbf{u}(\mathbf{v}),\mathbf{v}))^T \left((\partial_{\mathbf{u}}\mathbf{g})^{-T}_{\hat{\mathbf{u}},\tilde{\mathbf{v}}} \hat{\mathbf{F}}_u \, (\partial_{\mathbf{u}}\mathbf{g})^{-1}_{\hat{\mathbf{u}},\tilde{\mathbf{v}}}\right)(\mathbf{g}(\hat{\mathbf{u}},\mathbf{v}) - \mathbf{g}(\mathbf{u}(\mathbf{v}),\mathbf{v}))$$

(où $\tilde{\mathbf{v}}$ est un estimateur consistant de $\mathbf{v}_0$) conduit à un estimateur *efficace* et asymptotiquement Gaussien $\hat{\mathbf{v}}$ de $\mathbf{v}_0$:

$$\begin{cases} \dfrac{1}{\sigma}(\hat{\mathbf{v}} - \mathbf{v}_0) \xrightarrow{a.d.} N(\mathbf{0}, \mathbf{F}_v^{-1}) \\ \sigma \longrightarrow 0 \end{cases}.$$

On notera $\hat{\mathbf{F}}_v$ la MIF estimée à partir de $\hat{\mathbf{F}}_u$:

$$\hat{\mathbf{F}}_v = (\partial_{\mathbf{v}}\mathbf{u}(\mathbf{v}))^T_{\hat{\mathbf{v}}} \, \hat{\mathbf{F}}_u \, (\partial_{\mathbf{v}}\mathbf{u}(\mathbf{v}))_{\hat{\mathbf{v}}} \quad . \qquad (5)$$

Démonstration : Considérons la fonction $F_W(\mathbf{u},\mathbf{v}) = (g(\mathbf{u},\mathbf{v}) - g(\mathbf{u}(\mathbf{v}),\mathbf{v}))^T \mathbf{W} (g(\mathbf{u},\mathbf{v}) - g(\mathbf{u}(\mathbf{v}),\mathbf{v}))$ où $\mathbf{W}$ est une matrice définie positive. Cette fonction atteint une valeur minimale pour $\mathbf{u}=\mathbf{u}_0$ et $\mathbf{v}=\mathbf{v}_0$. Maintenant, il est facile de montrer que, pour $\mathbf{u}=\hat{\mathbf{u}}$, le théorème des fonctions implicites appliqué au gradient de F_W fournit la relation suivante entre la valeur $\mathbf{v}$ (qui minimise $F_W(\hat{\mathbf{u}},\mathbf{v})$) et $\hat{\mathbf{u}}$:

$$\Delta\mathbf{v} = \left(\partial_{\mathbf{v}}\mathbf{u}^T \partial_{\mathbf{u}}\mathbf{g}^T \mathbf{W} \, \partial_{\mathbf{u}}\mathbf{g} \, \partial_{\mathbf{v}}\mathbf{u}\right)^{-1} \left(\partial_{\mathbf{v}}\mathbf{u}^T \partial_{\mathbf{u}}\mathbf{g}^T \mathbf{W} \, \partial_{\mathbf{u}}\mathbf{g}\right)\Delta\mathbf{u} + o(\Delta\mathbf{u}) \quad (6)$$

où :

- $\Delta\mathbf{u} = \hat{\mathbf{u}} - \mathbf{u}_0$ et $\Delta\mathbf{v} = \mathbf{v} - \mathbf{v}_0$;
- $\partial_{\mathbf{v}}\mathbf{u}$ et $\partial_{\mathbf{u}}\mathbf{g}$ sont évalués à $\mathbf{v}_0$ et $(\mathbf{u}_0, \mathbf{v}_0)$.

La consistance de $\mathbf{v}$ se déduit immédiatement de l'expression (6) ; en particulier, la consistance de $\tilde{\mathbf{v}}$ affirmée dans le théorème correspond au choix $\boldsymbol{W=I}$. De plus, nous pouvons écrire pour le terme $o(\Delta\mathbf{u})$ de l'équation (6) :

$$\left\|\frac{1}{\sigma} o(\Delta\mathbf{u})\right\| = \left\|\frac{1}{\sigma}\Delta\mathbf{u}\right\| \frac{\left\|o(\Delta\mathbf{u})\right\|}{\left\|\Delta\mathbf{u}\right\|}.$$

Quand $\sigma \longrightarrow 0$ dans l'expression ci-dessus, $(1/\sigma).\Delta\mathbf{u}$ converge en loi vers un vecteur aléatoire gaussien (hypothèse (3)) et $\dfrac{\left\|o(\Delta\mathbf{u})\right\|}{\left\|\Delta\mathbf{u}\right\|}$ converge en probabilité vers 0. Donc, $\left\|(1/\sigma).o(\Delta\mathbf{u})\right\|$ converge en probabilité vers 0 et nous concluons d'après (6) que $(1/\sigma).\Delta\mathbf{v}$ est asymptotiquement gaussien de moyenne nulle et de matrice de covariance :

$$\left(\partial_{\mathbf{v}}\mathbf{u}^T \partial_{\mathbf{u}}\mathbf{g}^T \mathbf{W} \partial_{\mathbf{u}}\mathbf{g} \, \partial_{\mathbf{v}}\mathbf{u}\right)^{-1} \left(\partial_{\mathbf{v}}\mathbf{u}^T \partial_{\mathbf{u}}\mathbf{g}^T \mathbf{W} \partial_{\mathbf{u}}\mathbf{g}\right)\mathbf{F}_u^{-1}$$
$$\left(\partial_{\mathbf{u}}\mathbf{g}^T \mathbf{W} \, \partial_{\mathbf{u}}\mathbf{g} \, \partial_{\mathbf{v}}\mathbf{u}\right)\left(\partial_{\mathbf{v}}\mathbf{u}^T \partial_{\mathbf{u}}\mathbf{g}^T \mathbf{W} \partial_{\mathbf{u}}\mathbf{g} \, \partial_{\mathbf{v}}\mathbf{u}\right)^{-1}.$$

Le choix $\mathbf{W}_{opt} = \partial_{\mathbf{u}}\mathbf{g}^{-T} \mathbf{F}_u \, \partial_{\mathbf{u}}\mathbf{g}^{-1}$ conduit à la valeur $\left(\partial_{\mathbf{v}}\mathbf{u}^T \mathbf{F}_u \partial_{\mathbf{v}}\mathbf{u}\right)^{-1}$ de la covariance asymptotique ci-dessus de $(1/\sigma).\Delta\mathbf{v}$: ceci est juste la borne de Cramer Rao comme le montre l'expression (2) de la MIF. Donc, la minimisation de $F_{W_{opt}}(\hat{\mathbf{u}},\mathbf{v})$ conduit à une estimation efficace et asymptotiquement gaussienne de $\mathbf{v}_0$. Finalement, nous pouvons remplacer $\mathbf{W}_{opt}$ par n'importe quel estimé $\hat{\mathbf{W}}$ de $\mathbf{W}_{opt}$ tel que $\hat{\mathbf{W}} = (\partial_{\mathbf{u}}\mathbf{g})^{-T}_{\hat{\mathbf{u}},\tilde{\mathbf{v}}} \, \hat{\mathbf{F}}_u \, (\partial_{\mathbf{u}}\mathbf{g})^{-1}_{\hat{\mathbf{u}},\tilde{\mathbf{v}}}$ sans changer les conclusions : la fonction critère résultante $F_{\hat{W}}(\hat{\mathbf{u}},\mathbf{v})$ est juste le critère $C(\mathbf{v})$ du théorème.

L'originalité de ce théorème par rapport à la méthode EXIP [1] tient à l'introduction de la fonction $\mathbf{g}$ dans le critère : en effet, la méthode EXIP minimise un critère du type $\|\hat{\mathbf{u}} - \mathbf{u}(\mathbf{v})\|^2$ (dans une métrique appropriée). Notre théorème est particulièrement utile quand la fonction $\mathbf{g}(\hat{\mathbf{u}},\mathbf{v}) - \mathbf{g}(\mathbf{u}(\mathbf{v}),\mathbf{v})$ est linéaire par rapport à $\mathbf{v}$ puisque les critères $\tilde{C}(\mathbf{v})$ et $C(\mathbf{v})$ peuvent être minimisés exactement dans ce cas. Par exemple, il est facile de montrer que l'estimateur asymptotiquement efficace développé par Chan [2] pour la localisation de sources (voir aussi [3-4] pour des applications de localisation de mobiles) est une conséquence directe de notre théorème. Dans notre article, nous en déduisons un algorithme original pour localiser une cible à l'aide d'un radar multistatique.

3. Application à la localisation

Le radar multistatique comporte *N* émetteurs, localisés en $(x_i, y_i)_{1\le i\le N}$ et un seul récepteur, localisé en *(0,0)*. La cible est supposée se déplacer dans un plan. Ses coordonnées exactes sont notées (x_0, y_0) et ses coordonnées de vitesse (v_{x0}, v_{y0}). Notre vecteur de mesures $\mathbf{m} = (\delta_1 \cdots \delta_N \, \dot{\delta}_1 \cdots \dot{\delta}_N)^T$ consiste à *N* distances de propagation entre le récepteur et les émetteurs *1, 2, …, N*, et *N* dérivés de la distance de propagation correspondantes obtenues par des mesures Doppler. Introduisons quelques notations :

- $r_i = \sqrt{(x_0 - x_i)^2 + (y_0 - y_i)^2} \qquad (1 \le i \le N)$ et $r = \sqrt{x_0{}^2 + y_0{}^2}$
- $d_i = \dfrac{v_{x0}(x_0 - x_i) + v_{y0}(y_0 - y_i)}{\sqrt{(x_0 - x_i)^2 + (y_0 - y_i)^2}} \qquad (1 \le i \le N)$ et $d = \dfrac{v_{x0}.x_0 + v_{y0}.y_0}{\sqrt{x_0{}^2 + y_0{}^2}}$

Les expressions des mesures exactes (sans bruit) pour l'émetteur *i* sont :

$$\delta_i = r_i + r \text{ et } \dot{\delta}_i = d_i + d$$

Nous supposons qu'un bruit Gaussien, centré, de matrice de covariance connue $\mathbf{Q} = \sigma^2 \mathbf{Q}_n$ est ajouté aux mesures. La méthode développée ci-après peut être vue comme une application multiple de notre théorème. Dans les différentes étapes de la méthode, nous nous référons à $\mathbf{u}$, $\mathbf{v}$, $\tilde{\mathbf{u}}$, $\tilde{\mathbf{v}}$, $\hat{\mathbf{u}}$ et $\hat{\mathbf{v}}$ en mettant un exposant $^{(i)}$ pour la paramétrisation *i*.

- étape 1 :

Paramétrisation Π_u : $\mathbf{u}^{(1)}$ est égal au vecteur de mesures $\mathbf{m}$ (vecteur de IR^{2N}). Sa MIF est $\hat{\mathbf{F}}_{u^{(1)}} = \mathbf{Q}^{-1}$. Le but est maintenant de trouver une relation linéaire entre les paramètres $(x_0, y_0, v_{x0}, v_{y0})$ et $\mathbf{u}^{(1)}$ quitte à introduire de

nouveaux paramètres inconnus. Dans un autre contexte, Chan [2] fournit la méthode pour obtenir la relation suivante pour l'émetteur i :

$$u_i^{(1)2} - 2.u_i^{(1)}.r = x_i^2 + y_i^2 - 2.x_0.x_i - 2.y_0.y_i \ . \qquad (7)$$

On note que la première partie des mesures (distances de propagation) et certains paramètres exacts (x_0, y_0) sont liés par une relation linéaire si nous considérons r comme un paramètre inconnu. Pour les mesures des dérivées de la distance de propagation, nous avons la relation suivante :

$$r_i.d_i - r.d = -v_{x0}.x_i - v_{y0}.y_i \ .$$

Nous pouvons obtenir une autre relation :

$$r_i.d_i - r.d = r_{\cdot i}\,d_i + r.d_i - r.d - r.d_i$$
$$r_i.d_i - r.d = u_{i+N}^{(1)}.u_i^{(1)} - d.u_i^{(1)} - r.u_{i+N}^{(1)}$$

En combinant les deux équations obtenues, on obtient la relation suivante :

$$u_{i+N}^{(1)}.u_i^{(1)} - d.u_i^{(1)} - r.u_{i+N}^{(1)} = -v_{x0}.x_i - v_{y0}.y_i \qquad (8)$$

Comme pour la relation (7), l'équation (8) est une relation linéaire entre $\mathbf{u}^{(1)}$ et les paramètres exacts si nous traitons d comme un paramètre inconnu.

Paramétrisation Π_v : $\mathbf{v}^{(1)}$ est un vecteur de IR^6. Sa valeur exacte $\mathbf{v}_0{}^{(1)} = (v_{01}\ v_{02}\ v_{03}\ v_{04}\ v_{05}\ v_{06})^T$ a des composantes reliées à $(x_0, y_0, v_{x0}, v_{y0})$ par $v_{01} = x_0$, $v_{02} = y_0$, $v_{03} = v_{x0}$, $v_{04} = v_{y0}$, $v_{05} = r$ et $v_{06} = d$.

Le côté gauche des équations (7) et (8) nous permet d'obtenir $\mathbf{g}(\mathbf{u}(\mathbf{v}), \mathbf{v})$ tandis que le côté droit nous fournit $\mathbf{g}(\mathbf{u}, \mathbf{v})$. Il est alors possible de minimiser le critère $\tilde{C}(\mathbf{v})$ et donc de fournir un estimateur consistant $\tilde{\mathbf{v}}^{(1)}$. A l'aide de cette estimation consistante, nous pouvons calculer la matrice (4) $\partial_{\mathbf{u}}\mathbf{g}$:

$$\frac{\partial \mathbf{g}}{\partial \mathbf{u}} = \begin{pmatrix} 2u_1^{(1)} - 2\tilde{r} & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 2u_N^{(1)} - 2\tilde{r} & 0 & \cdots & 0 \\ u_{N+1}^{(1)} - \tilde{d} & \cdots & 0 & u_1^{(1)} - \tilde{r} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & u_{N+N}^{(1)} - \tilde{d} & 0 & \cdots & u_{N+1}^{(1)} - \tilde{r} \end{pmatrix}$$

Nous obtenons ensuite l'estimation efficace $\hat{\mathbf{v}}^{(1)}$ de $\mathbf{v}_0{}^{(1)}$ en minimisant le critère $C(\mathbf{v})$. Enfin, nous calculons la MIF de $\mathbf{v}_0{}^{(1)}$ (2) en utilisant la matrice suivante $\partial_{\mathbf{v}}\mathbf{u}(\mathbf{v})$:

$$\frac{\partial \mathbf{u}(\mathbf{v})}{\partial \mathbf{v}} = \begin{pmatrix} -2x_1 & -2y_1 & 0 & 0 & 2u_1^{(1)} & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -2x_N & -2y_N & 0 & 0 & 2u_N^{(1)} & 0 \\ 0 & 0 & -x_1 & -y_1 & u_{N+1}^{(1)} & u_1^{(1)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & -x_N & -y_N & u_{N+N}^{(1)} & u_N^{(1)} \end{pmatrix}$$

A la fin de cette première étape, le problème est sur-paramétré (il existe des liens entre les paramètres) et l'estimation peut être améliorée en tenant compte des relations entre paramètres : ceci fait l'objet des étapes suivantes qui sont obtenues par des applications successives du théorème proposé. Par manque de place et pour plus de clarté, nous donnons ici uniquement le squelette des différentes étapes.

- étape 2 : cette étape est uniquement un changement de variable.

Paramétrisation Π_u : $\mathbf{u}^{(2)}$ est égal à l'estimé $\hat{\mathbf{v}}^{(1)}$ (vecteur de IR^6). Cela signifie que $\hat{\mathbf{F}}_{u^{(2)}} = \hat{\mathbf{F}}_{v^{(1)}}$. Nous introduisons deux nouveaux paramètres : $z_x = x.v_x$ et $z_y = y.v_y$.

Paramétrisation Π_v : $\mathbf{v}^{(2)}$ est un vecteur de IR^6. Sa valeur exacte $\mathbf{v}_0{}^{(2)} = (v_{01}\ v_{02}\ v_{03}\ v_{04}\ v_{05}\ v_{06})^T$ a des composantes reliées à $(x_0, y_0, v_{x0}, v_{y0})$ par $v_{01} = x_0$, $v_{02} = y_0$, $v_{03} = z_{x0}$, $v_{04} = z_{y0}$, $v_{05} = r$ et $v_{06} = d$.

Les relations précédentes permettent de déterminer $\mathbf{g}(\mathbf{u}(\mathbf{v}), \mathbf{v})$ et $\mathbf{g}(\mathbf{u}, \mathbf{v})$, et donc l'estimation efficace $\hat{\mathbf{v}}^{(2)}$ de $\mathbf{v}_0{}^{(2)}$. Finalement, nous calculons la MIF $\hat{\mathbf{F}}_{v^{(2)}}$ of $\hat{\mathbf{v}}^{(2)}$.

- étape 3 :

Paramétrisation Π_u : $\mathbf{u}^{(3)}$ est égal à l'estimé $\hat{\mathbf{v}}^{(2)}$ (vecteur de IR^6). Cela signifie que $\hat{\mathbf{F}}_{u^{(3)}} = \hat{\mathbf{F}}_{v^{(2)}}$.

Paramétrisation Π_v : $\mathbf{v}^{(3)}$ est un vecteur de IR^5. Sa valeur exacte $\mathbf{v}_0{}^{(3)} = (v_{01}\ v_{02}\ v_{03}\ v_{04}\ v_{05})^T$ a des composantes reliées à $(x_0, y_0, v_{x0}, v_{y0})$ par $v_{01} = x_0$, $v_{02} = y_0$, $v_{03} = z_{x0}$, $v_{04} = z_{y0}$, $v_{05} = r$ et $v_{05}.d = z_{x0} + z_{y0}$.

Nous obtenons l'estimé efficace $\hat{\mathbf{v}}^{(3)}$ et sa MIF $\hat{\mathbf{F}}_{v^{(3)}}$.

- étape 4 :

Paramétrisation Π_u : $\mathbf{u}^{(4)}$ est égal à l'estimé $\hat{\mathbf{v}}^{(3)}$ (vecteur de IR^5). Cela signifie que $\hat{\mathbf{F}}_{u^{(4)}} = \hat{\mathbf{F}}_{v^{(3)}}$.

Paramétrisation Π_v : $\mathbf{v}^{(4)}$ est un vecteur de IR^4. Sa valeur exacte $\mathbf{v}_0{}^{(4)} = (v_{01}\ v_{02}\ v_{03}\ v_{04})^T$ a des composantes reliées à $(x_0, y_0, v_{x0}, v_{y0})$ par $v_{01} = x_0^2$, $v_{02} = y_0^2$, $v_{03} = z_{x0}$, $v_{04} = z_{y0}$ et $v_{01} + v_{02} = r^2$.

Nous obtenons l'estimé efficace $\hat{\mathbf{v}}^{(4)}$ et sa MIF $\hat{\mathbf{F}}_{v^{(4)}}$.

- étape 5 : cette étape est uniquement un changement de variable.

Paramétrisation Π_u : $\mathbf{u}^{(5)}$ est égal à l'estimé $\hat{\mathbf{v}}^{(4)}$ (vecteur de IR^4). Cela signifie que $\hat{\mathbf{F}}_{u^{(5)}} = \hat{\mathbf{F}}_{v^{(4)}}$.

Paramétrisation Π_v : $\mathbf{v}^{(5)}$ est un vecteur de IR^4. Sa valeur exacte $\mathbf{v}_0{}^{(4)} = (v_{01}\ v_{02}\ v_{03}\ v_{04})^T$ a des composantes reliées à $(x_0, y_0, v_{x0}, v_{y0})$ par $v_{01} = x_0$, $v_{02} = y_0$, $v_{03} = v_{x0}$ et $v_{04} = v_{y0}$.

Nous obtenons l'estimé efficace $\hat{\mathbf{v}}^{(5)}$ (et donc $\hat{\boldsymbol{\theta}}$) et sa MIF $\hat{\mathbf{F}}_{v^{(5)}}$.

4. Résultats de simulation

Nous supposons disposer de 6 émetteurs localisés en *{(0,0.5);(10,0);(8,7);(0,3);(4,9);(1,2)}* km. La cible est présente à la position *(5,5)* km et sa vitesse est égale à *(80,80)* m/s. Un bruit blanc Gaussien, de matrice de covariance $\mathbf{Q} = \sigma^2 \mathbf{I}_6$, est ajouté aux mesures. Les variances sont estimées à partir de 1000 essais. Elles sont comparées à la Borne de Cramer Rao (BCR) pour chaque paramètre et tracées en fonction de $20.\log_{10}\sigma$. Les figures 1 et 2 montrent les résultats pour $\hat{x}$ et $\hat{v}_x$ alors que les résultats pour $\hat{y}$ et $\hat{v}_y$ sont montrés sur les figures 3 et 4. Les performances sont très proches de la BCR et montrent le caractère asymptotiquement efficace de la méthode proposée.

5. Conclusion

La méthode proposée permet une estimation efficace sans procédure itérative. Après la présentation du théorème clé sur lequel elle s'appuie, nous l'avons appliquée à un problème de localisation à l'aide d'un radar multistatique. Les résultats de simulation confirment la théorie.

Références

[1] P. Stoica, and, A. L. Swindlehurst, "Maximum Likelihood Methods in Radar Array Signal Processing," *Proceedings of the IEEE*, vol. 86, n°2, pp. 419-441, February 1998.

[2] Y. T. Chan, and, K. C. Ho, " A simple and Efficient Estimator for Hyperbolic Location," *IEEE Trans. On Signal Proc.*, vol. 42, n°8, pp. 1905-1915, August 1994.

[3] A. Wan, and, Y. N. Peng, "Enhanced Linear Estimator for Mobile Location Using Range Difference Measurements," *IEICE Trans. on communications*, vol. E85-B, n°5, pp. 946-950, May 2002.

[4] Grosicki, E., *Techniques Avancées du traitement d'antenne pour la localisation de sources*, PhD. Thesis, December 2003.

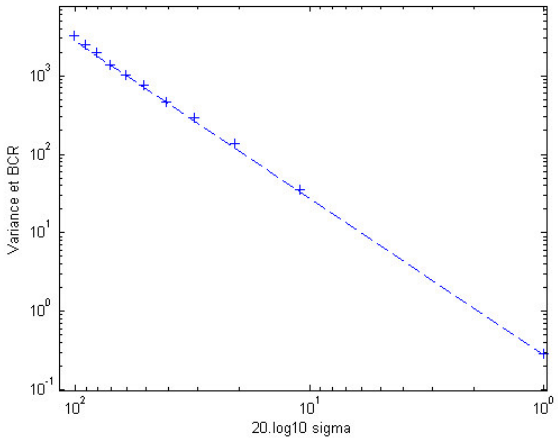


FIG. 1 : BCR de $\hat{x}$ (-) et variance de $\hat{x}$ (+).

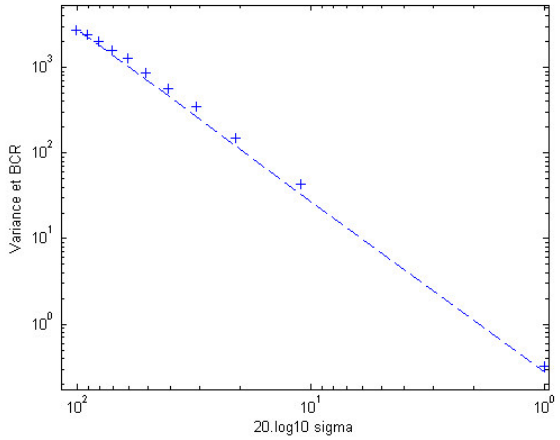


FIG. 2 : BCR de $\hat{v}_x$ (-) et variance de $\hat{v}_x$ (+).

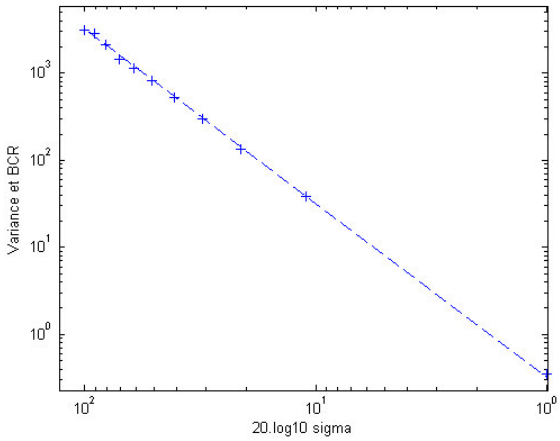


FIG. 3 : BCR de $\hat{y}$ (-) et variance de $\hat{y}$ (+).

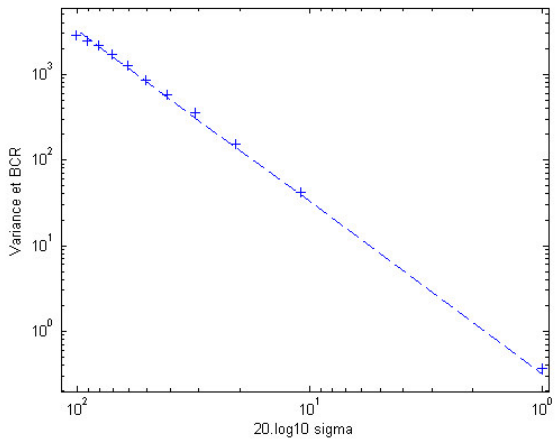


FIG. 4 : BCR de $\hat{v}_y$ (-) et variance de $\hat{v}_y$ (+).

Conception fonctionnelle d'une caméra intelligente pour le traitement vidéo temps réel

Barthélémy HEYRMAN[1], Michel PAINDAVOINE[1], Renaud SCHMIT[2], Laurent LETELLIER[2]

[1]LE2I, UMR CNRS 5158
Université de Bourgogne Aile des Sciences de l'Ingenieur BP 47870 - 21078 Dijon Cedex, France

[2]CEA/DRT-LIST/DTSI/SARC/LCEI,
CEA-SACLAY, Bat 528, F-91191 Gif-sur-Yvette Cedex

barthelemy.heyrman@u-bourgogne.fr, paindav@u-bourgogne.fr
renaud.schmit@cea.fr, laurent.letellier@cea.fr

Résumé – La vision assistée par odinateur occupe un rôle de plus en plus important dans notre société, comme dans la sécurité des biens et des personnes, la production industrielle, les télécommunications, la robotique, ... Cependant, les développements techniques sont encore timides et ralentis par des facteurs aussi différents que le coût des capteurs, le manque de flexibilité, la difficulté à développer rapidement des applications complexes et robustes ou encore la faiblesse de ces systemes à interagir entre eux ou avec leur environnement. Ce papier présente notre concept de caméra intelligente dotée de capacités de traitement video temps réel. Un capteur CMOS, un processeur et une unité reconfigurable sont associés sur le même chip pour offrir flexibilité et hautes performances.

Abstract – Computer-assisted vision plays more and more a part in our society, in various fields such as personal and goods safety, industrial production, telecommunications, robotics... However, technical developments are still rare and slowed down by various factors linked to sensor cost, lack of system flexibility, difficulty of rapidly developing complex and robust applications, and lack of interaction among these systems themselves, or with their environment.This paper describes our proposal for a smart camera with real-time video processing capabilities. A CMOS sensor, processor and, reconfigurable unit associated in the same chip will allow scalability, flexibility, and high performance.

1 Introduction

Grace aux progrès de la technologie, il est possible aujourd'hui de concevoir des caméras qui intègrent des capacités de traitement temps réel de façon à fournir non seulement la vidéo mais aussi les informations pertinentes de la scène. Essentiellement, deux approches de caméras intelligentes ont été proposées dans la littérature : les rétines et des circuits plus génériques. Les rétines intègrent au niveau du pixel des traitements bas niveau. Ces éléments sont souvents constitués de quelques dizaines de transistors par pixel permettant de réaliser différents traitements tels que :

- l'amélioration d'image, du filtrage [7] [2]
- la détection [11] ou de l'estimation de mouvement
- des opérations morphologiques et logiques [8]

Pour autoriser plus de souplesse, une autre solution consiste à coupler un capteur avec un processeur généraliste. Cela permet l'exécution d'applications plus complexes. Dans [5], chaque pixel contient une ALU bit série, une mémoire RAM de 24 bits et des entrées/sorties sur 8 bits. Ce circuit réalise une détection de contours en $5.6\ \mu s$ et un lissage en $7.7\ \mu s$. L'architecture SCAMP [3] est un réseau de type mesh d'APE (Analog Processing Unit). Un capteur de 21×21 SCAMP a été réalisé et permet d'effectuer un lissage en $5.6\ \mu s$ et une détection de contours en $11.6 \mu s$. Le capteur présenté dans [4] est plus particulierement conçu pour des opérations sur des blocs (DCT) ou des convolutions spatiales en utilisant des noyaux de taille 8×8. D'autres approches, comme celle présentée dans [9], utilisent des réseaux de neurones cellulaires. Ce capteur de taille 128×128 est capable de réaliser des convolutions 3×3, des opérations arithmétiques et booléennes.

Nous avons vu dans cette section que les travaux en ce qui concerne les capteurs CMOS et leurs possibilités sont nombreux. Cependant, les solutions présentées (Rétines et capteurs dotés de possibilités de traitements plus génériques) semblent avoir des difficultés à répondre efficacement à tous les besoins des applications de vision. Les rétines offrent l'avantage de la compacité et sont très efficaces pour les traitements bas niveau, mais pour des raisons de place limitée, leurs fonctionnalités sont souvent pauvres et figées. De leur coté, les capteurs associés à des éléments de calculs programmables apportent plus de souplesse mais leurs performances sont moindres et ils sont moins compacts. Nous présentons, dans la seconde partie de cette article, l'architecture que nous avons proposée afin de répondre aux problèmes soulevés dans l'introduction. La troisième partie présente les applications portées sur ce système et donne les résultats en terme de performances. La conclusion présente les travaux à venir d'un point de vue "outils" et "architecture".

2 Architecture proposée

Comme nous l'avons vu précédemment, les deux approches présentées ne sont pas satisfaisantes pour des applications de vision. En effet, les rétines, même si elles offrent des possibilités de traitements proches du pixel, ne peuvent pas être utilisées seules pour des applications complexes et les performan-

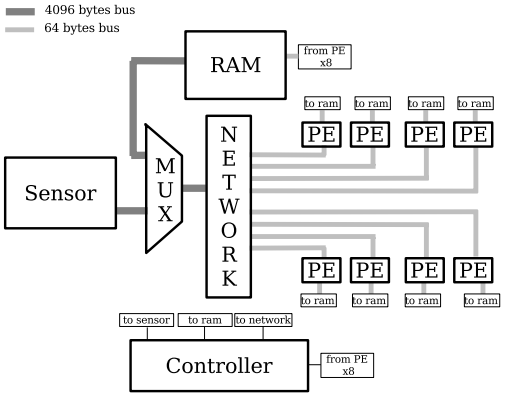


FIG. 1 – Synoptique de la caméra intelligente

ces de processeurs généralistes sont souvent insuffisantes.
La seconde remarque que nous faisons concerne les sorties des capteurs ou rétines. Même si ceux-ci réalisent leurs traitements de manière parallèle, les données pré-traitées sont souvent envoyées de manière séquentielle. Ceci constitue un important goulot d'étranglement dans le chemin de données et réduit considérablement la bande passante.
La troisième remarque s'appuie sur le faible nombre de pixels pertinents dans une image à traiter. La plupart du temps, seules de petites régions d'intérets (quelques dizaines de pixels) sont utiles comme dans le cas du décodage de "data matrix" ou de "détection d'amers".
Pour répondre à toutes ces questions, nous avons choisi de concevoir un capteur avec une sortie massivement parallèle capable de fournir des régions d'une image. Il est couplé à une RAM et un réseau de processeurs capables, respectivement, de stocker/traiter, les données émises. La figure 1 donne le synoptique de ce système.

2.1 Le capteur à sortie massivement parallèle

Le capteur est basé sur une technologie CMOS qui autorise un accès aléatoire aux pixels. Une autre caractéristique plus importante est la possibilité d'accéder à une sous-région de l'image en une seule fois. Pour répondre à ceci, le capteur est conçu pour fournir 4096 pixels en parallèle. En effet, l'évolution des techniques submicroniques permet de réduire la taille des transistors et des liens sur les couches de métal. On peut ainsi disposer en sorties d'un capteur avec beaucoup plus de fils puisqu'on reste dans le même composant. Ainsi, en technologie 90nm, la largeur d'un lien en métal 2 fait $0.30\ \mu m$ et l'écartement entre deux liens est de $0.28\ \mu m$. Il faut donc une longueur d'environ 20mm de silicium pour sortir 4096×8 fils soit un carré de $25mm^2$. Le circuit de lecture autorise en sortie des fenêtres carrées jusqu'à 64×64 pixels. Un autre avantage, de cette sortie en parallèle, est le fonctionnement réduit en fréquence des convertisseurs A/D. Ce concept est aussi présenté dans [10]. Les auteurs présentent un capteur capable de fournir en sortie des ROIs (région d'intéret) et qui permet par ailleurs de réaliser des trackings. Pour limiter la taille de la logique de commande, le déplacement des fenêtres se fait par pas de 32 pixels. Les pixels sont lus 4 par 4.

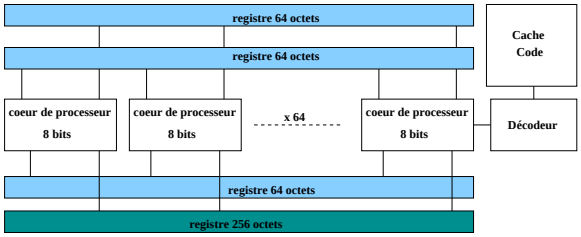


FIG. 2 – Schéma d'un processeur élémentaire

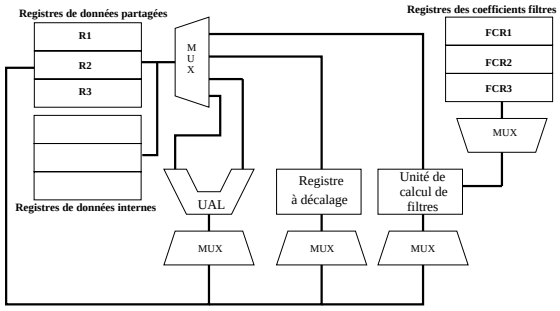


FIG. 3 – Structure détaillée d'un des cœurs 8 bits d'un PE

2.2 Description d'un PE

Ces PEs sont conçus pour exécuter aussi efficacement que possible des opérations simples que l'on retrouve souvent dans des applications de vision. En effet, ces algorithmes ne se concentrent généralement que sur une ou plusieurs petites parties de l'image. Par exemple, la compression JPEG utilise des fenêtres 8×8. C'est pourquoi chaque processeur est capable d'effectuer des opérations sur une fenêtre de 8×8 pixels au maximum. On peut imaginer regrouper des processeurs entre eux pour traiter une plus grande fenêtre. Les fonctionnalités implémentées sont les opérations arithmétiques et logiques, les filtrages par masque, les convolutions et les corrélations mais aussi les calculs d'histogrammes et les seuillages. Une structure de type MAC permet de réaliser efficacement ces traitements. Chaque PE est un groupe de 64 cœurs de processeurs (8 bits en entrée et 16 bits en sortie) qui contiennent des fonctions prédéfinies. Un cœur de processeur contient une ALU qui réalise tous les calculs "classiques" (additions, soustractions, opérations logiques), un registe à décalage qui permet d'effectuer des multiplications/divisions par puissance de deux, et une unité de calcul de filtre qui réalise des opérations de type somme de produits, ou (ET de OU). Les processeurs élémentaires contiennent aussi 3 registres de 64 mots de 16 bits pour le stockage des données entre deux opérations entre pixels, un registre de 256 octets pour les opérations effectuées sur des tables (calcul d'histogramme par exemple), et une mémoire cache pour stocker le code à exécuter. Les applications sont codées en assembleur comme ceci :

- Un mot clé : avr pour moyenne, add pour addition
- l'emplacement de la fenêtre 64×64 que le capteur doit fournir
- la banque mémoire à accéder
- 4 nombres donnant la configuration du réseau

Par exemple :
add r0 m0 0 0 8 8 // addition
hst // histogramme du résultat précédent

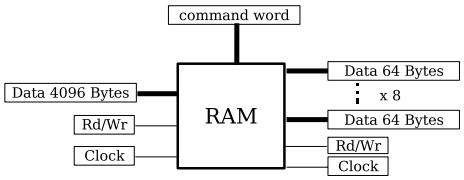


FIG. 4 – Schéma de la RAM

thr 45 // seuillage en fonction de la valeur donnée
va réaliser la somme entre la fenêtre en haut à gauche du capteur et le coin en haut à gauche d'une image stockée en mémoire. Certaines instructions ne requièrent pas de paramètres comme **hst** ou un nombre réduit comme **thr**.

2.3 Choix du réseau

C'est une partie importante de notre système. En effet, si nous ne sommes pas capables de fournir les données efficacement aux éléments de calculs, c'est à dire suffisamment rapidement et en quantité, l'architecture souffrira de pertes de performances importantes. Les réseaux de type cross-bar sont très efficaces pour dispatcher les données mais leur coût en silicium devient vite rédhibitoire quand le nombre d'entrées/sorties grandit. Nous avons choisi un cross-bar partiel qui prend en entrée 4096 pixels et ne fournit en sortie que 512 pixels. Ces derniers sont en fait 8 fenêtres de 8×8 pixels prises n'importe où dans la fenêtre d'entrée.

2.4 Conception de la RAM

Nous avons choisi pour notre système une RAM double port. En effet, elle permet de lire et d'écrire des données simultanément. La capacité de la mémoire est de 128×64 octets. L'un des ports de 4096 octets est relié au capteur ou au réseau via le multiplexeur. Ceci permet de stocker une fenêtre en un coup d'horloge. Son fonctionnement est classique : une horloge, un bus d'adresse et un signal de lecture/écriture. Le second port est différent par son adressage. Il est large de 8×64 octets et il est relié à chaque PE. Il possède une horloge, un signal de lecture/écriture et un mot de commande. Un schéma de cette mémoire est donné figure 4.

2.5 Le contrôleur

Son rôle est de superviser le fonctionnement du système. Il génère les signaux de commandes en fonction des requêtes des PEs. Il gère aussi les acquisitions d'images et leur stockage en mémoire. Il envoie les signaux de démarrage pour les traitements et gère les accès à la mémoire. Il peut être comparé à un contrôleur d'interruption et ne réalise aucun traitement. Ses réactions dépendent de l'application implantée. Une description plus précise de son fonctionnement est donnée dans la section 3.

3 Modélisation et validation de l'architecture

Dans le but d'étudier notre concept, nous avons décidé de le modéliser en SystemC. Les temps de simulation sont grandement améliorés par rapport à une description VHDL grâce à cette librairie C++. Un autre avantage est le haut niveau de description possible. Nous avons fait dans un premier temps un modèle fonctionnel software qui sera par la suite raffiné en utilisant la co-simulation SystemC-VHDL. Afin de valider cette architecture, nous l'avons testée sur trois applications de traitements d'images. Nous présentons rapidement ces application dans une première partie, puis la seconde partie présente les résultats obtenus.

3.1 Les applications implantées sur cette architecture

La première application utilisée pour valider le modèle, détecte le mouvement dans une image par soustraction. Pour cela, une moyenne sur 4 images est réalisée pour réduire le bruit, on soustrait ensuite l'image de référence. Le résultat est seuillé et des opérations morphologiques sont réalisées pour éliminer le bruit résiduel. Le nombre d'objets est comptabilisé et si il est différent du nombre qui apparaît dans l'image de référence, cette image est mise à jour.

La seconde application concerne la localisation et la reconnaissance de visages dans une image. L'algorithme utilisé repose sur les réseaux de neurones RBF. Il s'agit tout d'abord d'extraire de l'image un vecteur de données (moyenne glissante de 4 pixels) que l'on présente au réseau de neurone. Nous avons remplacé l'utilisation d'une fonction gaussienne dans les neurones par une fonction "porte". Cette approximation donne des résultats convenables [6] et permet d'accélérer très efficacement les traitements.

La troisième application réalise une stabilisation d'image [1]. Pour ce faire, des points caractéristiques sont définis dans une première image et recherchés dans une seconde par corrélation. La variations des coordonnées des points retrouvés dans la seconde image permet de retrouver le mouvement. L'application du "mouvement inverse" stabilise les images. Nous n'avons implantés que les deux premières étapes de ce traitement.

3.2 Résulats

Le premier résultat que nous obtenons est l'accélération du stockage des données en mémoire qui se fait par paquets de 4096 pixels. Cela signifie qu'avec une horloge de $10\ ns$ pour le système et un capteur $1K \times 1K$, 256 cycles d'horloges sont nécessaires pour stocker l'image complète et la bande passante théorique maximum entre le capteur et la mémoire est $1K \times 1K/2.56\ \mu s = 409.6\ Goctets/s$. Du point de vue des opérateurs, chacun requiert 3 cycles d'horloge pour traiter les données dans le pire cas (des accès mémoires sont nécessaires pour chaque instruction). Par exemple, une application constituée de 5 tâches différentes requiert 15 cycles d'horloge pour traiter 8×64 pixels. Une image de $1K \times 1K$ est traitée en : $((1K \times 1K)/(8 \times 64)) \times 15 = 30720$ cycles. Si nous choisissons une horloge de $10\ ns$, il faut environ 0.3ms pour effectuer l'opération. Dans ce cas plus de 3000 images peuvent être traitées en une seconde.

Le tableau 1 rassemble les résultats obtenus pour notre architecture. Nous constatons que les performances atteintes sont importantes mais nous pouvons noter que le système est plus efficace sur des traitements correspondant à ses caractéristiques

(fenêtres 64×64, filtrages 8×8) qui exploitent pleinement les ressources disponibles.

TAB. 1 – Performances obtenues

Application	temps (ms)	Accélération	Efficacité
Détection de mouvement	0,52	7,63	0,95
Reconnaissance de visage	2	7,7	0,96
Stabilisation d'image	0,3	7,83	0,98

Le tableau 2 donne un récapitulatif des performances obtenues par notre architecture comparées avec celles d'autres architectures. L'architecture ARM est basée sur un ARM922T qui fonctionne à 100 MHz. Le processeur Athlon est cadencé à 1,43 GHz.

TAB. 2 – Récapitulatif comparatif des performances obtenues (en ms)

Application	Notre archi.	ARM	Athlon	ZISC
Détection de mouvement	0,52	76,9	8	-
Reconnaissance de visage	2	-	-	16
Stabilisation d'image	0,3	-	26,3	-

4 Conclusion et perspectives

Nous avons décrit, dans cet article, une nouvelle architecture de caméra intelligente offrant des possibilités de traitements très rapides. Pour réaliser ceci, un capteur à sorties massivement parallèles qui a la possibilité de lire des régions d'intéret a été proposé. Un réseau de processeurs prenant en compte les spécificités des applications de vision et capable d'effectuer des opérations comme les filtrages par masque ou les convolutions est également proposé. Un modèle fonctionnel a de plus été réalisé pour valider ce concept.
Des travaux futurs vont concerner l'implantation de l'architecture sur silicium. Des améliorations au niveau des processeurs peuvent être apportées notamment en utilisant de la logique reconfigurable afin de rendre l'implantation d'opérateurs spécifiques plus simple. De même, des outils de programmation haut-niveau et automatiques dédiés à l'architecture pour en simplifier l'utilisation sont nécessaires et vont être étudiés.

Références

[1] L. Damez, J.P. Drutin, F. Diaz, and N. Allezard. Stabilisation 3d temps rel d'images : aspects algorithmiques et implantation parallle. In *Journes Francophones sur l'Adquation Algorithme Architecture - JFAAA'05*, pages 203–208, 2005.

[2] T. Delbrück and C.A. Mead. *Analog VLSI phototransduction by continuous-time, adaptive, logarithmic photoreceptor circuits*, pages 139–161. IEEE Computer Society Press, 1995.

[3] P. Dudek and P. J. Hicks. A general-purpose cmos vision chip with a processor-per-pixel simd array. In *Proc. ESSIRC2001*, 2001.

[4] A. Graupner, J. Schreiter, S. Getzlaff, and R. Schüffny. Cmos image sensor with mixed-signal processor array. *IEEE Journal of Solid-State Circuits*, 38(6) :948–957, June 2003.

[5] M. Ishikawa, K. Ogawa, T. Komuro, and I. Ishii. A cmos vision chip with simd processing element array for 1ms image processing. In *Proc. Conf. ISSCC'99*, number TP2.2, 1999.

[6] N. Malasn. *Localisation et reconnaissance de visages en temps rel : algorithmes et architectures*. PhD thesis, Thse de l'universit de Bourgogne, 2002.

[7] Y. Ni, F. Devos, M. Boujrad, and J.H. Guan. A histogram equalization based adaptive image sensor for real-time vision. *IEEE J. of Solid State Circuits*, 32(7) :1027–1036, July 1997.

[8] F. Paillet, D. Mercier, and T. Bernard. Making the most of 15k lambda2 silicon area for a digital retina. In Proc. SPIE, editor, *Advanced Focal Plane Arrays and Elactronic Cameras*, volume 3410, 1998.

[9] A. Rodrigues-Vazquez, G. Linan-Cembrano, L. Carranza, E. Roca-Moreno, R. Carmona-Galan, F. Jimenez-Garrido, R. Domiguez-Castro, and S. Espejo Meana. Ace16k : The third generation of mixed-signal simd-cnn ace chips toward vsocs. *IEEE Transactions on Circuits ans Systems*, 51(5) :851–863, 2004.

[10] O. Schrey, J. Huppertz, G. Filimonovic, A. Bussmann, W. Brockherde, and B.J. Hosticka. A 1kx1k high dynamic range cmos image sensor with on-chip programmable region of interest readout. In *Proc. ESSCIRC2001*, 2001.

[11] S. Shigematsu, H. Morimura, Y. Tanabe, T. Adachi, and Machida K. A single-chip fingerprint sensor and identifier. *IEEE Journal of Solid-State Circuits*, 34(12) :1852–1859, December 1999.

Plate-forme de Prototypage et Conception pour les systèmes de Radio-Logicielle

FABIENNE NOUVEL, ARNAUD MASSIANI, PIERRE DE STEFANO
IETR/INSA de Rennes, Institut Electronique et Télécommunications de Rennes, 20 Avenue des buttes de Coësmes, 35043 Rennes,
fabienne.nouvel@insa-rennes.fr

Résumé – Cet article présente la plate-forme radio logicielle utilisée dans le projet PALMYRE, élaboré dans le cadre du Contrat de Plan Etat Région (CPER) pour une durée de 5 ans. La plate-forme hétérogène et commune à plusieurs laboratoires bretons va ainsi permettre le développement et l'évaluation de systèmes radioélectriques pour les futures générations. Pour faciliter le portage de multiples applications sur cette plate-forme, une méthodologie de conception est mise en place, offrant ainsi un couplage algorithme-architecture maximal.

Abstract *– In this paper, the architecture demonstrator for single carrier and multi-carrier radio system will be presented. This development is performed in the scope of the project PALMYRE together with 3 universities laboratories research : IETR, ENST and LESTER. The main objective of this project is the demonstration of a multi-standard system on a single prototyping board, including DSP and FPGA.*

1. Introduction : Besoins et objectifs d'une plate-forme radio logicielle

Les applications de radiocommunication connaissent une véritable explosion. L'émergence des futures générations de systèmes, notamment la 4^{ième} génération, se présente comme une réponse à l'accroissement permanent de la diversité des standards et des applications demandées et supportées. Cette recherche accompagne la notion de Radio Logicielle. Ce terme désigne un système flexible, multi-standards, reconfigurables dont le traitement numérique est plus en plus prépondérant et à proximité de l'antenne. La réalisation de tels systèmes reste pour l'heure trop optimiste, au vue des contraintes technologiques actuelles, et conduit à la définition de systèmes qualifiés de Radio Logicielle Restreinte (SDR). La SDR repose sur une transposition en fréquence intermédiaire avant le traitement numérique. Ceci nécessite une évolution des structures classiques dédiées de réalisation qui conduit à l'utilisation de plate-formes non dédiées dites de prototypage rapide, ouvertes, permettant d'accueillir plusieurs applications ou standards de communications. Dans la première partie de l'article, les critères de choix de la plate-forme Radio Logicielle sont présentés. Cette plate-forme matérielle doit être accompagnée d'une plate-forme logicielle, guidant le concepteur dans l'optimisation de l'utilisation de celle-ci en fonction des algorithmes à implanter. C'est l'objet du deuxième paragraphe. Des résultats d'implémentation sur la plate-forme retenue pour PALMYRE sont exposés dans le troisième volet.

2. Définition de la plate-forme. Critères de choix

2.1 Critères de définition de la plate-forme.

Le développement d'architectures hétérogènes et reconfigurables, basé sur une utilisation conjointe de processeurs de signaux (DSP) et composants programmables (FPGA), offre une réponse aux contraintes de la SDR. Par ailleurs, les applications de radiocommunications ont pour caractéristiques d'êtres régulières (flot de données) et intensives en calcul, nécessitant du parallélisme dans les exécutions. Ces contraintes nous conduisent à la définition de plate-formes dites de prototypage rapide. Facilité le développement et la disponibilité anticipée d'un démonstrateur pouvant conduire à un Soc sont deux éléments clef de ces plate-formes. Le prototype doit être capable de fonctionner dans des conditions proches du système final, sans avoir recours à un circuit spécifique. Les bénéfices d'une telle approche sont nombreux : vérification en environnement réel, preuves d'un concept, mesures et tests rapides,…

On peut lister un certain nombre de critères qui vont orientés le choix de la plate-forme :

- l'utilisation de composants non spécifiques : elle doit accueillir à la fois des DSP et des FPGA. Les DSP disposent d'une organisation adaptée au traitement du signal. Les FPGA permettent de supporter la diversité des traitements et les contraintes de temps réel. Cette association forme une architecture hétérogène,

- le couplage entre les composants : le parallélisme effectif de la plate-forme nécessite une synchronisation entre les cibles. Le réseau de communication ne doit pas pénaliser les performances globales du système. Il doit aussi être flexible et permettre différentes configurations matérielles. Cette flexibilité se traduit par des bus au nombre variable, de taille adaptée aux flux de données, …

- la politique d'ordonnancement et la distribution des différentes opérations sur les cibles. Une politique dynamique, impose une gestion dynamique des ressources qui peut rendre la prédiction des performances aléatoire. A l'inverse, un ordonnancement statique, nécessitant une définition préalable de l'exécution, permettra un dimensionnement adéquat de l'application, des ressources de calcul et de communication.

En plus de cette hétérogénéité des composants numériques, la plate-forme doit offrir une interface

analogique au travers de convertisseurs numériques-analogiques et inversement. Ils doivent communiquer de façon identiques avec les modules numériques (bus). Ces convertisseurs vont permettre par la suite de connecter le traitement bande de base (BB) aux modules radio fréquence (RF) analogique, travaillant dans les bandes de fréquences 2,4GHz à 5GHz. Cette transposition BB en bande RF nécessite un étage intermédiaire de passage en fréquence intermédiaire (FI). La combinaison composants numériques et composants de conversion constitue alors la plate-forme de prototypage que l'on peut qualifier de 'mécano-électronique'.

Plusieurs plate-formes ont été proposées dans le cadre de projets, tels MATRICE et 4MORE[1], RICE [2], PLATON [3], Picard [4]. Parmi ces systèmes, seul le système Picard est constitué d'une carte numérique au format PCI, qui s'interface avec une carte RF. Par ailleurs, ils sont très orientés vers un standard de télécommunication (2G, 3G). Ces plate-formes ne répondent donc pas à la contrainte de la radio logicielle, dont l'objectif est d'offrir plusieurs standards sur une même plate-forme. Par ailleurs, peu d'entre elles proposent une interface radio-fréquence analogique.

2.2 Présentation de la plate-forme retenue

Le marché des équipements de prototypage, bien que vaste, ne répond pas toujours à nos contraintes : mono-composant, outil dédié, système non modulaire,… Dans [3], l'auteur dresse un état des nombreuses plate-formes disponibles, à base de DSP, de FPGA ou de composants dédiés. Pour déterminer la ou les sociétés fournisseurs, plusieurs comparaisons ont été menées selon différents critères précisés précédemment. Le choix s'est finalement arrêté avec la solution Sundance [5] qui répond bien aux critères énoncés.

La plate-forme repose sur l'utilisation d'une carte mère compatible au format PCI, qui comporte 4 emplacements au format TIM-40 recevant les cartes filles. Les modules numériques (DSP/FPGA) mono ou multi-composants ou de conversion (CNA/CAN) échangent leurs données selon 2 types de média principaux : les communications selon les ports C4x (CP) et les interfaces propriétaires Sundance SDB (Sundance Data Bus). Ces communications point à point sont toutes bidirectionnelles, synchrones ou asynchrones. Les CP ont des débits maximum de 20Mo/s alors que les SDB offrent des débits de 200Mo/s à 100MHz. Différentes configurations de connexions sont possibles, sachant que l'on dispose de 6 CP et de 2 à 4 SDB par modules. La Figure 1 est un exemple de configuration de l'émetteur utilisé dans le projet PALMYRE pour les modem mono et multi-porteuses.

Dans le cadre du projet, les modules DSP sont uniquement des DSP C6x de Texas Instrument. Plusieurs composants sont disponibles sur les cartes TIM : du C62 à 200 MHz au C64 à 600 MHz, le choix du module étant effectué selon les performances visées. Pour les FPGA, il s'agit des composants des familles VirtexE et VirtexII de Xilinx. Quelque soit le module DSP et le module FPGA, les communications utilisent les CP ou SDB.

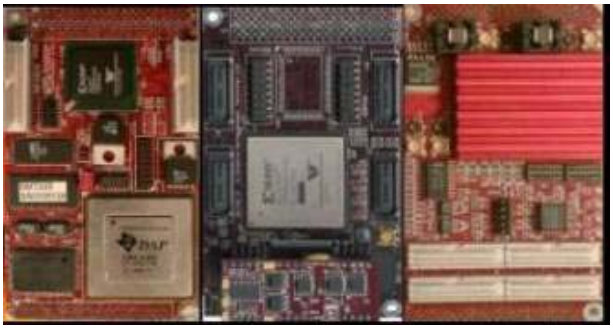


FIG. 1 : Carte Sundance avec 3 modules TIM : DSP C6x , FPGA VirtexII et CNA

En ce qui concerne la partie analogique, celle-ci est développée plus spécifiquement pour des applications Indoor.

Disposant d'une entrée en fréquence intermédiaire allant de 20 à 50 MHz, avec un niveau de –19 dBm, le signal est alors transposé dans les bandes 5.15 à 5.35 GHz, et assure une compatibilité ave les standards IEEE 802.11. Le signal est transmis avec une puissance de 3 à 23 dBm. En réception, la tête d'entrée dispose d'une sensibilité de –85 dBm, et réalise la transposition en fréquence intermédiaire avec un contrôle de gain de 65 dB. Par ailleurs, le concept de ce module RF permet une extension pour des applications MIMO (Multiple Input, Multiple Output).

L'utilisation d'une telle plate-forme 'matérielle' doit être accompagnée d'une plate-forme 'logicielle'. En effet, l'objectif de la plate-forme matérielle est de faciliter la validation et les tests d'applications et non la connaissance précise de l'architecture, notamment les moyens de communications entre modules. C'est l'objet du paragraphe suivant.

3. Méthodologies de conception

3.1 Présentation des différentes approches

L'hétérogénéité de l'architecture se retrouve dans l'hétérogénéité des applications visées : formes d'ondes, mono ou multi-porteuses, bande étroite ou large bande, …La mise en place de méthodologies de conception doit permettre de répondre aux points suivants : faciliter l'exploration architecturale et la ré-utilisation de développements existants, limiter le coût de développement et réunir les compétences des différents partenaires. Dans le cadre du projet PALMYRE, plusieurs méthodologies complémentaires sont proposées pour répondre à ces contraintes :

- L'approche MCSE/Cofluent [6] est une démarche globale qui s'articule autour de 4 étapes : spécifications, conception fonctionnelle, architecturale et implémentation. Les algorithmes et les architectures sont modélisées sous forme de graphes illustrés par les Figure 2.a et 2b. Les différentes fonctions sont testées tout en prenant en compte les caractéristiques des cibles. Par contre, la distribution des fonctions sur les cibles est manuelle et guidée par le concepteur,

- L'approche AAA/SynDEx [7] propose une démarche qui permet de réaliser automatiquement la distribution des fonctions sur l'architecture ainsi que l'ordonnancement de celles-ci. Les fonctions et cibles y sont modélisées.

- l'approche par l'outil de synthèse GAUT [3] : GAUT permet de réaliser la synthèse des unités de calcul et de communication pour des cibles FPGA. Partant d'une description comportementale, GAUT génère les différentes entités en y incluant les interfaces. Par contre, GAUT ne gère pas le partitionnement, ni l'adéquation algorithme-architecture qui doit être réalisée manuellement.

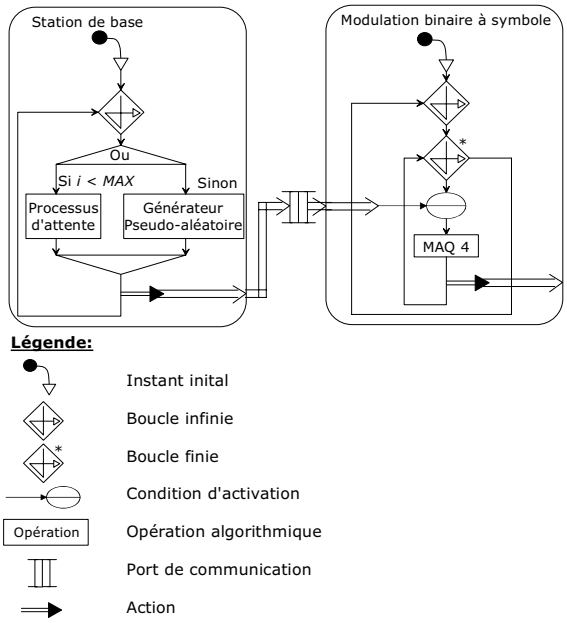


FIG 2.a Modélisation des algorithmes sous MCSE

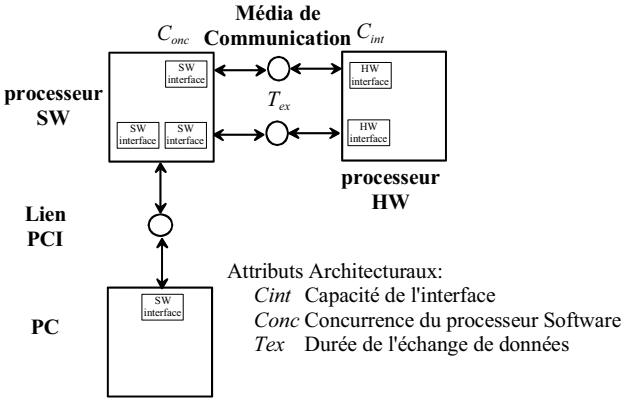


FIG 2.b Modélisation de l'architecture de la plate-forme Sundance sous MCSE

Ces trois méthodes ne sont pas concurrentes mais complémentaires. Elles n'interviennent que dans la phase conception numérique. La spécification fonctionnelle est décrite avec MCSE, aboutissant à la définition des modules fonctionnels et de leurs interfaces. Puis, l'approche AAA avec SynDEx permet une répartition optimale des algorithmes sur l'architecture cible. A ce niveau, il n'est pas nécessaire de disposer des algorithmes car SynDEx ne simule pas le système mais propose une distribution définie à partir des temps d'exécution de chaque module estimés sur chaque cible logicielle (DSP) ou matérielle (FPGA). Après l'adéquation, SynDEx génère un macro code séquentiel pour chaque cible, qui doit être traduit en C si la cible est un DSP ou VHDL s'il s'agit d'un FPGA. Partant de cette distribution et de l'ordonnancement, il est possible alors de développer les algorithmes, dans l'environnement MCSE. L'application peut alors être entièrement simulée, tout en tenant compte de l'architecture, et plus particulièrement des interfaces entre les différents composants (DSP/FPGA) : mode d'échanges, taille, évènements. Une fois le système validé fonctionnellement et matériellement, l'outil GAUT nous donne alors la possibilité de transcrire les algorithmes en langage VHDL pour les cibles FPGA. Cette traduction prend en compte aussi la génération des interfaces de communication et les synchronisations avec les unités de calcul. La Figure 3 illustre le flot ainsi mis en œuvre.

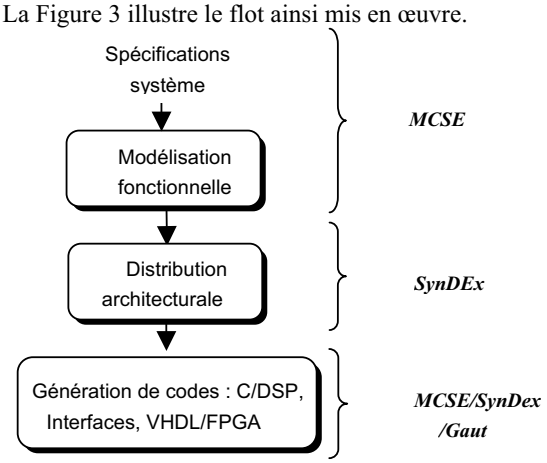


FIG 3. Flot de conception.

Dans le cadre du projet PALMYRE, l'équipe IETR de Rennes a travaillé sur les deux premiers outils MCSE et SynDEx [6] [7]. L'approche GAUT a été étudiée par le LESTER et fait l'objet des travaux présentés dans [3]. SynDEx et MCSE nous permettent alors de décrire le système avec une approche descendante, qui prend en compte à la fois les paramètres des algorithmes et ceux de l'architecture. Ce couplage évite la trop fréquente rupture entre les équipes de développement software et hardware.

4. Mise en œuvre de la plate-forme

4.1 Application mono-porteuse

La première application portée sur cette plate-forme est une chaîne de transmission mono-porteuse, offrant un débit utile de 18 Mbits/s. Elle met en œuvre 2 cartes Sundance dotées chacune d'elles d'un DSP C62, d'un FPGA Virtex XC2V1000 (1M de portes) et d'un couple CNA/CAN. La transposition en fréquence intermédiaire à 45 MHz est effectuée en numérique avant la phase conversion numérique-analogique, évitant dès lors l'utilisation de 2 convertisseurs à l'émission. En réception, le traitement symétrique est effectué. Disposant d'un rapport signal à bruit

de 11dB, un taux d'erreur binaire dans la plage de 1.10^{-6} à 1.10^{-4} est obtenu avec un rapport signal à bruit de 11 dB, selon les conditions initiales. Cette première application mono-porteuse a permis de valider les solutions architecturales retenues ainsi que les bibliothèques de communications CP et SDB. Ces résultats sont présentés dans [8].

4.2 Application multi-porteuses

La seconde application est un modem multi-porteuses large bande à étalement de spectre (MC-CDMA) [7]. Il s'appuie sur la même plate-forme matérielle et offre un débit de 30 Mbits/s à 60 Mbits/s avec une bande de transmission de 20 MHz. Celle-ci est découpée en 256 sous-bandes avec 192 utiles, le signal occupe alors une bande de 15MHz. La Figure 4 donne le synoptique général du modem MC-CDMA. Les paramètres ont été choisis de façon à se situer au delà de la norme Hiperlan utilisée dans les systèmes indoor. Après une modulation QPSK ou 16 QAM, le signal est étalé en fréquence avec un facteur d'étalement maximal de 32, puis transmis sur les 192 porteuses utiles. La modulation multi-porteuses utilise une IFFT de 256 points, basée sur une IP (Intellectual Property) de Xilinx [9]. Le débit obtenu avec une fréquence système de 20 MHz varie de 30Mbit/s à 60 Mbit/s. La transposition en FI à 10 MHz est effectuée aussi numériquement selon un schéma similaire à l'application mono-porteuse. La trame est alors constituée d'un symbole d'estimation puis de plusieurs symboles utilisateurs. La structure de la trame est totalement configurable (étalement, taille IFFT/FFT, modulation) offrant ainsi une adéquation selon l'environnement de communications. En réception, le symbole d'estimation permet de calculer les coefficients d'égalisation, qui restent constants sur l'ensemble de la trame. Le critère du ZF (Zero Forcing) [7] est appliqué, pour ses bonnes performances et sa simplicité. Pour la démodulation, les opérations inverses de celles de l'émetteur sont effectuées, comme indiqué sur la Figure 4.

Les émetteurs et récepteurs sont intégrés sur 2 modules FPGA constitués chacun d'un composant Virtex XC2V2000 (2 M de portes). Ces FPGA sont utilisés respectivement à 60% et 65% en émission et réception. Les modules FPGA communiquent également avec les DSP C62 de Texas par des liens SDB. Les DSP permettent de générer et récupérer le flux de données. Ils sont interfacés avec le poste PC par le lien PCI de la carte support Sundance.

4.3 Apport des méthodologies SynDEx et MCSE

L'outil SynDEx a été mise en œuvre lors de l'étape de partitionnement. Il est également utilisé pour la génération automatique du code pour les interfaces de communications entre modules. Le développement du modem multi-porteuse est réalisé dans l'environnement MCSE. Ainsi, l'ensemble des algorithmes ont pu être testés avant implémentation. Nous avons pu mettre en évidence les latences introduites par les liens, ainsi que les risques de rupture de flux en cas d'utilisation des liens de type CP. L'outil MCSE nous a par ailleurs permis de dimensionner de façon optimale les données manipulées sur le FPGA, tout en facilitant le passage virgule flottante - virgule fixe.

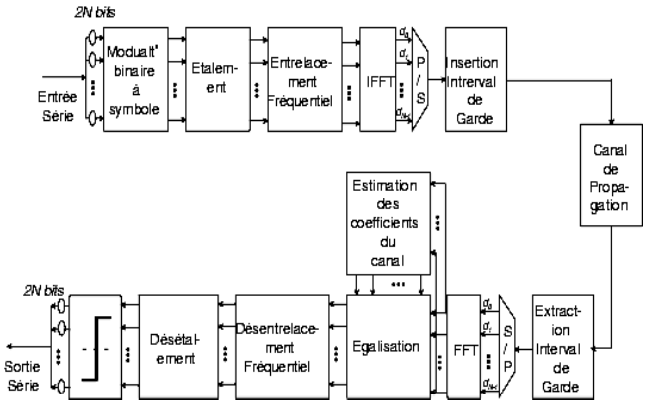


FIG 4 Synoptique du modem MC-CDMA

5. Conclusion

Plusieurs travaux sont actuellement en cours, visant à améliorer les performances du système et son intégration. La plate-forme de prototypage proposée dans le cadre du projet PALMYRE permet ainsi de supporter plusieurs modems. Des extensions vers d'autres standards sont envisageables, la seule contrainte étant les ressources des composants. Par ailleurs, l'approche méthodologique a permis de concevoir le modem de façon rigoureuse et optimisée.

Remerciements

Ce projet PALMYRE est supporté par le Ministère de L'industrie et de la Recherche, la région Bretagne, les collectivités locales et conseils généraux.

Références

[1]Site du Projet Européen IST *www.ist.matrice.org.*

[2] P. Murphy, *An FPGA Based Rapid Prototyping Platform for MIMO systems*, Asilomar Conference on signals system and Computers, Novembre 2003

[3] Bomel P., *Plate-forme de prototypage rapide fondée sur la synthèse de haut niveau pour applications de radiocommunications* , Thèse de Doctorat de l'Université de Bretagne Sud, N° 45, Décembre 2004.

[4] Picard Hardware Concept for WLAN applications, LABORATOIRE IMEC, www.imec.be, Belgique.

[5] SUNDANCE CORPORATION, Adresse web de la société *www.sundance.com*, Royaume Uni

[6] A. Massiani, F.Nouvel, Le Nours S., *Méthodologies de Conception appliquée aux systèmes de Radio Logicielle* , Conférence JFAAA 2005, 18-21 Janvier 2005, Dijon

[7] LE NOURS S., *Etude, optimisation et implémentation de systèmes MC-CDMA sur architectures hétérogènes,* Thèse de Doctorat de l'INSA de Rennes, Décembre 2003.

[8] PROJET REGION BRETAGNE CPER PALMYRE - Etat d'avancement – Juin 2005.

[9] Site de la société Xilinx *www.xilinx.com*. Rubrique FFT VirtexII

Synthèse Comportementale Sous Contraintes de Communication et de Placement Mémoire pour les composants du TDSI

Gwenolé Corre, Philippe Coussy, Pierre Bomel, Eric Senn, Eric Martin

Email : prénom.nom@univ-ubs.fr
LESTER LAB, UBS University, CNRS FRE 2734

Résumé – La conception de systèmes complexes en traitement de l'image et du signal implique de réduire les coûts architecturaux et de maximiser les performances temporelles tout en considérant les contraintes de communication et d'accès mémoire durant la conception et l'intégration d'accélérateurs matériels dédiés. Malheureusement, les blocs matériels utilisés dans les flots de conception semi-automatiques traditionnels n'autorisent pas une flexibilité suffisante pour garantir cet objectif. Dans cet article, nous présentons une méthodologie et un outil qui autorisent la synthèse d'applications en traitement du signal et de l'image sous contraintes de communication et de mémorisation. Basé sur un ensemble de modèles formels, notre outil GAUT aide le concepteur à trouver un compromis entre performance et complexité architecturale.

Abstract *– The design of complex Digital Signal Processing systems implies to minimize architectural cost and to maximize timing performances while taking into account communication and memory accesses constraints for the integration of dedicated hardware accelerator. Unfortunately, the traditional Matlab/ Simulink design flows gather not very flexible hardware blocs. In this paper, we present a methodology and a tool that permit the High-Level Synthesis of DSP applications, under both I/O timing and memory constraints. Based on formal models and a generic architecture, our tool GAUT helps the designer in finding a reasonable trade-off between the circuit's performance and its architectural complexity. The efficiency of our approach is demonstrated on the case study of a FFT algorithm.*

1. Introduction

En raison de la complexité des applications de traitement du signal numérique (TDSI), les concepteurs ont besoin d'un chemin plus direct entre la spécification et l'implémentation. Des flots de conception et les outils de CAO associés sont nécessaires pour maîtriser la complexité des systèmes et réduire les temps de conception. Ceci a mené au développement d'environnements d'aide à la conception explorant l'espace de solutions. Dans [1], [2] et [3] les auteurs proposent des approches qui utilisent des outils tels Matlab/Simulink/Stateflow [4] pour la spécification de l'application et qui produisent automatiquement une architecture VHDL RTL du système. Basés sur des macro-générateurs qui utilisent les mécanismes "generic"/"generate", la synthèse matérielle de l'application peut être résumée à une instanciation de blocs. Toutefois, bien que ce type de composant puisse être paramétrable, il repose sur un modèle architectural fixé dont la personnalisation reste limitée. Ce manque de flexibilité est particulièrement vrai pour (1) l'interface de communication dont les ordres d'acquisition et de production des données d'entrée/sortie ainsi que leurs caractéristiques temporelles (débit) sont définies et (2) pour l'unité de mémorisation dans laquelle la localisation des données est figée. La synthèse de haut niveau peut être utilisée pour résoudre ce manque de flexibilité. Quelques outils comme SystemC Compiler et Monet introduisent des modèles pour les E/S et la mémoire. Ils restent toutefois limités par la complexité des algorithmes à synthétiser [5], [6] et par la non dissociation de la spécification des contraintes de communication et de la spécification du traitement algorithmique.

Dans le domaine du temps réel et des applications utilisant de grande quantité de données, les ressources de calcul et de mémorisation doivent intégrer l'augmentation du flux de données. La conception d'architecture ou de système doit se concentrer sur (1) la réduction des goulets d'étranglement sur les bus et les tampons d'entrées/sorties, (2) la réduction du coût de stockage des données et (3) le respect de contraintes temporelles strictes en terme de cadence et de latence. La conception doit s'appuyer sur une modélisation des contraintes d'E/S et de mémoires de données internes, une analyse des contraintes pour vérifier l'existante d'une solution architecturale et l'utilisation de la synthèse de haut niveau pour générer l'architecture du composant.

Dans [7] et [8], nous proposons une méthodologie de conception de SoC basée sur la ré-utilisation d'IP algorithmique. Nous utilisons la synthèse de haut niveau sous contrainte d'E/S pour fournir des IP prenant en compte les contraintes d'intégration d'un système complet (cadence d'arrivée des données, technologie, format des bus …). Dans [9], nous introduisons une nouvelle approche prenant en compte l'architecture des mémoires interne et le placement des données lors de la synthèse de haut niveau. ***Dans cet article, nous proposons un flot de conception basé sur des modèles formels qui autorise la synthèse de haut niveau d'applications TDSI sous contrainte de communication et de mémorisation. Le concepteur peut ainsi spécifier les E/S, la distribution et le placement des données en mémoire interne et la cadence comme contraintes pour la synthèse de composants matériels. Cette approche peut être intégrée au flot de conception Matlab/Simulink pour augmenter la flexibilité des blocs matériels.***

Dans la section 2, nous formulons le problème de la synthèse de haut niveau sous contraintes d'E/S et de mémorisation. La section 3 présente les principales étapes de notre approche et les modèles formels qui y sont associés. Dans la section 4, nous démontrons l'intérêt de notre approche avec un exemple de synthèse d'une transformée de Fourier Rapide (FFT).

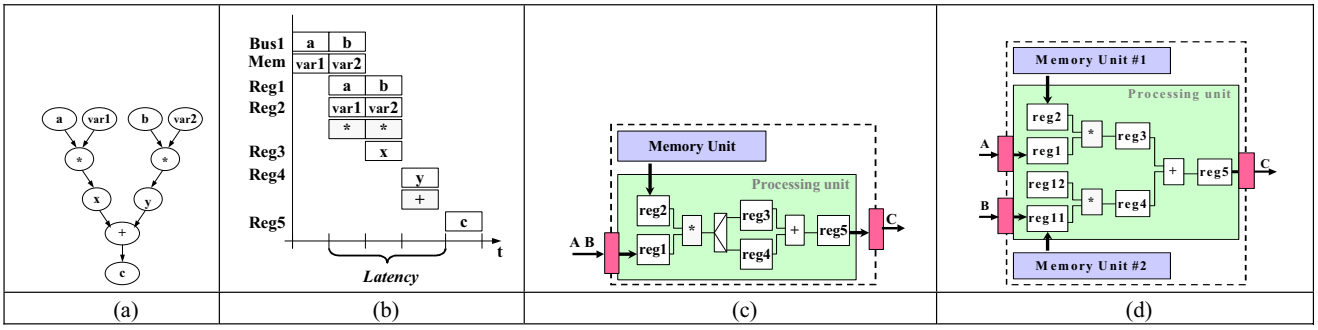


Fig. 1: (a) Signal Flow Graph SFG, (b) Comportement temporelle, (c) Architecture séquentielle, (d) Architecture parallèle

2. Formulation du problème

Considérons un composant matériel basé sur une architecture générique composée de deux unités fonctionnelles principales, une unité de traitement *UT* et une unité mémoire *UM*. Soit l'expression *c = (a*var1)+(b*var2)* où *var1* et *var2* sont des variables stockées en mémoire. La figure 1 (a) montre le Graphe flot de Signaux SFG de cette expression. Ce composant reçoit les données d'entrée *a* et *b* provenant via le *bus1* et retourne le résultat *c* sur le *bus2*. Toutes les E/S consommées et produites par l'unité de traitement sont lues et produites suivant un ordre de séquence fixe *S =(a,b,c)*, càd $t_a<t_b< t_c$. La séquence de lecture des variables internes est complètement déterministe, càd $t_{var1} < t_{var2}$. avec $t_{var1} = t_a$ et $t_{var2} = t_b$. Dans ce contexte, un seul banc mémoire est suffisant pour satisfaire les contraintes temporelles décrites dans la figure 1 (b) où la latence est de 2 cycles. La figure 1 (c) présente une architecture de composant qui intègre 1 multiplieur, 1 additionneur et 5 registres.

Considérons maintenant la séquence de transfert de données suivante $S_{busses} = (a \mid b, c)$ càd $t_a = t_b < t_c$. Si la latence requise pour produire le résultat est suffisamment grande (≥ 3 cycles), elle permet de réordonner (sérialiser) les données d'entrée *a* et *b*. La solution architecturale incluant un banc mémoire peut être réutilisée. Cependant, cette solution nécessite l'ajout d'un adaptateur (wrapper) d'entrée composé d'un registre, d'un multiplexeur et d'un contrôleur. Si la latence est faible (≤ 2 cycles), le concepteur doit concevoir un nouveau composant intégrant 2 multiplieurs, 1 additionneur, 7 registres et 2 bancs mémoire (voir figure 1 (d)). Le concepteur peut utiliser des composants préconçus ou des macro-générateurs, qui reposent toutefois sur des modèles architecturaux fixes avec des possibilités de personnalisation restreintes.

Un nouveau flot de conception basé sur la synthèse HLS sous contraintes est nécessaire pour introduire de la flexibilité et faciliter la conception de composants orientés TDSI. Il doit inclure (1) des modèles formels pour représenter les contraintes d'E/S et de mémorisation, (2) des phases d'analyse pour vérifier la faisabilité en fonction des contraintes et (3) des méthodes et techniques pour optimiser la synthèse.

3. Approche de conception

Le point d'entrée de notre outil de synthèse est une description algorithmique qui spécifie l'algorithme d'une fonction du domaine du TDSI. La spécification est d'abord compilée pour obtenir une représentation intermédiaire, le Graphe Flot de Signaux (*SFG*) (voir Figure 1).

3.1 Prise en compte des contraintes temporelles

Nous générons un graphe de contraintes algorithmiques (*ACG*) à partir des temps de traversée des opérateurs et des dépendances de données exprimées dans le *SFG*. Les latences des opérateurs sont assignées aux nœuds opération du graphe *ACG* pendant l'étape de sélection de la synthèse de haut niveau. A partir de la description du système et de son modèle architectural, le concepteur connecte le composant au reste du système en spécifiant les cadences des entrées et sorties, les séquences d'accès et les informations temporelles sur les transferts. Nous définissons un modèle formel appelé *IOCG* (IO Constraint Graph) qui supporte l'expression des contraintes d'intégration pour chaque bus ou port qui au travers desquels le composant communique avec son environnement.

Enfin, nous générons un Graphe de Contrainte Globale (*GCG*) en fusionnant les graphes *ACG* et *IOCG*. La fusion est réalisée intégrant les nœuds de *IOCG* et leurs contraintes sur les nœuds d'entrées et de sorties du *ACG*. Une contrainte minimum sur les nœuds de sortie (date au plus tôt pour le transfert de donnée) est transformée en une contrainte de temps maximum sur le *GCG* (date au plus tard pour le calcul et la production d'une donnée).

Nous vérifions dans une première étape que la cadence d'itération spécifiée pour l'application est compatible avec les dépendances de données de l'algorithme et les latences des opérateurs pour la technologie spécifiée. Nous vérifions ensuite que les dates de production des sorties sont toujours respectées pour le jeu de contraintes temporelles fournies pour les données d'entrée.

3.2 Prise en compte de la mémoire interne

Dans notre approche, une table mémoire est extraite à partir du *SFG*. Tous les nœuds de données du *SFG* (hors E/S) sont présents dans la table. Le concepteur peut distribuer et placer les données de l'application et ainsi définir un mapping mémoire. A partir de la table, nous construisons un Graphe de Contraintes Mémoire, *MCG* qui représente tous les conflits d'accès aux données. Le *MCG* est ensuite utilisé pour résoudre les conflits d'accès aux données placées en mémoires internes lors de l'ordonnancement des opérations. Des informations complémentaires sur les modèles formels et la conception des mémoires peuvent être trouvées dans [7], [8] et [9].

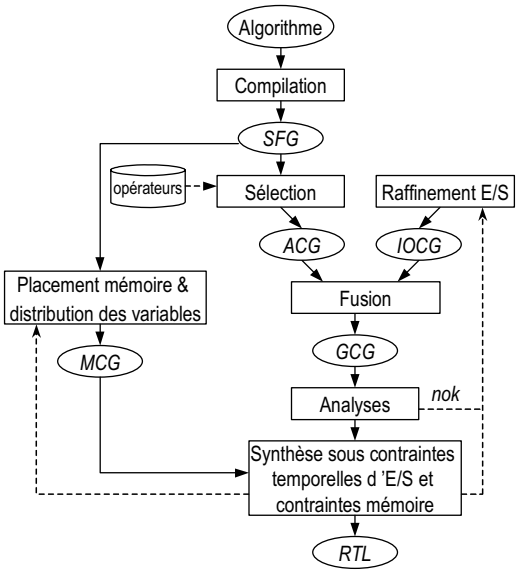


Fig.2: Flot de conception

```
Ordonnancement
1)   Calcul de la mobilité(GCG)
2)   pour (temps = 0; temps < fin; temps = temps + t_cycle)
3)      liste = lister les opérations par priorité(GCG)
4)      Ops_exe = trouver les opérations executables (Liste, temps)
5)      assigner(Ops_exe, ens_operateurs, MCG, temps)
6)   fin pour

assignation
1)   tant que (Ops_exe!= NULL)
2)    Ops_faible_mobilité = prendre-premier(Ops_exe)
3)   if(Op_faible_mobilité->marge > 0)
4)     If(Trouver_mem_conflic(MCG, Ops_faible_mobilité) = FALSE)
5)       If(ens_opérateurs != NULL)
6)          Assignation_Ops(sh_liste, opérateur)
7)       else //pas de conflit de resources ni d'accès à la mémoire
8)          Retarder(Ops_faible_mobilité)
9)   else // marge = 0
10)    If(Trouver_conflit_mem (MCG, Ops_faible_mobilité) = FALSE)
11)      creation_Opérateur_()
12)      Ops_Assignation(sh_liste, opérateur)
13)    else
14)      Exit(cycle, opérateur, operation, banc_mémoire, …)
15)  end if
16)  End while
```

Fig.3: Pseudo code de l'algorithme d'ordonnancement

3.3 Ordonnancement sous contraintes d'E/S et mémoire

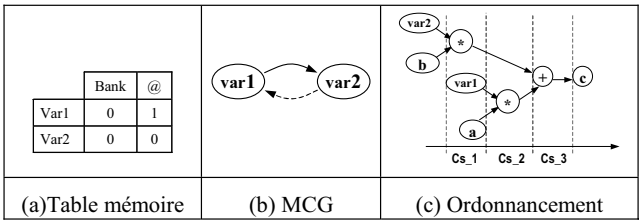
Un ordonnancement classique par liste de priorité repose sur une heuristique de type glouton dans laquelle les opérations exécutables (opérations à ordonnancer) sont classées suivant un ordre de priorité. Dans notre outil, un premier ordonnancement est réalisé à partir du *GCG*. Dans cet ordonnancement, la fonction de priorité dépend de la mobilité des opérations (date ALAP – date ASAP), (voir figure 3). Pour les opérations ayant la même mobilité, le second critère est la marge (date ALAP – temps_courant). Ensuite les opérations sont ordonnancées et assignées aux opérateurs.

Une opération peut être ordonnancée si la date du cycle courant est supérieure à la date ASAP de l'opération. Toutefois, si deux opérations exécutables ont besoin de la même ressource matérielle, l'opération qui a la mobilité la plus faible (donc la priorité la plus forte) est ordonnancée, la seconde est retardée. Lorsque la marge devient nulle, de nouveaux opérateurs sont alloués aux opérations. Pour réaliser un ordonnancement sous contraintes mémoire, nous introduisons des opérateurs d'accès mémoire et utilisons un critère d'accessibilité aux données sur le *MCG*. A chaque cycle d'accès à la mémoire, il y a autant d'accès mémoire possibles qu'il y a de ports d'accès libres. Les opérations exécutables sont toujours classées suivant la mobilité, mais les opérations dont les opérandes ne respectent pas le critère d'accessibilité à la mémoire sont retirées de la liste. Ainsi, lorsque la marge est égale à zéro et que les opérations ne peuvent être ordonnancées à cause des conflits d'accès à la mémoire, le processus de synthèse est arrêté et le concepteur doit cibler une solution alternative pour l'architecture du composant matériel en révisant le mapping mémoire ou en modifiant les caractéristiques des communications.

Notre technique d'ordonnancement est illustrée en figure 3 et reprend l'exemple précédent. Les contraintes sont maintenant les suivantes : *S =(a|b,c)*. La table mémoire (figure 4 (a) est extraite à partir du *SFG*. Le concepteur a défini un mapping mémoire dans lequel les données sont distribuées et placées dans un seul banc mémoire. Les accès à cette mémoire peuvent être séquentiels (1 cycle) ou aléatoires (2 cycles). Les données internes *var1* et *var2* sont respectivement placées aux adresses *@1* et *@0* dans le banc mémoire *bank0*. Notre outil construit le *MCG* (figure 4 (b)) correspondant. En plus des contraintes mémoire, le concepteur spécifie deux latences *Lat1=3 cycles* et *Lat2=2cycles*.

Pour la latence *Lat1*, les accès séquentiels s'effectuent entre *var2* → *var1* (un arc pointillé avec le poids Wseq entre *var2* → *var1*). Pour gérer les conflits d'accès à la mémoire, nous définissons un tableau d'accès pour chaque port. Dans notre exemple, le tableau n'a qu'une ligne puisque nous utilisons une seule mémoire simple port (bank0). Le tableau des accès mémoire contient N éléments, avec N = cadence/temps d'accès mémoire séquentielle. La valeur de chaque élément du tableau indique si la mémoire est accessible ou non au temps de cycle courant(c_step). Nous utilisons le *MCG* pour produire un ordonnancement qui favorise les accès en rafale à la mémoire. Si deux opérations ont la même priorité (margin = Lat1-T(+)-T(*) = 1 cycle) et requièrent un accès à la même mémoire, l'opération qui sera ordonnancée sera celle dont l'adresse de l'opérande est consécutive à celle du précédent accès. Par exemple, les opérations de multiplication *(a*var1)* et *(b*var2)* ont la même

mobilité. Au cycle c_step *cs_1*, elles sont toutes deux exécutables et leurs opérandes *var1* et *var2* sont en dans le banc mémoire bank0. Le *MCG_1* indique que la séquence *var2* → *var1* plus courte que *var1* → *var2*. Nous ordonnançons *(b*var2)* au cycle c_step *cs_1* et *(a*var1)* au cycle c_step *cs_2* pour favoriser les accès (voir figure 4 (c)).



(a)Table mémoire (b) MCG (c) Ordonnancement

Fig. 4: Scheduling under I/O timing and latency constraint

Pour la latence *Lat2*, les opérations de multiplication *(a*var1)* et *(b*var2)* ont la même mobilité (mobilité nulle). Les deux opérations doivent obligatoirement être ordonnancées au cycle c_step *cs_1*. A cause des conflits d'accès à la mémoire, il n'y a pas de solution d'ordonnancement. Le concepteur doit donc revoir les contraintes de conception. Il peut cibler une solution alternative en ajoutant un banc mémoire ou en augmentant la latence du traitement.

4. Résultats expérimentaux

Nous présentons les résultats de synthèse sous contraintes utilisant notre outil de synthèse de haut niveau GAUT [10]. L'algorithme utilisé pour ces expériences est une FFT. Cette FFT lit 128 échantillons réels *X(k)* en entrée et produit des sorties *Y(k)* composées d'une partie réelle *Yr(k)* et d'une partie imaginaire *Yi(k)*. Le *SFG* de cet algorithme est composé de 16897 arcs et 8451 nœuds. Plusieurs synthèses sont réalisées en utilisant une fréquence d'horloge à 200 MHz et une bibliothèque dont les temps de traversée des opérateurs sont de 2 cycles pour les multiplieurs et de 1 cycle pour les additionneurs, les soustracteurs et les accès mémoire.

Dans la première expérience, nous synthétisons le composant FFT sous contrainte d'E/S, puis analysons les besoins mémoire. La latence de la FFT est définie par la contrainte temporelle maximale entre le premier nœud d'entrée et le dernier nœud de sortie du *GCG*. La latence spécifiée (la plus courte en considérant les dépendances de données et le temps de traversée des opérateurs) est de 261 cycles. Les contraintes d'E/S impose au composant FFT une seule lecture d'échantillon *X* et une seule production d'échantillon *Y* par cycle d'horloge. Le composant résultant est composé de 20 multiplieurs, 8 additionneurs et 10 soustracteurs (Voir Table 1, Exp1). Pour respecter les contraintes d'E/S, l'architecture mémoire est relativement complexe (8 bancs mémoire).

Dans la deuxième expérience, la synthèse est réalisée uniquement sous contraintes mémoire. C'est le parallélisme d'accès aux données qui fixe la latence minimale du composant. Nous réalisons la synthèse en allouant le même nombre d'opérateurs que pour la première expérience et analysons la latence obtenue et les besoins en bus d'E/S. Les contraintes mémoires sont les suivantes, deux mémoires SRAM simple port sont utilisées pour stocker respectivement les 128 coefficients réels et les 128 coefficients imaginaires. La latence imposée par le mapping mémoire est de 215 cycles (Table 1, Exp 2). Dans ce cas, l'architecture nécessite 50 bus (ports) d'E/S ce qui complique fortement les échanges avec le système. L'ajout d'une unité de communication est alors nécessaire.

La dernière expérience synthétise la FFT sous contraintes d'E/S et sous contraintes mémoire. Nous conservons le mapping mémoire de l'expérience 2 et définissons la latence permettant de garantir la cadence des E/S de la première expérience. L'architecture contient 17 multiplieurs, 8 additionneurs et 10 soustracteurs (Table 1, Exp 3). La latence est plus grande mais la complexité de l'architecture est moindre puisque qu'il n'y a que deux bancs mémoire et que l'introduction d'une unité de communication n'est pas nécessaire. La synthèse sous contraintes mémoire et contraintes d'E/S permet de gérer efficacement les communications et la mémorisation en conservant une complexité d'architecture raisonnable.

Table 1 : Résultats de synthèse

	Banc mémoire.	Bus d'entrée	Bus de sortie	Sous.	Add.	Mult.	latence *(en cycle)*
EXP1	8	1	2	10	8	20	261
EXP2	2	36	14	10	8	20	215
EXP3	2	1	1	10	8	17	343

5. Conclusion

Dans ce papier, une méthodologie de conception de composants matériels sous contraintes d'E/S et mémoire est présentée pour les applications TDSI. Elle repose sur la modélisation des contraintes, leurs analyses et la synthèse de haut niveau. Elle aide le concepteur à implanter efficacement des applications complexes sur des composants matériels. Les résultats produits pour une application TDSI montrent l'intérêt de la méthodologie et de la modélisation des contraintes qui permettent de réaliser un compromis entre la latence, la cadence des entrées sortie et le mapping mémoire.

References

[1] Codesimulink, http://polimage.polito.it/groups/codesimulink.html

[2] J. Ruiz-Amaya, and Al., " *MATLAB/SIMULINK-Based High-Level Synthesis of Discrete-Time and Continuous-Time □□ Modulators*", In Proc. of DATE 2004.

[3] L. Reyneri, F. Cucinotta, A. Serra, and L. Lavagno. « A hardware/software co-design flow and IP library based on Simulink. », In Proc. of DAC, 2001.

[4] Mathworks, http://www.mathworks.com/

[5] H. Ly, D. Knapp, R. Miller, and D. McMillen, « Scheduling using behavioral templates," in *Proc. Design Automation Conference DAC'95*, June 1995

[6] D. Knapp, and *al.*, « Behavioral synthesis methodology for HDL-based specification and validation," in *Proc. of DAC*, 1995.

[7] P. Coussy, E. Casseau, P. Bomel, A. Baganne, E. Martin, "Constrained Algorithmic IP Design for System-On-Chip", *in Integration, the VLSI journal, 2004 (to appear)*.

[8] P. Coussy , D. Gnaedig, and al., " A Methodology for IP Integration into DSP SoC: A Case Study of a MAP Algorithm for Turbo Decoder", *In Proc. of ICASSP, 2004*

[9] G. Corre, E. Senn, and al., "Memory accesses Management During High Level Synthesis", *In Proc. of CODES-ISSS, 2004.*

[10] GAUT tool: http://web.univ-ubs.fr/gaut/

Rétroprojection 2D sur plateforme SOPC

Stéphane Mancini, Nicolas Gac , Michel Desvignes

Laboratoire des Images et des Signaux
46 avenue Félix Viallet, 38031 Grenoble Cedex 1, France
{stephane.mancini, nicolas.gac, michel.desvignes}@lis.inpg.fr

Résumé – Le développement et la diffusion des équipements TEP passent par la réduction des temps de calcul de la reconstruction des images acquises. Aussi cet arcticle présente une solution mixte logicielle/matérielle pour l'accélération de la reconstruction 2D sur une plateforme SoPC (System on Programmable Chip), la nouvelle génération de circuits reconfigurables. Le verrou technologique posé par la latence des accès mémoire est levé grâce au cache 2D Adaptatif et Prédictif (cache 2D-AP).

Abstract – Reduction of image reconstruction time is a key point for the development and spreading of PET scans. Thus this article presentes a hardware/software architecture which aims at accelerating the 2D reconstruction on a SoPC (System on Programmable Chip) plateform, the new generation of reconfigurable chip. Issue posed by the latency of memory accesses has been solved thanks to the 2D Aptative and Predictive cache (2D-AP cache).

1 Introduction

La reconstruction d'images en tomographie à émissions de positons (TEP) s'effectue à l'aide de processeurs classiques et demande un temps de calcul important. Elle est ainsi souvent découplée du processus d'acquisition. Or une reconstruction rapide des données acquises en imagerie TEP faciliterait le positionnement du patient et aiderait à détecter des problèmes potentiels survenus lors de l'acquisition. L'obtention rapide d'une image permettrait de réduire sensiblement la durée des examens cliniques TEP pour la détection précoce du cancer, l'évaluation de la propagation de la maladie et de la réponse au traitement. Cela permettrait de réduire les coûts de ces examens et favoriserait la diffusion de cette technique efficace d'imagerie médicale moléculaire. De plus, son utilisation pour la mammographie ou pour l'étude des petits animaux (micro PET) nécessite un système simple, flexible et rapide.

La reconstruction d'image s'effectue généralement en deux étapes : acquisition des données puis rétroprojection filtrée. Les travaux de ces dernières années portant sur l'acquisition en TEP humaine ont augmenté la qualité et la vitesse d'acquisition en utilisant du matériel dédié [1]. En effet, les processeurs de traitement numérique (DSP) et/ou les nouvelles technologies de logique programmable (FPGA) offrent une solution modulaire, adaptable et programmable, qui réduit les coûts et le temps en recherche et developpement. La rétroprojection n'est pas uniquement utilisée en TEP, elle l'est également dans d'autres domaines de la tomographie que ce soit en tomodensitométrie à rayons X ou en Tomographie à Emission Monophotonique (TEMP). Une autre famille d'algorithme se développe de plus en plus : les algorithmes itératifs. Ils augmentent sensiblement la qualité des images (artéfacts, ratio signal sur bruit) mais nécessitent un temps de reconstruction notablement supérieur aux techniques classiques de rétroprojection filtrée.

Il existe plusieurs implémentations de rétroprojection sur des clusters de PC [2, 3], sur du matériel dédié comme les ASICS ou les FPGA [4] (pour les scanners X) ou avec une architecture logicielle/matérielle [5]. Une autre stratégie est d'utiliser les processeurs 3D classiques comme les GPU (Graphic Processor Unit) pour accélérer la reconstruction [6]. Le principal goulot d'étranglement de tels systèmes est l'accès à la mémoire stockant les sinogrammes. En effet, pour reconstruire un pixel il faut parcourir entièrement tous les sinogrammes selon une courbe sinusoïdale. La recherche de parallélisation des calculs est rendu difficile par la limitation des accès mémoire par la bande passante de la mémoire principale.

Cet article présente un système de reconstruction par rétroprojection 2D implémenté sur une plateforme SOPC (System On Programmable Chip). L'originalité de ce système réside dans son architecture, solution efficace au problème soulevé par les accès aux mémoires externes. En effet, ces mémoires de type SDRAM sont peu coûteuses mais lentes et créent ainsi un goulot d'étranglement. Le cache 2D adaptatif et prédictif (cache 2D-AP) décrit en [7] constitue la base cette architecture. Cette stratégie s'est avérée efficace : elle permet de réduire le temps de reconstruction d'un ordre de magnitude par rapport aux solutions logicielles.

2 Objectifs

2.1 L'algorithme

L'algorithme implémenté effectue la rétroprojcetion des données acquises par le scanner. Ces dernières, appelées sinogrammes, forment la transformée de Radon de la fonction f représentant la densité du taux d'émission radioactive du volume observé. Le sinogramme est une matrice

image à deux dimensions à K colonnes. Chaque colonne k correspond à la projection orthogonale de f sur r détecteurs placés orthogonalement sur une ligne inclinée de l'angle $\theta_k = \frac{k*\pi}{K}$ par rapport à l'axe x. Ainsi, le point du sinogramme $S(\theta_k, r)$ est la somme des pixels sur une Ligne De Réponse (LDR) du scanner qui est perpendiculaire à l'axe des détecteurs et passe par le détecteur r. A partir de ce sinogramme, l'algorithme reconstruit l'image f^* en rétroprojetant les K lignes du sinogramme dans l'espace image.

$$f^*(x,y) = \sum_{k=0}^{K} S(k, r_k) \quad (1)$$

$$r_k = x * \cos(\frac{k\pi}{K}) - y * \sin(\frac{k\pi}{K}) + \text{offset} \quad (2)$$

Cette méthode n'est pas l'exacte transformée de Radon inverse. En effet, elle produit des artefacts en étoile et l'image reconstruite devient floue. Pour améliorer la reconstruction, on peut filtrer le sinogramme mais cette étape étant indépendante de la rétroprojection 2D, nous ne la développerons pas dans cet article.

2.2 Le "challenge"

Le sinogramme étant stocké en mémoire externe de type SDRAM, nous devons avoir une gestion efficace de la mémoire pour compenser la latence et permettre un haut degré de parallélisme. Cette stratégie diffère de celle de M. Leeser [8] qui utilise plusieurs bancs de mémoires SRAM indépendants avec une latence nulle pour améliorer le débit mémoire. Cette dernière constitue cependant une solution coûteuse et il est possible d'obtenir de meilleures performances avec des mémoires SDRAM à faibles coûts malgré une latence plus importantes. Les caches standards ayant pour principe un accès linéaire à la mémoire ne peuvent être une solution satisfaisante étant donné leur complexité technologique et leur faible taux de charge de données utiles. En effet, les accès mémoires nécessaires pour reconstruire un pixel image $f^*(x,y)$ suivent une sinusoïde dans le sinogramme ce qui constitue une faible localité spatialle pour les adresses mémoire. Pour accélérer ces accès mémoire, un nouveau mécanisme de cache est nécessaire. Une prédiction des points du sinogramme dont a besoin l'unité de calcul permettra au cache de charger les données pendant le processus de calcul.

3 Architecture

3.1 Architecture système

Une plateforme SoPC (System on Programmable Chip) a l'avantage de permettre une implémentation efficace et économique de la rétroprojection 2D. Pour davantage de flexibilité, et à coût réduit, le sinogramme et l'image reconstruite sont stockées en SDRAM, ou DDR-SDRAM, et la hiérarchie mémoire indispensable au recouvrement des calculs et accès mémoires, à grande latence, est construite à l'aide des blocs de mémoire embarquée dans le circuit SoPC.

L'architecture présentée figure 1 est évaluée sur une carte qui dispose d'une interface PCI avec une station hôte pour permettre l'échange de données avec le système. Cette carte dispose d'un circuit Virtex 2 Pro 2VP20, de 32 MO de SDRAM et 128 MO de DDR-SDRAM et d'un bridge PCI, réalisé sur un FPGA Spartan. Les données sont échangées entre le PC hôte et la carte grâce à un mécanisme de synchronisation.

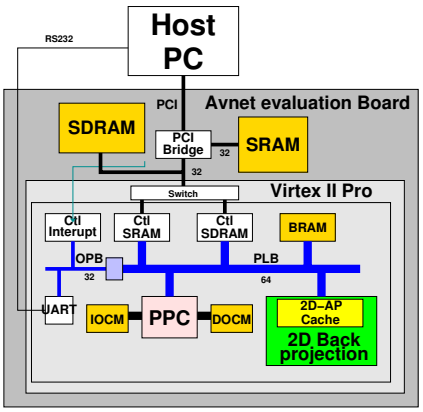


FIG. 1 – Architecture Système

L'architecture système est constituée :
- d'un processeur PPC en charge de la synchronisation des calculs et des communications
- d'une unité de rétro-projection avec son cache 2D-AP illustré en figure 2.
- de mémoire externe

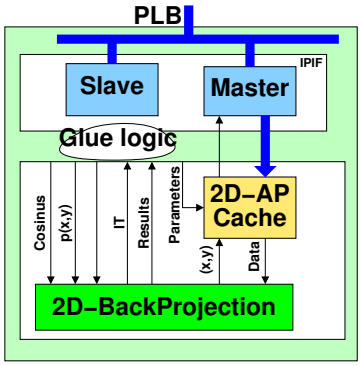


FIG. 2 – Unité de rétro-projection

Pour bénéficier au maximum de la cohérence spatiale 2D et temporelle des calculs, l'unité de rétroprojection reconstruit un bloc de forme quelconque de pixels voisins de f^*. Par souci de simplicité les blocs sont des carrés de pixels mais des hexagones pourraient être utilisés. Les données du sinogramme sont fournies à l'unité de rétroprojection par le cache 2D-AP qui se charge de leur transfert depuis la mémoire externe, le module étant maître sur le bus PLB.

Pour réduire les temps de calculs, l'unité de rétroprojection est parallélisée et une hiérarchie mémoire fournit les

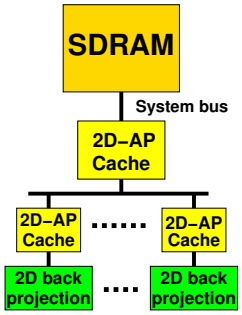


FIG. 3 – Architecture parallèle et Cache 2D-AP hierarchique

données aux modules : chaque module obtient ses données d'un cache 2D-AP qui lui même les met à jour à partir d'un cache 2D-AP de niveau supérieur. Dans notre exemple, nous nous arrêtons à un cache de niveau 2, comme illustré figure 3.

Cette architecture mémoire originale permet de n'utiliser qu'une seule mémoire externe qui contient tout le sinogramme, contrairement à la solution proposée par [8], tout en autorisant :

- le recouvrement de la latence de la mémoire externe et du bus système
- la réduction du débit à la mémoire externe

Les performances de cette hiérarchie mémoire sont améliorées lorsque la latence se réduit et une connexion directe de la mémoire à l'unité de rétro-projection permettrait de paralléliser massivement les unités de rétroprojection.

3.2 Le cache 2D-AP

3.2.1 Objectifs du cache 2D-AP

Le cache 2D-AP est l'élément original de l'architecture proposée en permettant aux unités de calculs d'accéder aux données sans temps mort. Ceci est rendu possible grâce à un mécanisme générique de prédiction des prochaines données utilisées. Le cache 2D-AP anticipe les accès au sinogramme 2D par une analyse de la séquence des coordonnées de pixels requis pour prédire les prochains accès à partir d'une hypothèse à vitesse de déplacement constante sur l'image. La reconstruction d'un bloc de pixels permet de produire une séquence d'accès au sinogramme qui tire partie au mieux des caractéristiques du cache 2D-AP.

La séquence produite par l'unité de rétroprojection pour reconstruire un bloc est l'union de toutes les sinusoïdes nécessaires à la reconstruction de chaque pixel. Dans le cas d'un carré de pixels, la séquence obtenue est une sorte de "tube" dont la figure 4 donne un exemple. Pour chaque angle θ, on accède à tous les points du sinogramme, sur une même colonne, nécessaires à la reconstruction de chaque pixel du bloc.

3.2.2 Fonctionnement

La prédiction de la zone en cache permet de réduire le taux de défaut de cache lorsque les accès pixels suivant se

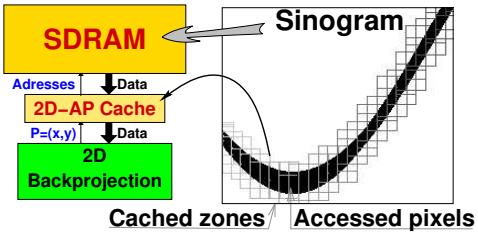


FIG. 4 – Concept du cache 2D-AP

déplacent sur l'image. Le centre et la taille de la zone sont déterminés par le calcul de la moyenne et du pseudo écart type (PSD=pseudo standard deviation) des accès pixels. La moyenne et le PSD sont calculés à l'aide de filtres passe-bas récursifs du premier ordre. En supposant une distribution uniforme des accès pixels autour de la moyenne, nous pouvons estimer que, pour une courte durée, ils sont dans un rectangle centré sur la moyenne et dont les demi-largeurs valent deux fois le PSD sur chaque axe. La zone ainsi calculée est mise à jour lorsque la moyenne varie.

Un mécanisme de prédiction permet d'anticiper la position du centre du cache lorsque les accès pixels suivent un chemin complexe dans l'image et réduit ainsi le coût de la latence des accès mémoire. La mise à jour ne se produit que lorsque la moyenne a suffisament variée et la nouvelle zone est centrée sur une anticipation de la moyenne, supposée se déplacer à vitesse constante. Ainsi, une zone de garde est définie autour du centre de la zone en cache et la mise à jour se fait lorsque la moyenne calculée sort de cette zone de garde, dans la direction de sortie, comme illustrée figure 5.

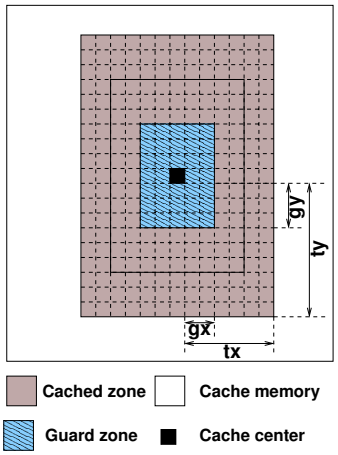


FIG. 5 – Zones du cache 2D-AP

Pour réduire le coût matériel du contrôle et obtenir de bonnes performances les filtres passes bas pour le calcul des moyennes sont réalisés à l'aide de simples additions et décalages. Ce sont des filtres RII du premier ordre dont l'équation est $s_n = ce_n + (1-c)s_{n-1}$, e signal d'entrée et s signal filtré, et nous notons $s = \mathrm{f}^c(e)$. Le paramètre $1-c$ correspond à la constante de temps du filtre RC équivalent. Lorsque les suites s et e sont représentées en virgule fixe et c une puissance de $\frac{1}{2}$, nous ne réalisons plus que des additions et décalages.

4 Résultats

Les mesures effectuées sur la plateforme et les résultats de simulation nous montrent que la hiérarchie mémoire choisie est efficace et nous permet des accélérations importantes par la parallélisation des opérateurs et ceci en n'ayant qu'une mémoire externe qui contient toutes les données. Les métriques obtenues sont données dans le tableau 4 et montrent une accélération quasi-linéaire avec le nombre d'opérateurs. Elles correspondent à la reconstruction d'une image $x_{max} * y_{max} = 320 * 320$ à partir d'un sinogramme de résolution angulaire $K = 512$.

Système	**Cycles**	**Temps**
Logiciel		
Pentium 3 (1 GHz)		3,5 s
PPC (VirtexII-Pro)		94 s
Matériel simple		
Idéal	52.10^6 $(320 * 320 * 512)$	1,1 s
Sans cache	52.10^6*28 (Idéal*Latence)	30 s
1 unité	78.10^6	1,5 s
Matériel parallélisé		
2 unités	42.10^6	0,8 s
4 unités	21.10^6	0,42 s
9 unités	11.10^6 (simulé)	0,22 s

TAB. 1 – Accélération par le module de rétroprojection@50 Mhz

Idéalement, nous devrions pouvoir réaliser une reconstruction sans temps mort, c'est à dire en $x_{max} * y_{max} * K = 320 * 320 * 512$ cycles d'horloge. Ces performances idéales sont altérées du fait d'une part de la synchronisation entre l'opérateur matériel et le PPC et d'autre part des performances du Cache 2D-AP sur le bus système. Les mesures effectuées nous montrent que les simulations réalisées pour la reconstruction d'un bloc sont fiables et peuvent être extrapolées à la reconstruction d'une image complète.

L'ensemble du module a été décrit en VHDL générique paramétrable pour explorer rapidement les différentes configurations possibles. Le cache est entièrement modulaire de façon a pouvoir mesurer l'efficacité de différents type d'estimateurs prédictifs et pour pouvoir construire une hiérarchie par simple assemblage de blocs. Le tableau 4 donne la complexité du système en nombre de LUT pour une synthèse sur cible Xilinx VirtexII-Pro.

Module	CLB	FG
1 unité	1078	2155
2 unités	2253	4506
4 unités	3877	7753

TAB. 2 – Synthèse de l'IP et son cache 2D-AP.

5 Conclusion

Nous avons présenté dans cet article un système de reconstruction par rétroprojection 2D implémenté sur une plateforme SoPC. Ce système a été conçu de manière à réduire les erreurs de reconstruction dues aux calculs en virgule fixe. Le cache 2D-AP permet de réduire le goulot d'étranglement existant lors des accès en mémoire. Lors de la reconstruction, les données nécessaires sont prédits statistiquement afin que le cache puisse les charger avant que l'unité de rétroprojection les utilise. De plus, l'algorithme de rétroprojection a été implémenté de façon à augmenter la localité spatiale et temporelle, l'utilisation du cache en devient plus pertinante.

La rétroprojection est utilisée par les algorithmes itératifs plus sophistiqués. Ces derniers offrent une meilleure qualité d'image mais en contrepartie ont un temps reconstruction beaucoup plus long. La réalisation présentée dans cet article est une première étape pour construire un système de reconstruction itératif, adaptable et flexible. Implémenté sur un SOPC, le module de rétroprojection 2D peut être dupliqué pour paralleliser les calculs. Un même module de cache peut être partagé ou une hiérarchie de caches mémoire peut être mise en place. Ainsi en s'appuyant sur la localité spatiale, nous pouvons reconstruire simultanément plusieurs pixels et la vitesse de reconstruction est alors notablement améliorée.

Références

[1] M.S. Musrock et al. Performance characteristics of a new generation of processing circuits for pet applications. *IEEE Tr. Nucl. Sci.*, 50(4) :974 – 978, August 2003.

[2] D.W. Shattuck, J. Rapela, E. Asma, A.Chatziioannu, J.Qi, and R.M. Leahy. Internet2-based 3D PET image reconstruction using a PC cluster. *Phys. Med. Biol.*, 47(15) :2785–2795, August 2002.

[3] S. Vollmar, C. Michel, J.T. Treffert, D.F. Newport, M. Casey, C. Knoss, K.Wienhard, X. Liu, M.Defrise, and W.-D. Heiss. Heinzelcluster accelerated reconstruction for FORE and OSEM3D. *Phys. Med. Biol.*, 47(15) :2651–2658, August 2002.

[4] Nikolay Sorokin. *An FPGA-Based 3D Backprojector*. PhD thesis, Universität des Saarlandes, Allemagne, 2003.

[5] J. Müller, D. Fimmel, R. Merker, and R. Schaffer. Hardware- software system for tomographic reconstruction. *J. Circuits Syst. Comp.*, 12(2) :203–229, April 2003.

[6] F. Xu and K. Mueller. Ultra fast 3D filtered back projection on commodity graphics hardware. In *Proc. IEEE Int. Symp. Biomedical Imaging (ISBI'04)*, pages 571–574, Arlington, USA, April 2004.

[7] S. Mancini, N. Gac, and M. Desvignes. Etude d'un cache 2D adaptatif et prédictif pour le traitement d'image. In *Journées Francophones sur l'Adéquation Algorithme Architecture (jFAAA'05)*, Dijon, France, 18-21 Janvier 2005.

[8] M. Leeser, S. Coric, E.Miller, H. Yu, and M.Trepanier. Parallel-beam backprojection an FPGA implementation optimized for medical imaging. *J. VLSI Signal Proc.*, 39(3) :295–311, March 2005.

Optimisation énergétique d'un système de communication dédié à un réseau de capteurs

Mickaël CARTRON[1], Olivier SENTIEYS[1]

[1]ENSSAT/IRISA
6 rue de Kerampont, BP 80518, 22305 Lannion Cedex, France
cartron@irisa.fr, sentieys@irisa.fr

Résumé – Cet article présente une méthode d'optimisation énergétique d'un système de communication mobile utilisé dans un contexte de réseau de capteurs. Plus particulièrement, il se focalise sur l'optimisation de l'expression de l'énergie consommée par les couches liaison et physique réalisant la communication entre deux nœuds en liaison directe. Les fonctions du niveau liaison comprennent dans notre cas la correction d'erreurs et la retransmission automatique des paquets erronés. Les moyens de modéliser l'énergie totale (dynamique et statique) d'un système sont détaillés. Des points de fonctionnements optimaux sont ensuite déterminés à l'aide d'expressions liant les performances et la consommation de notre système.

Abstract – This article presents an energetic optimization method of a sensor network's communication system. It focuses on the optimization of the expression of the energy consumed by the link and physical levels. These two levels realize a direct communication between nodes sharing the same channel. The link level functions include error correction and automatic retransmission of wrong packets. Several methods of evaluation of the total power (dynamic and static) of systems are exposed. Then, optimal functioning points are found using expressions involving the performance and the consumption of our system.

1 Introduction

La recherche dans le domaine des capteurs subit actuellement une révolution importante, ouvrant des perspectives d'impacts significatifs dans de nombreux domaines d'applications (sécurité, santé, environnement, sécurité alimentaire, fabrication, télécommunications, robotique, etc.). Les progrès dans le domaine des communications numériques sans fil nous permettent d'imaginer des réseaux de capteurs ad-hoc totalement autonomes avec des durées de vie importantes. L'augmentation de l'autonomie des capteurs est le principal centre d'intérêt de la recherche dans ce domaine et c'est celui qui est développé dans cet article. Dans une première partie, nous présenterons le concept de réseau de capteurs après quoi nous décrirons l'architecture générique et globale d'un capteur. Dans une troisième partie, nous ferons la description algorithmique des couches basses de la pile de protocole réseau d'une chaîne de communication adaptée à un réseau de capteurs. Enfin nous montrerons comment optimiser la partie radio du capteur, ce qui permettra la définition de modules globaux de contrôle d'énergie, selon une méthodologie de conception globale telle que celle décrite dans [1]. Cette méthodologie consiste à envisager un système de transmission dans sa globalité plutôt qu'un ensemble de blocs séparés, afin de se situer dans un espace d'optimisation plus vaste. Notre méthode d'optimisation est similaire à celle présentée dans [2], mais dans un cas plus général et sans imposer la norme IEEE 802.11a. De plus, nous allons plus loin, comme dans [3], car nous intégrons la retransmission automatique des erreurs dans nos calculs.

2 Présentation générale des réseaux de capteurs

Un réseau de capteurs est un ensemble de capteurs communicants qui est dispersé dans un environnement afin de le surveiller. Depuis quelques années, les réseaux de capteurs ont suscité un intérêt croissant dans le monde des télécommunications, du traitement du signal et des réseaux sans fil. En effet, les progrès technologiques rendent possible sur le papier de concevoir des objets communicant sans fil et ayant néanmoins des autonomies de fonctionnement très élevées, de l'ordre d'une ou plusieurs années. Un réseau de capteurs est donc constitué d'un ensemble de nœuds qui remplissent deux rôles : un rôle de source d'informations et un rôle de relais pour le reste du réseau. Les réseaux de capteurs doivent être *ad-hoc*, tout d'abord dans un souci de simplicité d'installation, mais aussi et surtout dans le souci de permettre au réseau de rester opérationnel même après des défaillances ponctuelles de nœuds. Ils doivent pouvoir s'auto-gérer, en utilisant des protocoles permettant d'apprendre des éléments tels que la topologie du réseau, le positionnement relatifs des capteurs au sein du réseau [4], les routes possibles pour communiquer avec d'autres nœuds donnés [5]. Les communications multi-étapes de courte portée sont privilégiées par rapport à une transmission directe avec une station de base.

De nombreuses applications ont déjà été imaginées, dans des domaines variés (agriculture, bâtiments, transports, défense, sécurité civile, etc.).

3 Architecture générique d'un nœud de communication

L'architecture générique d'un capteur a été définie en grande partie à partir de [6]. Elle se compose d'une partie chargée de l'alimentation des circuits avec éventuellement un système de récupération d'énergie, d'un ou plusieurs capteurs et leurs convertisseurs analogiques numériques associés, d'un bloc de traitement numérique composé d'un processeur éventuellement associé avec un composant de traitement spécialisé de type ASIC ou FPGA avec des blocs mémoire associés. Enfin, un module radio analogique permet de réaliser les communications.

4 Description algorithmique d'un élément du réseau

Nous avons subdivisé l'ensemble des traitements en trois niveaux. Le premier niveau est appelé niveau application et il comprend trois classes de tâches haut niveau, utilisatrices de l'accès sans fil. On distingue trois tâches. La première est la tâche de positionnement qui permet à un capteur de déterminer sa position géographique en réalisant des mesures de distances et des triangulations. La deuxième tâche a pour rôle la gestion du routage des messages au sein du réseau ad-hoc. Enfin, la troisième tâche du niveau application est la tâche de gestion des valeurs mesurées, pour lancer des requêtes locales et calculer des moyennes, par exemple. Le niveau intermédiaire est le niveau liaison. Son rôle est de fournir aux niveaux supérieurs une communication réussie entre deux éléments en liaison directe. Il doit gérer le contrôle d'erreurs et les ré-émissions automatiques de paquets et l'accès au média. L'accès au média est contrôlé par un protocole de type CSMA/CA (Carrier Sense Multiple Access with Collision Avoidance). Ensuite, le niveau le plus bas du traitement est le niveau physique dont le rôle est de mettre en forme les signaux par la modulation et la génération de porteuses. Pour réaliser le contrôle d'erreurs, nous avons sélectionné deux techniques : un codage/décodage en bloc à l'aide d'un code de Hamming d'une part et un codage convolutif associé à un décodage de Viterbi d'autre part. Une chaîne sans correction d'erreur a également été envisagée. La ré-émission automatique employée est basée sur la technique SACK (Selective Acknowledgement) [3], avec laquelle les paquets erronés sont ré-émis et seulement ceux-ci. Ensuite, le niveau le plus bas du traitement est le niveau physique dont le rôle est de mettre en forme les signaux par la modulation et la génération de porteuses. L'algorithme que nous étudions inclus le niveau liaison et le niveau physique. Le niveau liaison est réalisé avec des composants numériques et sa consommation sera modélisée par les termes P_{dyn} et P_{stat}, respectivement pour les consommations dynamique et statique. Le niveau physique est constitué de composants analogiques et sa consommation sera modélisée par les termes P_{an} et P_{amp}, respectivement pour la puissance consommée par la mise en forme du signal analogique et les circuits de réception, d'une part et par la puissance consommée par l'amplificateur de puissance d'émission, d'autre part.

5 Performances conjointes des niveaux liaison et physique

Les objectifs de l'étude sont multiples : il faut exprimer l'énergie consommée par bit transmis en fonction des paramètres choisis dans le but de mettre en évidence une configuration des paramètres du système permettant d'obtenir un minimum en terme d'énergie par bit transmis. Ces résultats pourront être ensuite utilisés pour donner les moyens à une application de définir ses besoins en terme de qualité de transmission ou de taux de transmission et d'intervenir ainsi directement sur la puissance d'émission, les tailles de paquets, et d'autres paramètres via une interface système adéquate. Des travaux similaires on déjà été menés dans [3] et [2]. L'article [3] combine les corrections d'erreurs et la ré-émission automatique, mais envisage uniquement une cible processeur, et s'appuie sur des modèles de consommation assez grossiers, sans prendre en compte l'énergie statique dans le calcul. Mais dans le contexte d'un réseau de capteur, cette hypothèse ne peut être maintenue. L'article [2] s'appuie sur la norme IEEE 802.11a, et cherche à adapter au mieux les paramètres de choix de codage et de puissance d'émission, pour diminuer l'énergie sous diverses contraintes d'applications. Par contre, l'énergie due aux ré-émissions n'y est pas prise en compte.

5.1 Expressions de la consommation

Pour évaluer la puissance consommée par un traitement, on peut utiliser trois méthodes. Dans le cas général on peut donner l'expression suivante, en négligeant les pertes statiques :

$$E_{\mathsf{dyn}} = E_{\mathsf{op}}.N_{\mathsf{opérations}} = a.C_{\mathsf{s}}.A_{\mathsf{op}}.V_{\mathsf{dd}}^2 \qquad [J] \quad (1)$$

avec E_{dyn} qui représente l'énergie dynamique de l'algorithme, E_{op} l'énergie consommée pour une opération "moyenne", $N_{\mathsf{opérations}}$ le nombre d'opérations de l'algorithme, a l'activité du circuit, A_{op} la surface de l'opérateur "moyen" et V_{dd} la tension d'alimentation.

Si on souhaite raffiner l'évaluation de la consommation, il faut alors prendre en compte la puissance statique. La part de la puissance statique est d'ailleurs loin d'être négligeable dans un contexte comme celui d'un réseau de capteurs où le taux d'activité est très faible. Dans ce cas, il faut introduire un terme de puissance statique dans l'expression de l'énergie de l'algorithme. E_{algo} est l'énergie consommée pour exécuter une fois l'algorithme considéré, P_{stat} est la puissance statique totale du circuit réalisant le traitement, T_{algo} est le temps d'exécution de l'algorithme à l'aide du circuit considéré, F_{clk} représente la fréquence de l'horloge, $N_{\mathsf{opérateurs}}$ est le nombre d'opérateurs utilisés pour le calcul et β est un coefficient de consommation statique par unité de surface.

$$E_{\mathsf{algo}} = (\beta.A_{\mathsf{op}}.N_{\mathsf{opérateurs}} + P_{\mathsf{dyn}}).T_{\mathsf{algo}} \qquad [J] \quad (2)$$

$$T_{\mathsf{algo}} \geq \frac{N_{\mathsf{opérations}}}{N_{\mathsf{opérateurs}}.F_{\mathsf{clk}}} \qquad [s] \qquad (3)$$

Ce qui nous permet d'exprimer une *borne inférieure* pour l'expression de l'énergie de l'algorithme :

$$E_{\mathsf{algo}} \geq N_{\mathsf{opérations}}.\left(\frac{\beta.A_{\mathsf{op}}}{F_{\mathsf{clk}}} + E_{\mathsf{op}}\right) \qquad [J] \qquad (4)$$

Cette expression qui correspond à une description haut niveau est certes un peu grossière, mais elle est cependant riche d'enseignements quant aux directions architecturales à prendre pour réaliser un système réellement basse consommation. De plus ce modèle est bien adapté pour avoir une idée de la consommation d'un gros système. Cependant, si on souhaite plus de précision, il faut alors définir exactement le matériel employé pour chaque bloc de traitement. Dans ce cas, la puissance consommée pendant l'exécution de l'algorithme peut s'exprimer :

$$P_{\mathsf{algo}} = \sum_{j\in\mathcal{D}_f}\left(\sum_{i\in\mathcal{D}_{sys}} nb_{i,j}.E_{op_{i,j}}.F_j + \sum_{i\in\mathcal{D}_{sys}} nb_{i,j}.p_{\mathsf{stat}_i}\right)[W] \qquad (5)$$

avec $\mathcal{D}_f$ l'ensemble des fréquences de fonctionnement, F_j la fréquence de fonctionnement du domaine j, $\mathcal{D}_{sys}$ l'ensemble des sous-systèmes (e.g. opérateur arithmétique, registre etc.) fonctionnant à F_j, $nb_{i,j}$ le nombre d'opérateurs de type i fonctionnant à la fréquence du domaine j, $E_{op_{i,j}}$ l'énergie dynamique d'une opération du type i et du domaine j et p_{stat_i} la puissance statique dépensée par un système de type i sous tension. Nous avons utilisé ce dernier modèle, plus précis, pour les calculs de consommation des circuits de codage et de décodage de canal.

5.2 Modélisation fonctionnelle et énergétique du problème

Une retransmission est nécessaire lorsqu'un paquet transmis comporte au moins une erreur. Soit lp le nombre d'éléments binaires d'un paquet, TEB le taux d'erreur binaire et $nberr$ le nombre d'erreurs dans un paquet donné, la probabilité qu'un paquet transmis comporte au moins une erreur est le taux d'erreur par paquet (TEP) et s'exprime :

$$TEP = 1 - \text{prob}(nberr = 0) = 1 - (1 - TEB)^{lp} \qquad (6)$$

Soit NbP, le nombre de paquets d'information à transmettre et deb le débit binaire de transmission du système actif, le temps mis pour une transmission réussie du premier coup d'un paquet d'information s'exprime de la façon suivante :

$$T_{\mathsf{ideal}} = \frac{NbP.lp}{deb} \qquad [s]\,. \qquad (7)$$

Le temps réel de transmission doit donc tenir compte du pourcentage de temps additionnel nécessaire pour les retransmissions. Lors d'une tentative d'envoi, le paquet a la probabilité TEP d'être faux, l'expression du temps réel de transmission est donc :

$$T_{\mathsf{reel}} = T_{\mathsf{ideal}}.\sum_{i=0}^{\infty} TEP^i = \frac{T_{\mathsf{ideal}}}{1 - TEP} \qquad [s] \qquad (8)$$

Afin de réaliser une optimisation énergétique du système de communication, la grandeur à étudier est l'énergie par bit transmis avec succès. Celle-ci est définie par l'expression :

$$E_{bit} = \frac{P}{deb.(1 - \frac{1}{2}.\text{erfc}(\sqrt{RSB}))^{lp}} \qquad [J] \qquad (9)$$

dans laquelle nous avons considéré une modulation de type BPSK, permettant ainsi d'exprimer TEB en fonction du rapport signal sur bruit RSB. P est la puissance globale du système en communication incluant les puissances dynamique et statique consommées par les circuits numériques et analogiques.

$$P = P_{\mathsf{algo}} + P_{\mathsf{an}} + P_{\mathsf{amp}} \qquad [W]\,, \qquad (10)$$

avec P_{an} représentant une constante associée à l'ensemble des circuits analogiques et P_{amp} représentant la puissance consommée par le système d'amplification de puissance précédant l'antenne d'émission.

P_{amp} doit tenir compte de l'efficacité du système d'amplification et dépend de son régime de fonctionnement. Nous avons utilisé un modèle exponentiel tel que celui proposé dans [2] pour représenter cette efficacité. Soit eff l'efficacité du couple amplificateur/antenne et P_{diff} la puissance du signal diffusée par l'antenne, on a alors, avec $A = 0.02$ et $B = 0.140$:

$$P_{\mathsf{amp}} = P_{\mathsf{diff}}.eff = P_{\mathsf{diff}}.A.\exp(B.10\log_{10}(P_{\mathsf{diff}})) \qquad (11)$$

Nous avons également utilisé un autre modèle obtenu à partir des données constructeur d'un composant commercial, le Chipcon CC1020. Nous avons déterminé la somme $P_{\text{an}} + P_{\text{amp}}$ pour ce composant. On a alors $P_{\text{an}} = 0.102$ W, $A = 0.008$ et $B = 0.001$. On peut tout de suite remarquer l'ordre de grandeur très différent pour P_{an}. La consommation très importante du CC1020 peut s'expliquer de plusieurs façons. Tout d'abord, le CC1020 gère sa propre horloge, est doté d'une PLL, de plus il s'agit d'un composant très polyvalent.

5.3 Résultats

L'expression 9 que nous avons déterminée précédemment nous permet de tracer l'énergie par bit transmis avec succès en fonction de paramètres de la chaîne de transmission. Ces paramètres d'étude sont la distance entre les nœuds de communication, le type de canal, le bruit au niveau du récepteur, la taille des paquets, le type de codage de canal employé et la puissance d'émission. L'analyse des tracés nous permet de déduire le point de fonctionnement optimal de la transmission d'un point de vue de l'efficacité énergétique. Un module de gestion de l'énergie peut alors être défini en utilisant ces résultats et être utilisé au moment du déploiement du réseau et après chaque modification des conditions comme après des mouvements ou des pannes.

La figure 1 représente un résultat de simulation obtenu à partir des expressions analytiques décrites précédemment, pour une distance de 10 mètres, une puissance de bruit de -90 dB, une taille de paquets correspondant à la norme ATM. La puissance d'émission est en abscisse. Le

choix de cette distance a été fait dans l'optique de transmissions multi-étapes. En effet, l'objectif est de réaliser des communications sur de longues distances par plusieurs transmissions à courte distance, avec de faibles puissances d'émission, plutôt que de rares transmissions plus gourmandes en énergie. En Wi-Fi, le réglage le plus faible de la puissance d'émission (-20 dBm) permet d'assurer des taux d'erreurs raisonnables jusqu'à une distance d'environ 10 mètres, c'est pourquoi cette distance nous intéresse. La droite verticale représente la limite de Shannon, c'est à dire la limite théorique minimale de la puissance d'émission pour pouvoir avoir une transmission aussi fiable que nécessaire. Les trois autres courbes représentent l'énergie par bit transmis en fonction de la puissance du signal diffusé à l'antenne. Toutes les courbes ont le même comportement global. La partie gauche, pour des puissances d'émission faibles, est décroissante. Dans cette zone, les paquets sont souvent transmis avec des erreurs, et les retransmissions augmentent le temps nécessaire pour transmettre une quantité d'information donnée. La partie droite des courbes correspond à une puissance d'émission relativement plus élevée, qui permet de réduire le TEP. Les paquets nécessitant une ré-émission deviennent rares, mais, en contrepartie, la puissance instantanée du système en fonctionnement est plus élevée puisque l'amplificateur doit délivrer plus de puissance. Le bon compromis se situe au minimum de chacune des courbes. On voit sur l'exemple qu'il est possible de gagner de l'ordre de 30 % d'énergie en choisissant un codage de canal de type Viterbi et en se positionnant au point optimal comparé à une absence de codage de canal. En considérant cette fois une gestion optimale de la puissance d'émission comparée à un choix au pire cas dans lequel on maximiserait la puissance d'émission pour assurer la connectivité, on peut alors économiser 87 % d'énergie.

La figure 2 montre le même style de courbe pour une couche physique différente, dont le modèle a été extrait d'un composant du commerce. On peut constater que la mise en forme du signal consomme la plus grande part de l'énergie et prédomine sur les autres facteurs de consommation. Cette part fixe de consommation de fonctionnement et de mise en forme du signal est peu adaptée à des réseaux de capteurs. Il y a donc beaucoup à gagner à optimiser cette partie. Dans cette optique, de nouvelles normes émergent notamment IEEE 802.15.4 et Zigbee, redéfinissant les couches basses de réseaux mieux adaptées aux réseaux de capteurs.

6 Conclusion

Après avoir présenté la problématique des réseaux de capteurs, une architecture générique d'un élément d'un réseau de capteurs a été décrite. Dans le but d'optimiser sa puissance, nous avons établi des modèles de consommation utilisables en pratique et nous avons utilisé ces modèles pour caractériser un système de communication candidat pour un réseau de capteurs. Ensuite, nous avons élaboré des modèles théoriques liant les performances et la consommation de notre système, en fonction de paramètres et des besoins applicatifs. Nous utilisons ensuite ces modèles pour trouver les points de fonctionnement optimaux.

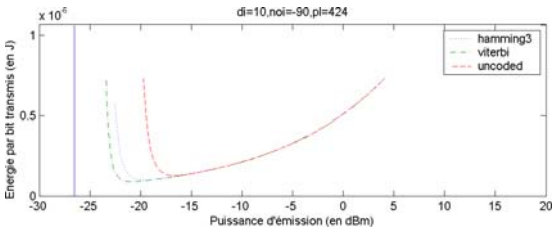


FIG. 1 – Énergie par bit transmis en fonction de la puissance d'émission, pour plusieurs types de codeurs/décodeurs et pour des paquets de type ATM, modèle de couche physique fin

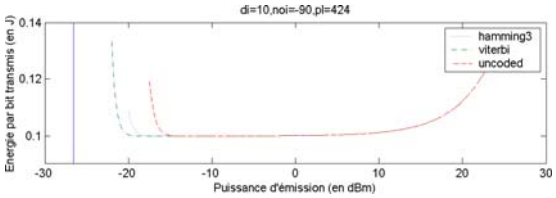


FIG. 2 – Énergie par bit transmis en fonction de la puissance d'émission, pour plusieurs types de codeurs/décodeurs et pour des paquets de type ATM, modèle de couche physique typique du commerce

Références

[1] Andrea Goldsmith and Stephen Wicker, "Design Challenges for Energy-Constrained Ad Hoc Wireless Networks," *IEEE Wireless Communications*, vol. 9, no. 4, pp. 8–27, August 2002.

[2] Daji Qiao, Sunghyun Choi, Amjad Soomro, and Kang G. Shin, "Energy-Efficient PCF Operation of IEEE 802.11a Wireless LAN," in *INFOCOM'02*. IEEE, 2002.

[3] Paul Lettieri, Christian Fragouli, and Mani Srivastava, "Low Power Error Control for Wireless Links," in *MOBICOM'97*. ACM, 1997.

[4] Chris Savarese, Jan Rabaey, and Jan Beutel, "Locationing in Distributed Ad-Hoc Wireless Sensor Networks," in *ICASSP*. IEEE, 2001.

[5] Jamal Al-Karaki and Ahmed Kamal, "Routing Techniques in Wireless Sensor Networks : A Survey," *IEEE Wireless Communications*, vol. 11, no. 6, pp. 6–28, December 2004.

[6] Rex Min, Seong-Hwan Cho, Manish Bhardwaj, Eugene Shih, Alice Wang, and Anantha Chandrakasan, "Power-aware Wireless Microsensor Networks," in *Power-aware Design Methodologies*, Massoud Pedram and Jan Rabaey, Eds., chapter 12, pp. 335–372. Kluwer Academic Publishers, 2002.

Un cluster de calcul hybride pour les applications de vision temps réel

Joël FALCOU[1], Jocelyn SÉROT[1], Thierry CHATEAU[1], Frédéric JURIE[2]

[1]LASMEA, UMR6602 UBP/CNRS, Campus des Cézeaux, 63177 Aubière, France.

[2]GRAVIR, INRIA/CNRS, 55 avenue de l'Europe, Montbonnot 38330, France

Joel.Falcou@lasmea.univ-bpclermont.fr Jocelyn.Serot,@lasmea.univ-bpclermont.fr
Thierry.Chateau@lasmea.univ-bpclermont.fr, Frederic.Jurie@inrialpes.fr

Résumé – On décrit l'implantation d'une chaîne de traitement et de reconstruction de points 3D en temps réel sur un cluster de PowerPC G5 biprocesseurs associant trois niveaux de parallélisme (SIMD grain fin, SMP et MIMD à passage de messages) et un dispositif dédié de diffusion de flots vidéo. Nous montrons que ce type d'architecture, permettant d'exploiter plusieurs formes de parallélisme, répond aux exigences des applications de vision temps réel en alliant puissance de calcul, facilité de programmation et extensibilité. Des résultats préliminaires sont fournis et discutés.

Abstract – We describe the implementation of an application performing real-time reconstruction of 3D points from stereo video streams on a cluster of G5 processors. The goal is to show that the combined use of a dedicated broadcasting mechanism for input streams and a three-level parallel programming model involving SIMD,SMP and MIMD/MP parallelism makes this type of platform a suitable solution for the implementation of real-time vision applications involving complex algorithms and requiring large amounts of computing power. Preliminary results are shown and discussed.

1 Introduction

La vision artificielle est une discipline exigeant de fortes puissances de calcul, particulièrement lorsqu'il s'agit de traiter des flots d'images en opérant à la volée des capteurs. Pour ces applications, les traitements doivent se faire en temps réel pour permettre aux systèmes d'interagir avec leur environnement (applications de robotique mobile, interfaces homme machine, etc.).

Conscient à la fois de l'importance de ce type d'applications et des contraintes qui y sont liées, les chercheurs du domaine ont utilisé diverses stratégies pour accroître la réactivité de tels systèmes : dégradation des algorithmes, utilisation de matériels spécifiques pour les traitements de bas niveau (FPGA, processeurs des cartes graphiques, etc.) ou d'architectures distribuées. Ces stratégies se sont parfois montrées efficaces en raison des propriétés spécifiques des images et des structures des algorithmes associés. Cependant, la progression constante et rapide de la puissance de calcul des ordinateurs personnels rend l'utilisation de solutions spécifiques de moins en moins attractive : les délais supplémentaires liés au temps de développement de ces solutions et de leurs logiciels, leur coût important, plaident en effet désormais en faveur de l'utilisation d'architectures fondées sur l'usage systématique de composants standards.

Dans le domaine du calcul scientifique hautes performances (simulation, modélisation, ..) cette classe de solution a vu son audience exploser ces dix dernières années avec la généralisation des clusters (grappes, fermes) de calcul bâties à partir d'ordinateurs individuels. Mais ce type de solution n'a pratiquement jamais été mis en œuvre dans le contexte d'applications de vision complexes devant opérer à la volée de capteurs. Le travail décrit ici vise donc à évaluer cette opportunité en proposant une architecture de type cluster dédiée aux applications de vision temps-réel. Nous décrivons l'architecture matérielle d'une telle plate-forme – en justifiant les choix technologiques effectués à partir des exigences du domaine d'application et de l'état de l'art – puis l'infrastructure logicielle associée, en insistant sur les aspects programmabilité. La validité de l'approche est démontrée en décrivant l'implantation d'une application de reconstruction de points d'intérêt 3D à partir d'un flot d'images stéréo et les performances obtenues avec une première implémentation.

2 Architecture du cluster

2.1 Architecture matérielle

L'architecture matérielle du cluster est présentée sur la figure 1 et comprend 14 noeuds de calcul. Chacun de ces noeuds est composé d'un PowerPC G5 biprocesseur[1] cadencé à 2 GHz et équipé de 1Go de mémoire. Les noeuds sont interconnectés par un réseau Ethernet Gigabit et reçoivent un flux vidéo stéréoscopique en provenance d'une paire de caméras numériques via un bus *FireWire IEEE1394a*[2]. Cette structure permet une diffusion simultanée et synchronisée des images sur tous les noeuds, éliminant ainsi le goulet d'étranglement classique consistant à redistribuer à l'ensemble des noeuds les images de travail via le réseau Ethernet. Dans notre configuration, la bande passante Ethernet reste disponible pour le transfert de message entre les noeuds. Le choix du processeur G5 est motivé par les points suivants :

– **La présence de l'extension multimédia Altivec [7].** Cette extension peut fournir des accélérations de l'ordre de 2 à 12 pour de nombreuses classes d'algorithmes bas et moyen niveau. [5]. Sans cette extension, la taille du

1. Apple XServe Cluster Node
2. Fonctionnant en mode isochrone à 400Mo/s

cluster nécessaire à l'obtention de l'accélération voulue (typiquement 50) aurait été prohibitive tant en termes de coût que de consommation.

– **La gestion native de l'Ethernet Gigabit**.
L'Ethernet Gigabit offre un des meilleurs ratio coût/performance. Des solutions plus onéreuses comme MyriNet ou InfiniBand ont été envisagées mais leur coût élevé aurait diminué sensiblement le nombre de noeud du cluster.

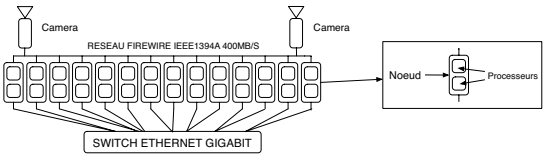


FIG. 1: *Architecture du cluster*

L'architecture proposée est suffisamment ouverte pour lui permettre de bénéficier des derniers progrès technologiques tout en supportant des modèles et des outils de programmation standard. En outre, l'utilisation de ces outils et modèles standards permet de rendre le cluster évolutif en permettant une mise à jour matérielle ponctuelle des noeuds.

2.2 Architecture logicielle

L'architecture logicielle du cluster est structurées selon trois niveaux.

Au niveau le plus bas, un **pilote *Firewire***[6] gère le transfert des images des caméras vers la mémoire du processeur en utilisant un canal DMA dédié. Pour l'utilisateur, les appels à ce pilote sont encapsulés dans une bibliothèque dédiée qui fournit des méthodes pour configurer les caméras, obtenir un pointeur vers l'image qui vient d'être reçue et synchroniser l'acquisition sur le cluster (sec. 3.3) Le second niveau est dédié à la mise en oeuvre des trois niveaux de parallélisme offerts par le cluster : un parallélisme de type SIMD à grain fin au sein de chaque processeur (Altivec), un parallélisme de type SMP entre les deux processeurs de chaque noeud et un parallélisme de type SPMD à passage de message entre les processeurs des différents noeuds. Le premier type de parallélisme est directement exploitable via l'interface C de l'extension AltiVec ou via EVE [5], une bibliothèque de vectorisation optimisée pour cette extension. Le parallélisme SMP nécessite la définition explicite des processus légers à exécuter sur chaque noeud via la bibliothèque PTHREAD. Enfin, l'application est décomposée en groupe de processus SPMD communiquant entre eux grâce à MPI. L'utilisation explicite de PTHREAD est due au fait que la gestion des pilotes Firewire n'est pas prévue pour être appelée de manière transparente par deux processeurs. Chaque noeud est vu comme un seul noeud Firewire par le pilote et comme deux noeuds par MPI. Cette dualité provoque des conflits de partage de ressources que MPI ne peut gérer facilement.

Enfin, un framework développé en C++ fournit, via l'utilisation de patrons prédéfinis – squelettes au sens de [8] – des fonctions de gestion des entrés/sorties vidéos et des mécanismes permettant des modifications dynamiques des modules algorithmiques de l'application, de développer et de mettre au point rapidement des applications pour le cluster. Créer une nouvelle application se résume alors à écrire des modules algorithmiques indépendants et à définir une stratégie d'exécution de ces modules. Ces stratégies peuvent êtres stockés sous forme de simple fichier texte et activés de manière dynamique au cours de l'exécution de l'application, soit de manière interactive ou pré-programmée. Le code présenté sur la figure 2 est un exemple simple d'utilisation de notre architecture logicielle. Dans cet exemple, nous construisons une application qui va traiter une image en provenances d'une caméra numérique et y appliquer un seuillage. Dans ce but, nous utilisons un squelette de parallèlisation classique : le *Scatter-Compute-Merge*.[3]

```
struct Data : public Singleton<Data> {
  REGISTER_ELEMENT(Frame, in);
  REGISTER_ELEMENT(Frame, out);
};

struct Seuil : public Task {
  virtual void run( int nodeID ) {
   Data::out() = where(Data::in() > 127, 255, 0);
  }
};

int main(int argc, const char** argv )
{
  Camera input( 30, res640x480xmono );
  Cluster c(argc,argv, MPI_COMM_WORLD);

  Task* a = SCM<RowSplit,Seuil,RowMerge>(0,5);
  c.assignTask(a);

  input >> Data::in();
  c.run(Data::Instance());
  return 0;
}
```

FIG. 2: *Une application de seuillage parallélisée sur notre cluster.*

La première partie de ce code consiste à définir une structure `Data` qui sera utilisée par tous les modules algorithmiques pour stocker les données de l'application. Dans cet exemple, `Data` contient deux instances de la classe `Frame` qui encapsule le concept d'image vidéo. Nous définissons ensuite la classe `Seuil` qui sera notre module de traitement. Cette classe fournit une méthode `run` qui contient le traitement proprement dit. Ici, le seuillage est effectué via la fonction `where` fournie par EVE . Cette fonction est une version SIMD de la construction `if then else`. L'application proprement dite définie une instance de la classe `Camera` ainsi que de la classe `Cluster`. Nous définissons alors la tache à effectuer sur ce `Cluster` en utilisant un des squelettes parallèles prédéfinis. `SCM` est un squelette de parallélisation qui va découper l'image d'entrée en bande, les diffuser sur un nombre fixe de noeuds de calcul, appliquer une opération et récupérer les résultats de ces calculs afin de reconstruire l'image de sortie. La fonction `SCM` prend en paramètres les classes définissant les stratégies de découpages et de fusion (`RowSplit` et `RowMerge`), celle définissant les opérations à effectuer (ici `Seuil`) et le nombre de noeuds de calculs (ici 5). Une fois cette tache définie, l'application commence à acquérir des images de la caméra et à les traiter via le `Cluster`.

3. Pour un souci de brièveté, nous avons volontairement omis les directives d'inclusions et les déclarations de *namespace*.

3 Application

Nous avons évalué cette plate-forme à partir d'applications vidéos temps réel pour démontrer que son architecture matérielle (double bus : Ethernet pour les communications et *Firewire* pour la diffusion vidéo) et logicielle (modèle de programmation parallèle à trois niveaux : SIMD, SMP, MP) convient aux besoins de cette classe d'applications tant en termes de performance que de flexibilité.

3.1 Description de l'algorithme

Nous avons implanté les premières étapes d'une chaîne de reconstruction et suivi 3D utilisant la vision stéréoscopique. L'application décrite ici génère une ensemble de points 3D qui seront par la suite traités par un algorithme d'interprétation et de suivi que nous ne décrivons pas ici. Cet algorithme peut se décomposer en quatre étapes :

1. **Rectification de l'image (RECTIF)** : Cette étape consiste à appliquer une homographie aux images d'entrée en mettant en correspondances les droites épipolaires et les lignes de pixels de l'image[9]. Cette étape permet ensuite de simplifier la recherche des correspondances en limitant cette recherche à une ligne de pixels.
2. **Détection de point d'intérêts (DETECT)** : Un détecteur de coins de Harris et Stephen [4] est utilisé pour extraire des points d'intérêts qui seront autant de germes pour l'étape de mise en correspondance
3. **Mise en correspondance (MATCH)** : Chaque point d'intérêt de l'image gauche (resp. de l'image droite) est mis en correspondance avec un point d'intérêt de l'image droite (resp. gauche) en utilisant un recherche du maximum de vraisemblance par corrélation centrée réduite.
4. **Reconstruction 3D (BUILD)** : On effectue une triangulation [10] des paires de points 2D correctement mis en correspondance.

Nous avons choisi cet algorithme pour trois raisons : le grand nombre d'opérations par image, la diversité de ces opérations et le fait qu'il soit à la base de nombreuses autres applications comme la navigation de robot mobile ou l'interface homme/machine. Un exemple de reconstruction est donné sur la figure 3.

3.2 Stratégie de parallélisation

L'ensemble des quatre étapes décrites ci-dessus est exécuté en découpant l'image source en cinquante bandes horizontales. Chacune de ces bandes est ensuite distribuée à un processeur[4]. Comme le nombre de points d'intérêts dans chaque bande ne peut être prédit et donc que la charge de calcul associée peut fluctuer de manière significative d'une bande à l'autre, le squelette `SCM` présenté dans l'exemple 2 est inadapté. Un autre squelette de parallèlisation classique, `Farm`, fournit une méthode simple d'équilibrage de charge. Dans ce squelette, un noeud maître distribue de manière dynamique des bandes d'image à une réserve d'esclaves. Dès qu'un des esclaves à terminé d'effectuer sa tâche sur une bande, le maître lui en fournit

4. Ce qui est possible grâce à l'étape de rectification qui force les points d'intérêts à s'aligner sur les lignes de pixels de l'image.

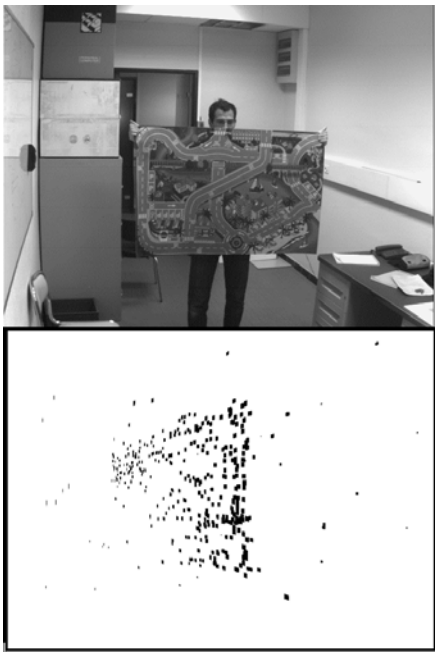


FIG. 3: *Résultat de la reconstruction 3D*

une nouvelle. Au final, les ensembles de points 3D en provenance de tous les esclaves sont fusionnés et rapatrié sur le processeur maître. PTHREAD permet de distribuer sur les processeurs de chaque noeud une bande provenant de l'image droite et la bande correspondante de l'image gauche. A un moment donné, chaque noeud traite ainsi des bandes d'images provenant bien d'une même portion des images initiales. Les étapes de détection, de mise en correspondance et de reconstruction mettent en jeu des opérations iconiques ou géométriques régulières (convolution, produit scalaire) Elles sont donc vectorisées par EVE [5] pour profiter de l'accélération fournie par l'unité SIMD du PowerPC G5.

3.3 Synchronisation des Cameras

Pour assurer une grande précision dans la reconstruction, l'algorithme doit être appliqué à des paires d'images décrivant le même événement. Si les images droites et gauches de la paire sont désynchronisées, la mise en correspondance et la reconstruction seraient de mauvaise qualité. En pratique, cela implique que l'acquisition des images droite et gauche sur un noeud doit être synchronisé. De même, la diffusion des images sur tous les noeuds du cluster doit être synchronisé. Cette double synchronisation est réalisée comme suit :

- **Synchronisation intra-noeud** : le pilote Firewire CFOX est capable par un système de datation des images et de triple-buffering de fournir des images synchronisées pour un nombre arbitraire de caméras. Cette synchronisation est totalement transparente pour l'utilisateur qui n'a pas à se soucier de connaître la datation des images. L'appel de la méthode d'acquisition gére ces problèmes de manière transparente.
- **Synchronisation inter-noeud** : En utilisant une synchronisation par passage de messages (via `MPI_Barrier`),

il est possible de forcer les noeuds à s'attendre et à se synchroniser pour acquérir la nouvelle image depuis le bus Firewire.

4 Résultats

Des résultats préliminaires, obtenus sur des images 640×480 ×8bits sont donnés dans le tableau 1. Les temps de calculs de chaque étape ainsi que le temps d'exécution total de l'algorithme sont donnés pour un nombre de processeurs variable en ms[5]. La première colonne donne pour référence le temps d'exécution pour une version séquentielle de l'algorithme sur un seul processeur.

Etape	n=1	n=2	n=4	n=8	n=16	n=26
RECTIF	246	139,1	70,5	36,1	19,5	13,1
DETECT	262	80,1	40,5	20,6	11,2	7,1
MATCH	304,2	180	91,5	47,4	22,4	13,8
BUILD	180	100	53	27,5	18,2	12,0
TOTAL	992,2	479,2	244,6	122,6	68,2	42,9

TAB. 1: *Résultats préliminaires de recon-3D*

La latence minimale est de 42.9ms, obtenue avec 26 processeurs, soit une cadence de 24 images par seconde (les caméras fonctionnant à 30 images/secondes). L'accélération est alors de 23.1 (soit une efficacité de 88%[6]). Deux facteurs expliquent ces résultats encourageants. Premièrement, grâce à la diffusion automatique des images sur le bus Firewire, l'application montre un ratio calcul / communication très élevé (entre 0.90 et 0.92). Deuxièmement, la parallélisation grain fin SIMD est exploité de manière efficace dans les étapes DETECT et MATCH. Sans AltiVec, la cadence tombe à 19 images/secondes, démontrant que la couche SIMD fournit une accélération non négligeable. Signalons que ces résultats préliminaires ont été obtenus avec une implantation naïve des algorithmes. Un grand nombre de paramètres et de modèles peuvent êtres affinés pour permettre une utilisation plus intensive de EVE et donc un gain final plus important. Le nombre d'esclaves dans le squelette `Farm`, la taille et le nombre de bande sont autant de paramètres dont les variations ont potentiellement un impact sur les performances de l'application.

5 Travaux en rapport

Pour la reconstruction 3D temps réel, une solution classique consiste à reconstruire l'enveloppe visuelle d'un objet à partir de silhouettes en utilisant N caméras chacune connectées à un noeud. Les caméras sont synchronisées afin d'assurer que chaque paire PC/Camera traite une vue du même événement. La parallélisation de ces algorithmes utilise un simple pipeline ou chaque étape de l'algorithme est exécutés en parallèle sur chaque noeud. Wu et Matsuyama [1] ou Franco et al. [2] aboutissent ainsi à des résultats temps réel avec un nombre restreint de cameras et de noeuds. Ces approches différent de celle présentée ici tant au niveau des stratégies de parallélisation que de l'utilisation des caméras.

Yoshimoto et Arita[3] décrivent quant à eux un cluster de PC utilisant un réseau Firewire pour le traitement d'images en temps réel. Ce cluster utilise un bus Firewire IEEE-1394a à la fois pour transmettre les images à traités et les communications inter-noeuds. Des applications temps réels sont montrées mais avec cette architecture, l'utilisation du bus devient rapidement un goulet d'étranglement, ce qui limite sérieusement les performances et l'extensibilité du cluster.

6 Conclusion

Cet article montre qu'un cluster doté d'un mécanisme de diffusion vidéo performant et offrant plusieurs niveaux de parallélisme permet de satisfaire les contraintes temporelles d'applications de vision temps réels, ce que nous avons démontré en développant une application de reconstruction 3D à partir d'un flux vidéo stéréoscopique. À notre connaissance, il s'agit de la première utilisation d'une telle architecture pour permettre l'exécution d'une application temps réel de cette complexité. Ces résultats restent néanmoins préliminaires. À court terme, la finalisation de la chaîne de reconstruction ainsi que le développement d'une application de suivi 3D est prévue.

Références

[1] Wu X., Matsuyama T. Real-time active 3d shape reconstruction for 3d video. In Proc. of 3rd International Symposium on Image and Signal Processing and Analysis, pages 186-191, 2003.

[2] Franco J.-S., Menier C. and Boyer E. A Distributed Approach for Real-Time 3D Modeling.CVPR Workshop on Real-Time 3D Sensors and their Applications, 2004

[3] Yoshimoto H. and al. Real-Time Image Processing on IEEE1394-based PC Cluster. In 15th International Parallel and Distributed Processing Symposium, 2001

[4] Harris C. and Stephens M. A combined corner and edge detector. In 4th Alvey Vision Conference, 1988, pp. 189-192.

[5] Falcou, J., Serot, J - E.V.E., An Object Oriented SIMD Library. In Practical Aspects of High-level Parallel Programming,ICCS 2004, pp. 323-330(3)

[6] Falcou, J. - CFOX - A Firewire Camera Driver for OS X `http://wwwlasmea.univ-bpclermont.fr/Personel/Joel.Falcou/software/cfox`

[7] Ollman I. AltiVec Velocity Engine Tutorial. `http://www.simdtech.org/altivec`. Mars 2001.

[8] Cole M. - Algorithmic Skeletons. In Research Directions in Parallel Functional Programming, Springer-Verlag, 1999.

[9] Fusiello A.,Trucco E., Verri A. - A compact algorithm for rectification of stereo pairs. In Machine Vision and Application, vol. 12, no. 1, pp. 16–22, 2000

[10] Hartley R. I. and Zisserman A. - Multiple View Geometry in Computer Vision. Cambridge University Press - ISBN 0521623049, 2000.

5. Les deux processeurs dédiés à l'affichage et à la synchronisation globale ne sont pas pris en compte.

6. 82% en tenant compte des deux processeurs de contrôle et d'affichage.

Implantation d'un récepteur Multi-Utilisateurs appliqué à la liaison montante UMTS

M. TOUZRI[1], S. SAOUDI, S. KEROUEDAN, P. ADDE

[1] GET / ENST-Bretagne / PRACOM / CNRS-TAMCIC. ENST-Bretagne, Technopole Brest-Iroise, CS 83818, 29238 BREST, France

Makram.Touzri@enst-bretagne.fr

Résumé – Cet article présente les étapes préparatoires à l'implantation d'un récepteur multi-utilisateurs appliqué à la liaison montante du système UMTS. Tout d'abord, la comparaison des résultats des simulations d'une détection mono-utilisateur à base de Filtre Adapté avec une détection multi-utilisateurs à suppression d'interférence successive (SIC) dans le cas synchrone permet de valider les choix faits. Puis dans un contexte asynchrone à trajets multiples, il est possible de comparer les performances d'un récepteur RAKE à celles d'un récepteur SIC/RAKE.

Les simulations sont faites avec un traitement à valeur réelle d'une part et à valeur entière d'autre part pour permettre l'étude et l'optimisation de la quantification.

Abstract *– In this paper, we introduce the first step for the implementation of a multi-user detection for the UMTS uplink channel. We discuss results of simulation performed with the use of a single detection and a multi-user detection. We compare performances of a Matched Filter vs. Successive Interference Cancellation receiver in the synchronous case and a RAKE vs. a SIC/RAKE receiver in the asynchronous and multi-channel case. Simulation results are compared in both real and integer computing.*

1. Introduction

La troisième génération de téléphonie mobile, l'UMTS (*Universal Mobile Telecommunication System*) est basée sur l'utilisation de l'Accès Multiple par Répartition de Code à large bande (AMRC) ou le *Wideband Code Division Multiple Access* (W-CDMA) à étalement direct de spectre à 3.84 Mcps en modulation QPSK pour atteindre des débits théoriques avoisinant les 2 Mbps [1].

Cependant, l'Interférence d'Accès Multiple (IAM) réduit fortement le nombre d'utilisateurs actifs. Ceci est dû principalement aux trajets multiples du canal de transmission qui dégradent l'orthogonalité des codes utilisés.

Pour accroître le nombre d'utilisateurs actifs, des résultats acceptables sont obtenus avec une détection mono-utilisateur et en perfectionnant le décodage canal [2]. Mais cette approche présente encore des limites vis à vis des taux de charge et débits élevés.

Dans ce contexte, des méthodes de détection multi-utilisateurs ont été élaborées [3] et ont atteint de bonnes performances théoriques. Elles présentent cependant, une complexité accrue en terme d'intégration sur circuit.

Dans cet article, on étudie l'implantation d'un récepteur multi-utilisateurs à suppression d'interférence successive (*Successive Interference Cancellation*) (SIC) appliqué au cas d'une liaison montante du système UMTS.

On s'intéresse à la comparaison des performances et à l'optimisation du traitement à valeur réelle et du traitement à valeur entière de la structure proposée [4] dans le cas d'un canal synchrone.

Ensuite, on traite l'architecture SIC/RAKE donnée dans [5] et on présente les premiers résultats en traitement à valeur entière dans le cas du canal asynchrone à trajets multiples.

Dans cette première approche, on étudie la détection multi-utilisateurs SIC et SIC/RAKE sans l'utilisation de codage canal.

2. Suppression d'interférence successive

2.1 Canal synchrone

Le SIC est un processus itératif de détection multi-utilisateurs. Sa structure est composée d'une série d'unités de suppression d'interférence (*Interference Cancellation Unit* ICU). Chaque ICU est composé de deux blocs (Fig. 1.b). Le premier est un Filtre Adapté (FA). Sa sortie est injectée dans le deuxième bloc qui est un Emetteur Local. Ce dernier re-étale la sortie du FA pour générer une estimation du signal émis par l'utilisateur *k*. La sortie de ICU est un signal résiduel *e* résultant de la différence du signal reçu et du signal généré. Ce processus est répété pour tous les utilisateurs *K* et pour les *M* étages (Fig. 1.a).

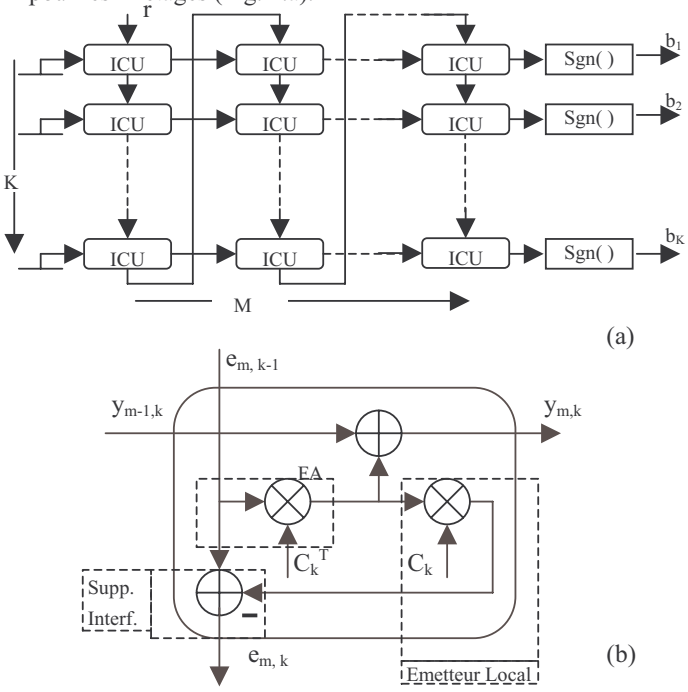


Fig.1. SIC : (a) Structure générale. (b) Structure de l' ICU

2.2 Simulation

Dans la simulation du cas synchrone, et pour permettre d'introduire un défaut d'orthogonalité, seuls des segments de codes de *Gold* de la longueur du facteur d'étalement, sont utilisés pour la discrimination entre utilisateurs.

Les paramètres des simulations sont : le nombre d'étages *M=5*, le facteur d'étalement *Nc=16* et le nombre d'utilisateurs *K=8*, ce qui signifie un taux de charge égal à *K/Nc=50%*. Les puissances reçues des utilisateurs sont supposées égales.

Le Taux d'Erreurs Binaire (TEB) est une moyenne pour tous les utilisateurs. Cependant, l'affichage d'un TEB par utilisateur pourrait montrer une différence entre le premier et le dernier utilisateur traité. En effet, celui-ci profite pleinement de la suppression d'interférence successive.

Les résultats obtenus (Fig. 2) montrent la convergence du SIC qui dépasse dès le premier étage, la performance du FA. Le dernier étage du SIC a tendance à s'approcher de la courbe de référence qui illustre le TEB pour la transmission d'un seul utilisateur.

La valeur optimale de *M* dépend du taux de charge *K/Nc*.

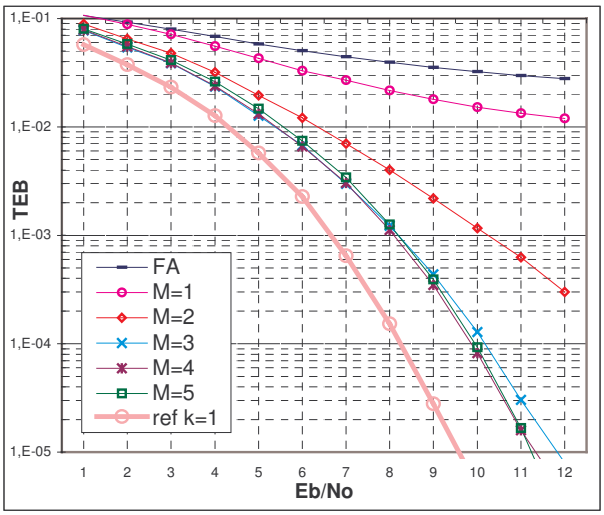


Fig. 2 : Performance du SIC en fonction du nombre d'étages *M*.

3. Quantification

3.1 Processus

L'implantation de la structure du SIC est précédée d'une adaptation de l'algorithme de traitement et son passage en calcul à valeur entière au lieu du calcul à valeur réelle. Ce qui se traduit par un calcul dans un intervalle à valeur entière signée et écrêtée à $\pm 2^{n-1}$ où n est le nombre de bits de quantification. La Fig. 3 illustre ce passage ainsi que la description du processus de la simulation avec une branche en calcul « réel » et une en calcul « entier » pour permettre une comparaison directe des performances. Le facteur d'échelle *F* utilisé est calculé en fonction de *K*, *Nc* et *n* qui sont connus au niveau du récepteur.

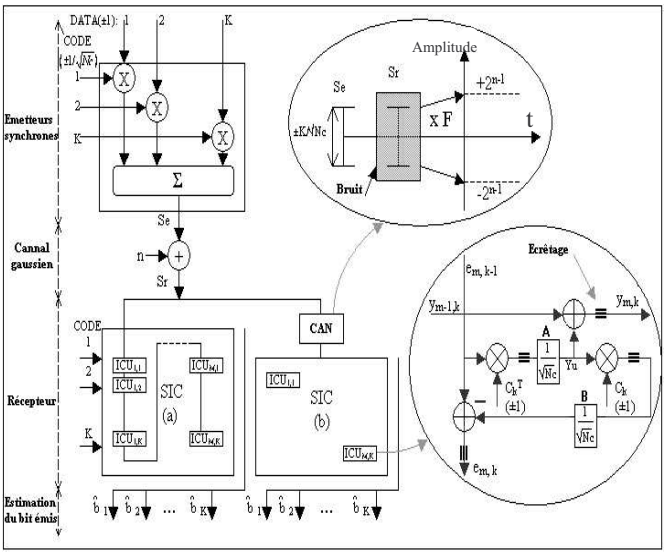


Fig. 3 : Processus des simulations : (a) réel. (b) entier.

3.2 Simulations

Les résultats de la Fig. 4 permettent de se fixer un nombre de bits de quantification *n* constant durant tout le processus du SIC et égal à 8 pour avoir des performances quasiment égales à celles obtenues en calcul à valeur réelle et ceci dans un intervalle de rapport signal à bruit donné.

Toutefois, des modifications peuvent être apportées pour réduire le nombre *n* à l'intérieur de chaque ICU sans dégrader les performances globales du SIC jusqu'à une limite en terme de rapport signal à bruit (autour de 8 dB).

En effet, la Fig. 3. montre une division par $\sqrt{Nc}$ dans un but de normalisation. Ceci est dû aux valeurs en ± 1 prises par les codes d'étalement dans le traitement à valeur entière contrairement au cas réel où ils prennent des valeurs en $\pm 1/\sqrt{Nc}$ [4]. Cette normalisation se traduit, dans le calcul à valeur entière, à faire une division par $\sqrt{Nc}$ à chaque multiplication par le code (deux fois par ICU).

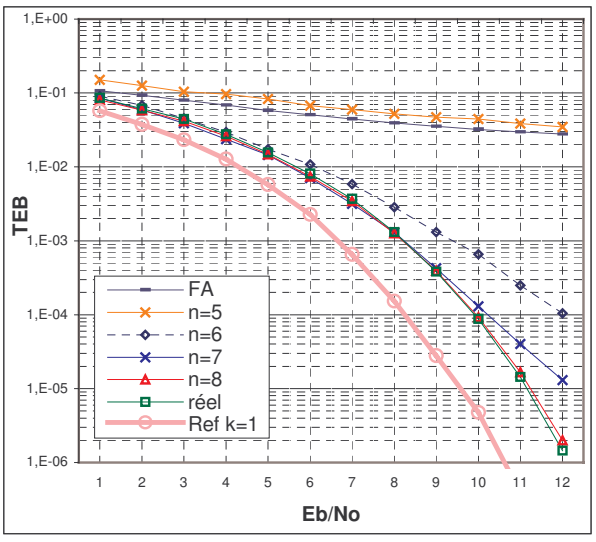


Fig. 4 : Résultats réel-entier ; *K*=8, *Nc*=16, *M*=5.

Cependant ces deux divisions partielles peuvent être groupées en une seule division par *Nc*. Ceci permet de simplifier la division en binaire et se traduit dans le cas général où $Nc=2^c$, par un décalage à droite de *c* cases.

«$Nc = 256 / 2^j$, avec j = 0,…,6», ce qui donne à *Nc* les valeurs {4, 8, 16, 32, 64, 128 et 256}.

Le choix de l'emplacement de cette division groupée au niveau de ICU donne des performances égales (réel-entier) dans la limite de 8 dB et favorise le point A (Fig.3).

En effet, les performances restent identiques avec l'augmentation du nombre d'étages *M* et les dynamiques des sorties du FA et de l'émetteur local sont petites et peuvent être quantifiées avec un nombre de bits plus petit.

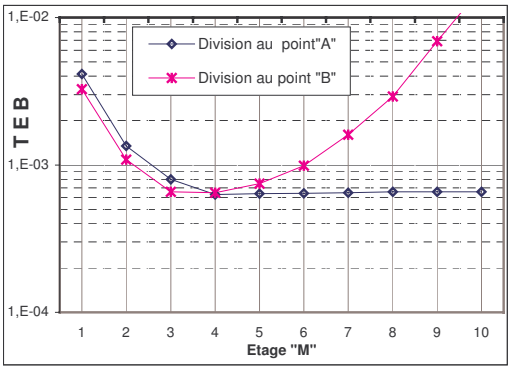


Fig. 5 : Résultat SIC à 8dB selon modification

En pratique et avec un nombre de bits de quantification global égal à 8, on peut réduire la quantification de la sortie du SIC ($Y_{m,k}$) à 6 bits et celle du filtre adapté (Y_u) à 4 bits.

Ces modifications apportées dans le traitement à valeur entière, permettent d'aller jusqu'à la suppression de cette division et de faire une soustraction d'interférence dure (Hard) dans l'ICU. La Fig. 6. montre le processus suivi.

Il est à noter l'addition d'une zone de Masquage (Mask) pour annuler la suppression d'interférence dans le cas de faibles valeurs de sortie du FA où on a une estimation médiocre du symbole envoyé. La Fig. 7. illustre l'avantage de cette modification, dans la limite de 8 dB, appliquée au cas « entier » et aussi au cas « réel » (Fig. 7. réel modif.) pour permettre une comparaison.

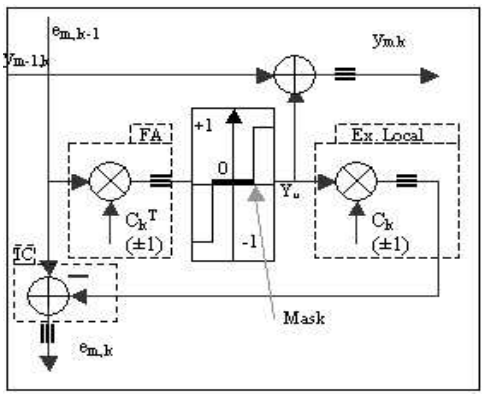


Fig. 6 : ICU modifié : Structure.

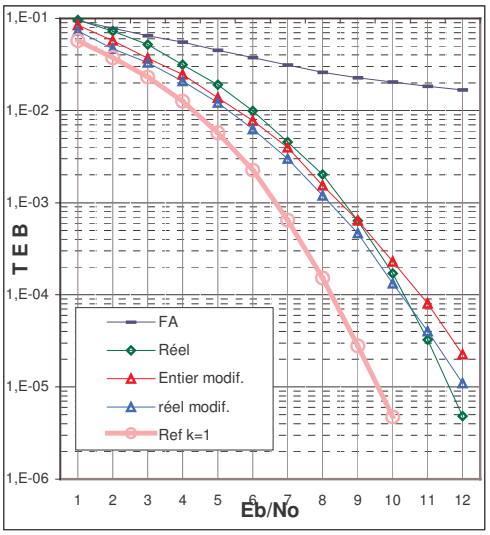


Fig. 7 : ICU modifié : Résultats

Toutefois, il est à noter que ces modifications affectent la robustesse du système vis à vis de l'effet d'éblouissement.

Cependant, le but est de réduire au maximum la complexité du SIC pour mieux maîtriser son comportement et permettre sa combinaison avec un décodage canal.

Les simulations faites dans le cas synchrone, permettent d'approuver une utilisation convenable du SIC avec 8 bits et avec 3 étages dans un rapport signal sur bruit inférieur ou égal à 8 dB.

4. SIC / RAKE

4.1 canal asynchrone à trajets multiples

L'architecture générale du SIC reste la même, tandis que l'ICU est modifié pour englober un récepteur RAKE à *L* branches. Celui-ci se charge d'aligner les signaux issus des trajets multiples. Ces signaux sont composés de deux voies en quadrature de phase, ce qui change tout le traitement en complexe. Les paramètres des simulations sont les suivants :

- des codes de *Gold* $C_k^{(s)}$ sont utilisés pour la discrimination entre utilisateurs,
- des codes de *Walsh* $C^w_{[I]k}, C^w_{[Q]k}$ sont utilisés pour séparer respectivement la voie de données « I » et la voie de contrôle « Q »,
- le facteur d'étalement *Nc* de la voie de contrôle est égal à 256,
- le *Nc* de la voie de donnée varie de 4 à 256,
- le nombre de trajets par utilisateur *L*=6,
- le modèle de *Clarke* est utilisé pour simuler un canal de Rayleigh et fournir les coefficients $C_{k,l}$ et les retards des trajets à partir, entre autre, du nombre de trajets et de la fréquence Doppler.

Dans la référence [5], les retards (en *Chips*) sont compris dans l'intervalle [0, 2×*Nc*] correspondant à un maximum de *Nc* retards entre les utilisateurs et *Nc* retards entre les trajets, d'où l'utilisation d'une fenêtre glissante dans le traitement pour optimiser les performances. La structure d'une unité de suppression d'interférence dans le cas asynchrone à trajets multiples est illustré à la Fig. 8.

4.2 Simulations

Les simulations sont réalisées avec n=8 bits et avec un écrêtage en entrée et en sortie de ICU.

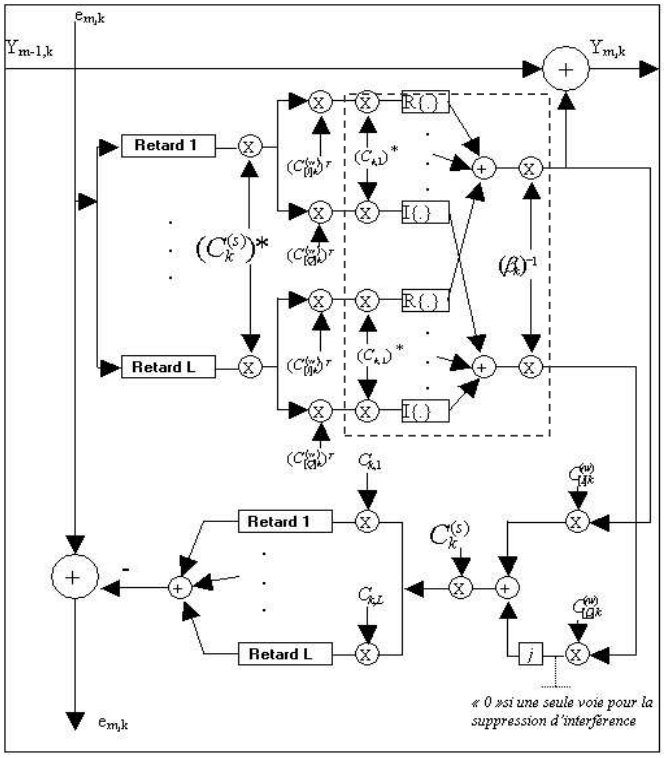


Fig. 8 : Structure d'une unité SIC/RAKE.

La Fig.9 montre les performances obtenues respectivement avec le traitement à valeur réelle, le traitement à valeur entière sans écrêtage interne et le traitement à valeur entière avec écrêtage interne. L'écrêtage externe au niveau des entrées et sorties de ICU est toujours effectué.

La performance du détecteur mono-utilisateur, correspondant à l'utilisation d'un simple RAKE par utilisateur, est la même en traitement à valeur réelle et entière. En effet, la dégradation due à la quantification ne se voit pas puisque l'estimation du symbole émis est une simple décision dure sur la sortie du RAKE. Par contre l'utilisation du SIC, qui est un processus itératif, propage l'erreur de quantification et affecte le signal résiduel qui est la sortie de chaque ICU. D'où l'utilisation d'un nombre de bits de quantification plus grand à l'intérieur de chaque ICU du SIC pour améliorer la précision sur l'estimation des bits reçus.

Le bloc en pointillé sur la Fig. 8 délimite l'utilisation de n=16 bits pour la quantification et se situe juste avant la normalisation par le facteur β_k et après l'utilisation des coefficients du canal.

En conséquence, et pour réduire le nombre de bits de quantification interne, la normalisation par β_k peut être placée sur chaque branche du RAKE. Ceci donne lieu à « L » normalisations différées par β_k et permet de réduire la quantification à 12 bits avec une légère dégradation des performances. Celles-ci sont illustrées par la courbe « Entier écrêté » dans la Fig. 9.

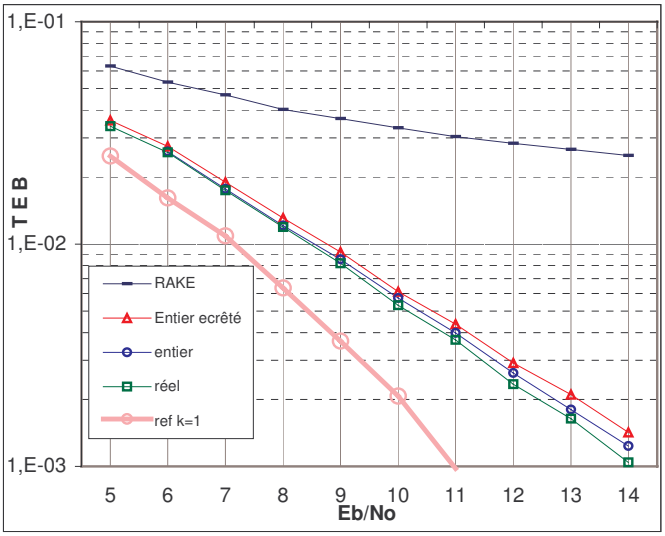


Fig. 9 : SIC/RAKE
K=8, NcI=32, NcQ=256, M=3. SIC sur une seule voie.

5. Conclusion

Dans cet article, on a préparé l'implantation d'un récepteur multi-utilisateurs appliqué à la liaison montante du système UMTS. Pour cela, on est passé du traitement à valeur réelle au traitement à valeur entière dans le cas synchrone et dans le cas asynchrone à trajets multiples. Les résultats obtenus sont prometteurs et il reste à optimiser l'algorithme du détecteur.

Cette première étape nous a permis de maîtriser la détection multi-utilisateurs, ce qui facilitera son association dans la prochaine étape avec un décodage canal à base de turbo-codes. On vise la conception d'une structure combinée performante avec un nombre réduit d'itérations au niveau du décodage canal et de la détection multi-utilisateurs.

L'objectif principal est l'étude et l'implantation d'une structure intégrant deux processus itératifs à savoir la Détection Multi-Utilisateurs et le Turbo-décodage pour des systèmes de réception à base de CDMA.

Références

[1] TS 25.213 UMTS. *spreading modulation (FDD)* 3GPP Technical specification, June 2001.

[2] TS 25.213 UMTS. *Multiplexing and channel coding (FDD)*. 3GPP Technical specification, June 2001.

[3] S. Verdu. *Multi-user detection*. Cambridge University Press 1998.

[4] M. Ammar, T. Chonavel, S. Saoudi. *Nouvelle structure SIC iterative avec estimation conjointe du canal pour la liaison montante du système UMTS-FDD*. GRETSI, Toulouse, France, Septembre 2001.

[5] M. Ammar. *Détection multi-utilisateurs pour l'UMTS*. Thèse, département Signal et communication, ENST-Bretagne, juin 2002.

[6] L.C.A Hui and K. Ben Letaief. *Successive Interference Cancellation for multi-user asynchronous DS-CDMA detectors in multi-path fading links*. IEEE Transactions on Communications, March 1998.

Un Modèle Hiérarchique pour les Architectures Reconfigurables en Radio Logicielle

Jean-Philippe Delahaye, Jacques Palicot, Pierre Leray

IETR/SUPELEC, Campus de Rennes, Avenue de la Boulais, Cesson-Sévigné, BP 81127, France
Jean-Philippe.Delahaye@supelec.fr,Jacques.Palicot@supelec.fr,Pierre.Leray@supelec.fr

Résumé – Cet article présente un modèle hiérarchique de la gestion de configuration d'architecture reconfigurable dédié aux systèmes multi-standards. A partir d'une étude au niveau algorithmique nous proposons de décrire un modèle fonctionnel de gestion de configuration tenant compte des besoins applicatifs. L'analyse algorithmique porte sur une chaîne d'émission multi-standards UMTS/FDD Uplink, GSM Uplink et 802.11g mode OFDM. Le but du modèle présenté est de minimiser les ressources matérielles à reconfigurer à chaque nouveau changement de contexte applicatif. Par le contrôle de la reconfiguration partielle et dynamique de la chaîne, ce modèle hiérarchique est bien adapté à la gestion des ressources hétérogènes d'un système en correspondance avec les besoins d'adaptabilité de l'application multi-standards.

Abstract – This paper presents a functional model based on a hierarchical architecture template meeting with Software Defined Radio System requirements (SDR System). The concepts and mechanisms required to design future reconfigurable system architectures are addressed in the paper. The definition of the new features requested in such architectures is based on a thorough multi-standards functional analysis of a transmitter (i.e. UMTS/FDD Uplink, GSM Uplink, and 802.11g OFDM mode). In addition to a classical datapath for processing a configuration management path has been integrated. This model aims at helping the design and management of a heterogeneous dynamically reconfigurable hardware architecture for SDR terminals.

1 Introduction

Les techniques radio logicielle introduitent par Mitola [7] visent à offrir un accès à un large choix de standards de radiocommunications sur une architecture matérielle unique. Une grande fléxibilité de l'architecture du système radio logicielle est donc nécessaire afin de répondre à la diversité des traitements à exécuter. Cette fléxibilité est assurée par l'utilisation d' architectures reconfigurables [1]. La diversité des traitements implique l'utilisation d'architecture matérielles hétérogènes [6] afin que l'exécution soit optimisée. Diversité des traitements et hétérogénéité de l'architecture induisent une grande variété de configurations. Il est donc nécessaire de définir une architecture de gestion et de contrôle de configurations adaptée. Les différents besoins en termes de gestion de configuration d'un terminal radio logicielle ont largement été justifiés dans la littérature, comme par exemple par Gultchev et al. dans [2]. De plus, suivant les besoins d'adaptabilité du terminal, les configurations sont de natures différentes, comme le décrit Kountoutis et al. dans [5]. Par exemple, la reconfiguration intervient lors d'un changement de standard, de mode, mais aussi en vue de la correction ou l'amélioration des performances d'algorithmes. Ceci correspond à différents niveaux de granularité. Le problème se pose de gérer cette multi-granularité de contexte en correspondance avec une gestion efficace des ressources matérielles dont la nature est hétérogène, par exemple bit-level configuration des FPGAs et word level configuration des DSPs. Notre modèle hiérarchique permet de faciliter cette gestion de configuration multi-granularité et ainsi minimiser les ressources à reconfigurer à chaque changement de contexte afin d'assurer une reconfiguration dynamique.

Ce papier s'appuie sur des travaux ayant commencés par une étude algorithmique des traitements bande de base effectués dans les chaînes d'émissions des 3 standards de radiocommunications (UMTS, GSM,WLAN 802.11g). Ces standards ont été choisi pour la grande diversité des traitements offerte. Les applications visées nous ont conduit à choisir une architecture orientée flot de données présentée dans le paragraphe suivant. L'étude algorithmique est détaillée dans les parties 2 et 3, elle a mené à la proposition du modèle hiérarchique présenté dans la section 4.

1.1 L'approche Config-Data Path

Le domaine des architectures reconfigurables a déjà largement retenue l'approche flot de données pour les applications télécoms, R. Hartenstein dans [3] en liste les principaux projets. Face à la diversité des applications en radio logicielle et les contraintes de surface, de rapidité de calcul et de basse consommation, chaque type de fonction nécessite une implémentation optimisée appelant des traitements spécialisés. La reprogrammation de ressources matérielles hétérogènes doit donc être pris en charge. Ces raisons nous ont conduit à associer à chaque bloc de traitement, un gestionnaire de configuration dédié, afin de spécialiser la gestion de reconfiguration. D'autres approches de gestion de configuration d'architecture orientée flot de donnée ont été proposées notamment par Srikanteswara et al. [9] dans un modèle en couche. Dans cette proposition, les données de configuration et les données à traiter sont transmises par le même chemin, elles sont donc

encapsulées et augmentées d'un en-tête les différenciant. Afin d'éviter ce surcoût de transmission d'un en-tête et l'ajout d'un protocole nous proposons de séparer les 2 flux de données, qui ont de plus des débits différents. La figure 1 illustre ce modèle basé sur la séparation de la voie de traitement et de la voie de configuration.

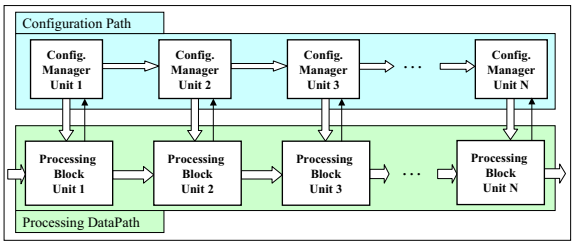


FIG. 1 – Functional Modeling of the Configuration-Processing Datapath

Avantages de l'approche : La séparation des 2 types de flux de données permet de dissocier plus aisément la partie contrôle et le traitement des données. Chaque Configuration Manager Unit (CMU) est associé à un bloc de traitement PBU (Processing Block Unit). Les ressources materielles des PBUs étant spécifiques à un type de traitement, chaque CMU est spécialisé dans la gestion de ces ressources cela permet de simplifier la conception de chaque entité de gestion et d'optimiser leur utilisation. L'architecture distribuée en composants CMUs/PBUs, permet une gestion efficace de la reconfiguration partielle. L'estimation du temps de reconfiguration de chaque PBU par les CMUs et la transmission de ces informations de configurations à travers le chemin de configuration permet de garantir le fonctionnement des reconfigurations dynamiques.

2 Cas de changement de contexte

L'adaptation d'un système radio logicielle à un environnement applicatif multi-standards supportée par une plate-forme d'exécution unique passe par l'utilisation de composants reconfigurables. Bien qu'un changement d'application ait un impact sur l'ensemble des couches protocoles de l'application de transmission, nous restreignons le terme changement de contexte à la couche physique. Afin d'optimiser l'adaptabilité du système reconfigurable, il faut diminuer le coût d'une demande de changement de contexte, en minimisant les ressources à reconfigurer. La première étape d'optimisation de la reconfiguration passe par une analyse au niveau algorithmique des différents traitements communs entre les différentes applications. A ce niveau, nous pouvons noter les travaux de R. Hoshyar et al. [4] proposant une méthode pour réaliser cette analyse des applications multi-modes. Par ailleurs, sur la définition et la proposition d'opérateurs communs efficaces et les études de parametrisation, nous nous reportons aux travaux de J. Palicot et al. [8]. Lors d'une reconfiguration optimisée, seul les blocs de traitements spécifiques au nouveau contexte par rapport au précédent se verront modifiés afin d'éviter un rechargement total du contexte d'exécution. Afin d'augmenter le taux de réutilisation des blocs de traitements (ou opérateurs communs) ceux-ci seront paramétrables. Enfin, face à la diversité de changement de contexte, nous en avons définit différents types.

- **Standard Switching** : Un changement complet de standard implique une refonte profonde de la chaîne de transmission. Les chaînes de transmissions entre les standards étant très différentes, c'est le type de changement de contexte qui demande les plus larges reconfigurations. Il est à noter qu'un standard switching concerne aussi les couches supérieures OSI.
- **Mode Switching** : Certains standards spécifient plusieurs modes de fonctionnement. Un mode switch est donc un changement de contexte intra-standard alors qu'un standard switch est un changement inter-standard. Dans la plupart des cas, le changement de mode ne concerne donc pas les couches hautes. Les paramètres du changement de mode pour la couche physique sont déterminés par la couche MAC. Par exemple le standard 802.11g possède plusieurs mode de fonctionnement DSSS, FHSS, OFDM.
- **Service Switching** : Un service switch sur la couche physique reste un changement de contexte intra-standard contrairement à un service switch au niveau applicatif pouvant entraîner un standard switch. Ce changement peut correspondre par exemple à une demande de changement de débit, la chaîne de traitement est généralement peu modifiée. Ce changement peut-être effectué par paramétrage ou par reconfiguration d'une partie des fonctions. Par exemple en mode OFDM du 802.11g, lors d'une demande d'augmentation de débit de 6Mbit/s à 54Mbit/s, la fonction de mapping change respectivement de BPSK à 64-QAM.
- **Performance enhancement, bug fixing** : La possibilité de changer un motif de traitement en vue de l'amélioration de ses performances ou dans le but de corriger une erreur peut correspondre à divers types de situations. La possibilité de réaliser ce type de petits changements nécessite la mise en place de règles de développement des algorithmes par composants de traitements rendant leurs reconfigurations possible de manière indépendantes. Ceci nécessite en outre la mise en place de mécanismes de contrôle afin de limiter la discontinuité de traitement occasionnée.

L'analyse des différents besoins de changement de contexte, effectuée ci-dessus, a mis en évidence qu'il existe différents niveaux de granularité des changements de contextes. Il est donc nécessaire de pouvoir gérer cette multi-granularité au niveau de la reconfiguration de l'architecture système afin d'en améliorer l'efficacité. Nous proposons dans la section 4, une architecture de gestion de configuration hiérarchique répondant efficacement aux besoins de gestion de la multi-granularité de configuration liées aux applications multi-standards.

3 Classification fonctionnelle

Cette seconde étape d'analyse porte sur les fonctions bande de base des 3 standards à l'émission. L'étude algorithmique nous a mené à constituer des classes suivant les caractéristiques et besoins matériels communs aux fonctions. Chacune des 3 classes fonctionnelles présentées par la suite, regroupe les fonctions sous forme générique. Par exemple la fonction générique codage convolutionnel dont les paramètres sont le taux de codage et les polynômes générateurs sont présentés pour les 3 standards étudiés dans le tableau 1.

TAB. 1 – Parameters for the 'Convolutional coding' Generic Function

Conv. Coding tri-standards	Coding Rate	Generator Polynomial
GSM	1/2	$G_0 = D^4 + D^3 + 1$
TCH/FH		$G_1 = D^4 + D^3 + D + 1$
TCH/HS	1/3	$G_4 = D^6 + D^5 + D^3 + D^2 + 1$
		$G_5 = D^6 + D^4 + D + 1$
		$G_6 = D^6 + D^4 + D^3 + D^2 + D + 1$
UMTS	1/2	$G_0 = D^8 + D^6 + D^5 + D^4 + 1$
BCH, PCH, RACH		$G_1 = D^8 + D^7 + D^6 + D^5 + D^3 + D + 1$
CPCH, DCH ...	1/3	$G_0 = D^8 + D^6 + D^5 + D^3 + D^2 + D + 1$
		$G_1 = D^8 + D^7 + D^5 + D^4 + D + 1$
		$G_2 = D^8 + D^7 + D^6 + D^3 + 1$
802.11g	1/2	$G_a = D^6 + D^4 + D^3 + D + 1$
(OFDM Mode)		$G_b = D^6 + D^5 + D^4 + D^3 + 1$

Une fonction est composée de blocs de traitements paramétrables. Ces blocs permettent de réaliser l'ensemble des fonctions paramétrées des différents standards. L'intérêt de la classification fonctionnelle est de grouper les fonctions nécessitant des ressources de traitement similaires. Les blocs de traitement paramétrables sont donc utilisables comme blocs ou motifs communs à plusieurs fonctions. Le groupement en classe fonctionnelle est proposé en

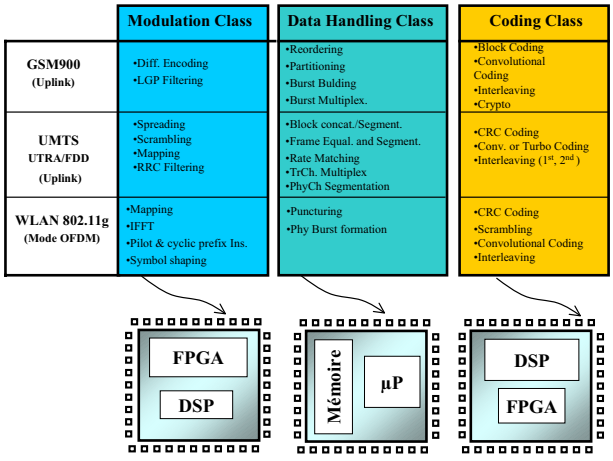


FIG. 2 – Functional Classification and HW/SW Needs

correspondance avec les ressources matérielles nécessaires, comme l'illustre schématiquement la figure 2. Ainsi, les implémentations de motifs d'exécution logiciels/matériels sont adaptées aux capacités de traitements requises par une classe de fonctions. Les 3 classes fonctionnelles sont présentées en détail ci-dessous :

- **Coding Class** : Cette classe regroupe les fonctions de codage canal, cyclic, convolutionnel turbo... . Elle est caractérisée par une grande variété de schémas de codage, ce qui nécessite une architecture d'exécution flexible. Une implémentation logicielle est possible car la capacité de traitement demandée n'est pas élevée sauf dans le cas du 802.11g à 54Mbps. En revanche, l'exécution des architectures de type processeurs 16/32 bits est sous-optimisée car les données à traiter sont de type bit. Une implémentation faible consommation sera basée sur l'utilisation de blocs de traitements matériel paramétrables.
- **Data Handling Class** : Cette classe regroupe les fonctions de manipulation de packets de données. Ces fonctions de transfert des données sont orientées contrôle. Les données sont de taille très différentes en entrée et sortie, dû à la nature des fonctions (concaténation, segmentation, multiplexage, etc.) . Très consommatrices de mémoire elles nécessitent une grande flexibilité de traitement, offerte par des architectures de type processeurs. Afin de réduire la consommation de puissance, il est nécessaire de faire attention à la gestion d'alimentation des blocs mémoires utilisés suivant les fonctions.
- **Modulation Class** : Cette classe rassemble les fonctions effectuées en bande de base juste avant la transposition en fréquence. Les fonctions sont souvent réalisées sur des données sur-échantillonnées demandant de grandes capacités de calcul notamment les fonctions de filtrage. Certaines de ces fonctions tirent avantages d'une implémentation sur des accélérateurs matériels.

4 Approche hiérarchique de gestion de configuration

Suite à l'analyse au niveau algorithmique, nous proposons ici de décrire un modèle fonctionnel de la gestion de configuration tenant compte des besoins de reconfiguration, déduit de cette étude. De plus, la gestion de configuration doit être en même temps adaptée aux capacités de reconfigurabilité des composants. Le but est de minimiser les ressources matérielles à reconfigurer à chaque nouvelle demande de changement de contexte de l'application. L'aspect hiérarchique de cette architecture de gestion de configuration permet de gérer efficacement la multi-granularité de configuration. Le gestionnaire de niveau 1 a un rôle de superviseur initiant une reconfiguration sur la chaîne de traitement. Cette reconfiguration est alors prise en charge par les gestionnaires de niveaux 2 à 3. Nous décrivons plus en détails ci-dessous le rôle de chaque niveau hiérarchique :

- **Niveau hiérarchique 1** : Le gestionnaire de configuration (*L1_CM*) fonctionne au niveau standard. Cette entité est en charge de sélectionner les fonctions gé-

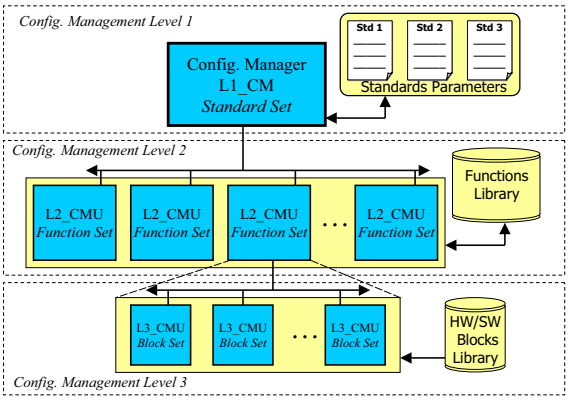


FIG. 3 – Hierarchical model of Configuration Management

nériques afin de constituer une chaîne dont l'ensemble des blocs fonctionnels correspond à un standard donné. A ce niveau, les fonctions sont vues comme des boîtes noires assemblées les unes par rapport aux autres et dotées d'un ensemble de paramères définit suivant les spécifications du standard choisi. Un changement de standard sera donc gérer par ce niveau hiérarchique.

– **Niveau hiérarchique 2** : Chaque fonction paramétrée par le *L1_CM* est prise en charge par une unité de gestion dédiée, appelée *Configuration Manager Unit (L2_CMU)*. Une fonction est constituée d'un ou plusieurs blocs de traitements. Le rôle d'un *L2_CMU* est de configurer cette fonction en sélectionnant les blocs de traitement en spécifiant aussi leurs interconnexions.
– **Niveau hiérarchique 3** : Chaque sous-bloc est paramétré, suivant les critères définies par le *L2_CMU*, puis instancié matériellement par un *L3_CMU*. Les interfaces de communications associées au bloc de traitement sont aussi configurées par le *L3_CMU*. La tâche principale des *L3_CMUs* est donc de configurer les ressources matérielles.

L'approche hiérarchique de gestion de configuration est appropriée pour gérer la multi-granularité de configuration et améliore par une gestion distribuée le contrôle de la reconfiguration partielle de la chaîne de traitement. Le partitionnement des fonctions en sous-bloc de traitement permet de garder une grande flexibilité de configuration. La mise en oeuvre de cette flexibilité est rendue efficace par la gestion hiérarchique. Cette découpe hiérarchique corresponds aussi aux différents besoins de changement de contexte identifiés dans le paragraphe 2. Suivant le type de changement de contexte demandé les différents niveaux de gestions de configurations sont utilisés. Le changement de standard utilisera tout les ressources de configuration du *L1_CM* aux *L3_CMUs*. Lors d'un changement de service où une seule fonction est changée, seul le *L2_CMU* associé est utilisé. Dans le cas d'une correction d'un programme (bug fixing) intervenant sur quelques motifs de traitements seuls les *L3_CMUs* concernés par le changement sont alors sollicités.

L'architecture d'exécution associée : L'architecture d'exécution est hétérogène pour répondre aux besoins des différentes classes de traitements présentées dans le paragraphe 3. Le partitionnement d'une application en composants puis en sous-blocs (PBU) simplifie la configuration. En effet, un PBU est soit matériel soit logiciel, le fait d'associer à chaque bloc de traitement un *L3_CMU* permet de spécialiser la gestion de configuration de ce dernier aux ressources qu'il contrôle.

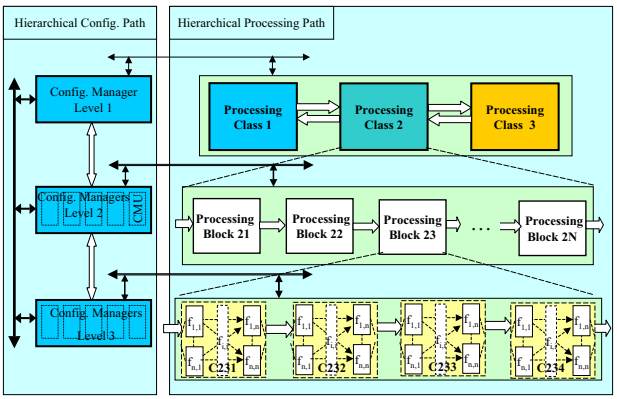


FIG. 4 – Hierarchical model for SDR System Design

5 Conclusions et perspectives

La conception système dans le contexte radio logicielle implique de définir une architecture ouverte et flexible prenant en compte la possibilité d'adapter le système aux futures évolutions des standards ou l'émergence de nouveaux. Nous pensons que notre approche est assez générique pour répondre à la diversité des standards. L'analyse fonctionnelle réalisée à l'émission et le modèle hiérarchique proposé peuvent être étendus côté réception. Le modèle de gestion hiérarchique sera validé sur une plate-forme de prototypage DSP/FPGA supportant les applications de reconfiguration d'un émetteur multi-standards.

Références

[1] M. Cummings and S. Haruyama., "Fpga in software radio," *IEEE Comms. Mag.*, pp. 108–112, 2 1999.

[2] S. Gultchev, K. Moessner, and R. Tafazoli, "Management and control of reconfiguration procedures in software radio terminals," in *Proc. 2^{nd} Workshop on Software Radios*, Karlsruhe, Germany, March 2002, pp. 125–129.

[3] R. Hartenstein, "A decade of reconfigurable computing : a visionary retrospective," in *Proc. Design, Automation, and Test in Europe*, Munich, Germany, 2001, pp. 642–649.

[4] R. Hoshyar, S. Gultchev, K. Seo, and R. Tafazolli, "Software reconfigurability - algorithm level approach," in *Proc. 4^{th} International Conference on 3G Mobile Communication Technologies*, London, UK, June 2003.

[5] A. Kountouris and C. Moy, "Reconfiguration in software radio systems," in *Proc. 2^{nd} Workshop on Software Radios*, Karlsruhe, Germany, March 2002, pp. 119–124.

[6] M. Mehta, N. Drew, and C. Niedermeier, "Reconfigurable terminals : an overview of architectural solutions," *IEEE Comms. Mag.*, pp. 82–88, 8 2001.

[7] J. Mitola, "The software radio architecture," *IEEE Comms. Mag.*, vol. 33, pp. 26–38, 5 1995.

[8] J. Palicot and C. Roland, "Fft a basic function for a reconfigurable receiver," in *Proc. 10th International Conference on Telecommunications*, Papeete, Tahiti, Febuary 2003, pp. 898–902.

[9] S. Srikanteswara, J. Reed, P. Athanas, and R. Boyle, "A soft radio architecture for reconfigurable platforms," *IEEE Comms. Mag.*, vol. 38, pp. 140–147, 2 2000.

Implantation du décodeur entropique de JPEG 2000 basée sur la reconfiguration dynamique des FPGAs

Sophie BOUCHOUX, El-Bay BOURENNANE

LE2I - Université de Bourgogne
Aile H - Sciences Mirande, B.P. 47870
21078 Dijon cedex, France
sbouchou@u-bourgogne.fr, ebourenn@u-bourgogne.fr

Résumé – Cet article a pour objet la mise en œuvre de la reconfiguration dynamique des FPGAs. L'application envisagée ici, est le décodeur entropique de JPEG 2000 et la procédure MQ-Décode. Des critères d'analyse pour les différentes implantations seront tout d'abord définis, puis nous présenterons la norme JPEG 2000 et l'architecture employée.

Abstract – This article is about dynamic reconfiguration of FPGA. The application implemented to illustrate this case is entropic decoder and arithmetic decoder of JPEG 2000. Some analysis criteria will be defined, JPEG 2000 norm described and the architecture used presented.

1 Introduction

Le standard de compression JPEG 2000 mis au point par le groupe JPEG a pour but d'offrir une alternative à l'actuel JPEG pour la compression d'images fixes. En effet, le standard actuel JPEG ne permet pas de palier à certains défauts ou artéfacts et n'est pas adapté au codage de certains types d'images. JPEG 2000 permet généralement d'obtenir une meilleure qualité d'image à taux de compression équivalent que JPEG. Nous avons choisi d'implanter une partie de cet algorithme sur une architecture à base de FPGA en utilisant la propriété de reconfiguration dynamique de ces derniers. La première partie de cet article est consacrée à la définition de critères d'analyse qui permettront d'évaluer les apports de la reconfiguration dynamique par rapport à la configuration statique. La seconde partie est consacrée à la description de l'algorithme JPEG 2000 et plus particulièrement à la partie que nous avons choisi d'implanter : le décodeur entropique. La troisième partie concerne la description de l'architecture utilisée, nommée ARDOISE. Enfin, la dernière partie présente les résultats obtenus par implantation du décodeur entropique en reconfiguration dynamique partielle sur l'architecture ARDOISE.

2 Critères d'analyse

2.1 Critères d'analyse des implantations statique et dynamique

L'analyse des implantations statique et dynamique va porter sur l'évaluation de deux critères [1], [2], l'un représente le coût de l'implantation et l'autre ses performances. Le coût sera désigné ici par le nombre de CLBs (Blocs Logiques Configurables) occupés par l'algorithme N_{CLBs_stat}. Le temps d'exécution T_{Exec_stat} nécessaire pour réaliser l'application permettra de définir le critère de performance (Perf.). De la même façon, nous définissons le coût dynamique et le critère de performance dynamique. A performances égales, l'apport de la reconfiguration dynamique sera donné par le rapport :

$$\frac{N_{CLB_Stat} - N_{CLB_Dyn}}{N_{CLB_Stat}} \times 100\% \quad (1)$$

Afin d'évaluer le degré d'optimisation de l'architecture mise en oeuvre, nous utiliserons aussi le rendement spatio-temporel pour chacun des types d'implantation (statique et dynamique). Le rendement spatio-temporel est donné par la relation 2.

$$\rho_{Spat_temp} = \sum_{i=1}^{M} (\rho_{Spat_tache_i} . \rho_{Temp_tache_i}) < 1 \quad (2)$$

avec

$$\rho_{Spat_stat} = \frac{N_{CLBs_actives}}{N_{CLBs_dispo}} \quad (3)$$

et

$$\rho_{Temp_stat} = \frac{T_{Exec_stat}}{T_{Dispo}} \quad (4)$$

N_{CLBs_dispo} est le nombre de CLBs disponibles dans le FPGA et T_{Dispo}, le temps maximal dont on dispose pour effectuer le traitement (en imagerie standard, T_{Dispo} = T = 40 ms).

2.2 La reconfiguration dynamique dans le cas général

Dans certains cas, les traitements d'une application utilisant la configuration statique des FPGAs peuvent être exécutés en un temps très court et par conséquent, les FPGAs se mettent en attente de l'image suivante et restent inactifs. La figure 1 montre ainsi une application pouvant présenter ou non un pipeline et la durée de son exécution sur un FPGA classique en mode statique peut être inférieure à la durée d'acquisition de l'image : $M.N.T_{Iteration_stat} < T$ (M étant le nombre de tâches à exécuter et N le nombre d'échantillons à traiter).

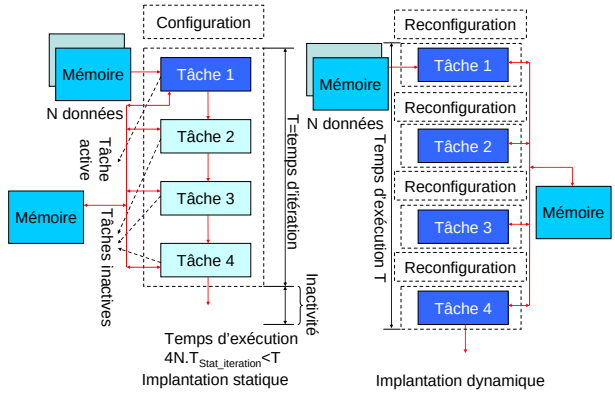


FIG. 1 – Implantation statique à rendement spatio-temporel inférieur à 1 et l'implantation dynamique correspondante

2.2.1 Discussion de l'implantation statique

Le coût et la performance sont respectivement donnés par les équations 5 et 6.

$$\begin{aligned} Cout_{Stat} &= N_{CLB_stat} \\ &= \sum_{i=1}^{M} N_{CLB_tache_i} + N_{CLB_controle} \end{aligned} \quad (5)$$

$$\begin{aligned} Perf_{Stat} &= T_{Exec_stat} \\ &= M.N.T_{Iteration_stat} + T_{Inactivite} = T \end{aligned} \quad (6)$$

2.2.2 Discussion de l'implantation dynamique

$$\begin{aligned} Cout_{Dyn} &= N_{CLB_dyn} \\ &= \max_{i=1}^{M}(N_{Tache_i_CLB} + N_{CLB_controle}) \end{aligned} \quad (7)$$

$$\begin{aligned} Perf_{Dyn} &= T_{Exec_dynam} = T_{Exec_stat} \\ &= M.N.T_{Iteration_stat} + T_{Inactivite} \\ &= M.N.T_{Iteration_dyn} + M.T_{Config} = T \end{aligned} \quad (8)$$

$$G_{CLB} = \frac{N_{CLB_stat} - \max_{i=1}^{M}(N_{CLB_tache_i})}{N_{CLB_stat}} \times 100\% \quad (9)$$

La période d'itération dynamique nécessaire est telle que :

$$T_{Iteration_dyn} = T_{Iteration_stat} + \frac{T_{Inactivite}}{M.N} - \frac{T_{Config}}{N} \quad (10)$$

Le temps d'inactivité de l'implantation statique est exploité pour réduire la fréquence d'itération de l'implantation dynamique. Si la relation 11 est vérifiée, alors la fréquence d'itération dynamique peut même être plus faible que la fréquence d'itération statique.

$$\begin{aligned} &(T_{Inactivite})_{Stat} \geq M.T_{Config} \\ \Rightarrow\ &T_{Iteration_dyn} \geq T_{Iteration_stat} \end{aligned} \quad (11)$$

On parvient ainsi à obtenir une réduction de la surface et une diminution de la fréquence des traitements grâce à la reconfiguration dynamique.

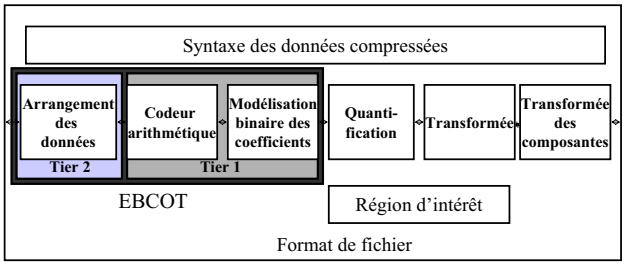


FIG. 2 – Les différents blocs de l'algorithme JPEG 2000

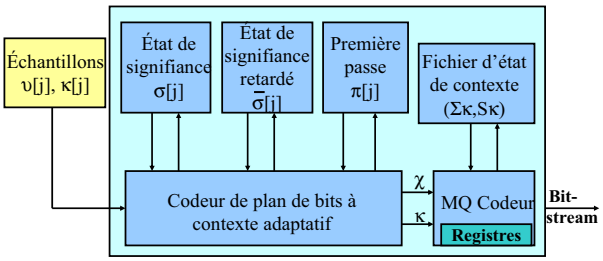


FIG. 3 – Schéma bloc de la partie Tier 1 de l'algorithme EBCOT

3 La norme JPEG 2000

La norme JPEG 2000 [3], [4] peut-être décomposée en plusieurs blocs successifs comme le montre la figure 2. Sur cette figure, les blocs sont disposés dans l'ordre « décodeur ». L'image d'origine est découpée en tuiles après la transformation des composantes. Toutes les tuiles subissent ensuite une transformation en ondelettes (transformation avec ou sans pertes), indépendamment les unes des autres. Tous les niveaux de résolution sont alors découpés en *code-block* de même taille (64×64, 32×32, ...). Les coefficients des *code-block* subissent une quantification et les coefficients quantifiés sont décomposés en plans de bits. En commençant par le plan de bits de poids fort, tous ces plans sont codés en fonction de leur « signifiance » et de leur contexte par trois passes successives. Les bits issus de ces différentes passes, ainsi que le contexte associé, sont ensuite envoyés à un codeur arithmétique (MQ-codeur identique à celui de JBIG2 [3]). Ces données codées sont ensuite mises en forme, en respectant la syntaxe définie dans la norme (*code-stream syntax*) pour former le *bit-stream* final. Ces trois derniers modules forment ce que l'on appelle EBCOT (*Embedded Block Coding with Optimisation Truncation*). L'EBCOT est lui-même composé de 2 parties distinctes : Tier 1 et Tier 2. Tier 1 regroupe la modélisation binaire des coefficients et le codeur arithmétique, et Tier 2 correspond à l'arrangement des données.

Nous allons nous intéresser plus particulièrement à la partie Tier 1 de l'EBCOT (Figure 3) et à la partie « décodeur ».

On trouve dans [3] les procédures correspondant à cette partie de l'algorithme (*Embedded Block Decoder*). Après initialisation, tous les éléments du *bit-stream* subissent les trois passes de décodage successivement : d'abord la passe de signifiance, puis la passe d'affinage et enfin la passe de nettoyage. Tous les plans de bits sont parcourus de la même façon. Les contextes, c'est-à-dire l'état de signifiance des 8 voisins du bit traité, sont initialisés puis modifiés au fil du décodage par la partie « décodeur de plan de bits à contexte adaptatif ». Il y a en tout 19 va-

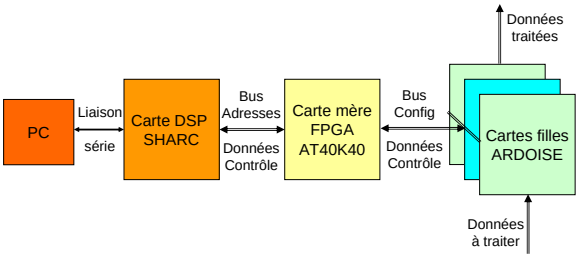


FIG. 4 – Les différents modules de l'architecture ARDOISE

leurs de contexte possibles : 9 pour la passe de signifiance (de 0 à 8), 5 pour le codage du signe (de 9 à 13), 3 pour la passe d'affinage (de 14 à 16) et 2 pour la passe de nettoyage (17 et 18). Dans toutes ces passes de décodage, une procédure est appelée régulièrement : la procédure MQ-Décode. Le codage arithmétique est basé sur la subdivision récursive de l'intervalle [0, 1) en fonction des symboles de la séquence d'entrée. La probabilité récursive de subdivision d'intervalle du codage Elias est la base de la procédure de codage arithmétique binaire. Quand le premier symbole de la séquence d'entrée est traité, l'intervalle est subdivisé en sous-intervalles associés aux symboles de l'alphabet. Chaque fois qu'un nouveau symbole est traité, le sous-intervalle associé est subdivisé de la même manière que l'intervalle initial. Cette procédure récursive cesse lorsque le dernier symbole de la séquence est traité.

4 L'architecture utilisée pour l'implantation : ARDOISE

Le projet ARDOISE a vu le jour au sein du GDR ISIS dans le but d'évaluer les performances d'une architecture dédiée au traitement des images en temps réel qui utiliserait la propriété de reconfiguration dynamique (totale ou partielle) des FPGAs [5]. Le FPGA choisit pour cette architecture est un FPGA ATMEL de la famille AT40K40. Cette famille de FPGAs a été choisie car elle permet des reconfigurations très précises (chaque CLB du FPGA est configurable indépendamment des autres contrairement à d'autres familles de FPGAs (par exemple XILINX) dont l'élément de configuration de base est la colonne de CLBs). Deux versions de l'architecture ARDOISE ont été réalisées. Actuellement, nous travaillons sur la seconde version. Nous disposons sur celle-ci d'une carte DSP SHARC Analog Devices, d'une carte mère à base d'un FPGA AT40K40 et de 3 modules carte fille (AT40K40 et 2 mémoires SRAM) (voir Figure 4).

La carte DSP sert d'interface entre la carte mère FPGA et le PC (via une liaison série). Elle est utilisée pour transmettre les fichiers de configuration des cartes filles et générer les signaux nécessaires. Le FPGA de la carte mère est chargé de gérer la configuration des différentes cartes filles. Sur la carte mère, on dispose d'une mémoire SRAM qui sert à stocker les fichiers de configuration reçus par le port série, ainsi qu'une mémoire Flash qui contiendra à terme le fichier de configuration du FPGA mère et les différentes configurations des cartes filles.

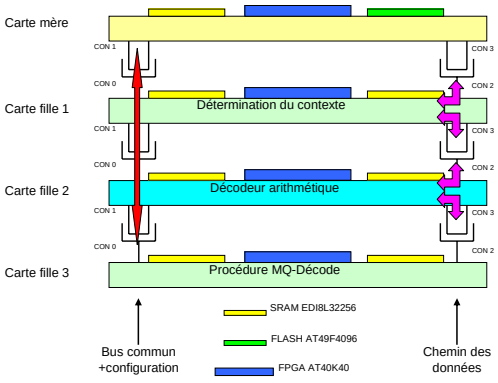


FIG. 5 – L'implantation des différentes parties du décodeur entropique sur l'architecture ARDOISE

5 Application de la reconfiguration dynamique

L'ensemble de l'algorithme JPEG 2000 nécessite beaucoup de ressources et de calculs. La proportion des besoins pour chacune des parties est de 70 % pour l'EBCOT, 20 % pour la transformation en ondelettes et les 10 % restants pour tout le reste des calculs [6]. La première partie de l'implantation concerne la procédure MQ-Décode [7]. Cette procédure est l'élément de base du décodeur entropique. Elle est utilisée une à plusieurs fois dans chaque passe de décodage. Cette partie de l'algorithme JPEG 2000 a été implantée sur une des 3 cartes filles de l'architecture ARDOISE.

La seconde partie de l'implantation concerne la détermination du contexte du symbole courant. En effet, la procédure de décodage utilise le voisinage du symbole à décoder. Cette partie de l'algorithme est elle aussi implantée sur une carte fille de l'architecture ARDOISE.

La dernière partie de l'algorithme du décodeur entropique de JPEG 2000 à implanter est la gestion des différentes passes de décodage qui interviennent dans l'algorithme EBCOT. Ces différentes passes de décodage sont également implantées sur une carte fille ARDOISE. L'implantation complète de ces 3 passes simultanément étant impossible sur un seul FPGA AT40K40, la reconfiguration dynamique de ce FPGA sera utilisée. Les différentes passes de décodage interviennent successivement pour chaque plan de bits : tout d'abord la passe de signifiance, puis la passe d'affinage et enfin, la passe de nettoyage, sauf pour le premier plan de bits signifiant qui est décodé uniquement par la passe de nettoyage. Ces différentes passes sont implantées les unes après les autres sur le même FPGA. Une autre configuration de gestion est alors nécessaire pour définir la configuration suivante à implanter. On utilise la mémoire interne du FPGA pour stocker les différentes informations nécessaires pour connaître le déroulement du décodage. Ces informations sont utilisées dans chacune des configurations, il faut donc que la reconfiguration du FPGA ne soit que partielle pour sauvegarder ces valeurs.

Au final, on utilise pour implanter le décodeur entropique de JPEG 2000 les 3 cartes filles de l'architecture ARDOISE. On peut voir sur la figure 5 la disposition des différentes parties implantées. Les FPGAs contenant les configurations du « MQ-Décodeur » et de la « détermination du contexte » sont confi-

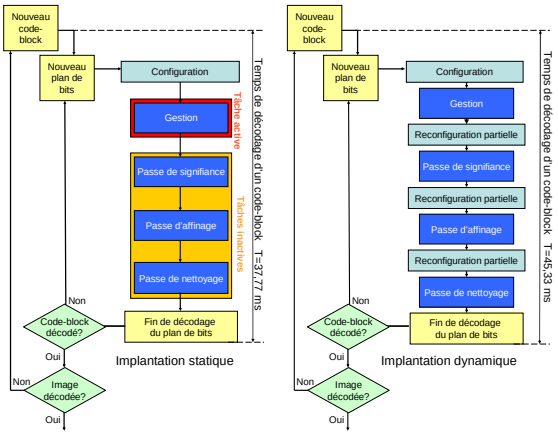


FIG. 6 – Représentation temporelle des fonctionnement en mode statique et en mode dynamique de l'implantation du décodeur entropique

gurées au début du décodage et le reste en permanence jusqu'à la fin. Le FPGA contenant les différentes passes de décodage est reconfiguré partiellement 4 fois pour chaque plan de bits à décoder.

La figure 6 représente la comparaison du fonctionnement des différentes configurations dans les cas statique et dynamique, au cours du temps.

La reconfiguration totale du FPGA ATMEL AT40K40 à la fréquence maximale de 33 MHz est effectuée en 337 μs. De plus, la fréquence de fonctionnement de ce FPGA est de 100 MHz [8]. La taille des *code-blocks* à décoder ici sera de 32×32 éléments. Pour chacune des trois passes de décodage, nous estimons le nombre maximal de cycles nécessaires pour traiter les 1 024 éléments. Si le traitement est effectué à la fréquence de 100 MHz, ce temps d'exécution est de 44,33 ms. Nous pouvons considérer que le temps total de reconfiguration du FPGA en reconfiguration dynamique pour les quatre reconfigurations partielles nécessaires est d'environ 1 ms. Nous obtenons donc pour la reconfiguration dynamique un temps total d'exécution de $Perf_{Dynam}$ = 45,33 ms pour chaque *code-block* 32×32. Dans le cas de la configuration statique, le temps d'exécution sera à peu de chose près identique à celui de la reconfiguration dynamique. En effet, nous nous trouvons ici dans le cas où il n'y a aucun parallélisme par les fonctions. Tous les éléments doivent être traités en respectant l'ordre de balayage défini dans la norme et le décodage de chaque élément est dépendant des éléments décodés précédemment. Le seul parallélisme que l'on peut avoir dans ce cas vient du fait que chaque *code-block* est indépendant des autres. Il est donc possible de traiter plusieurs *code-blocks* en même temps, à condition de disposer de la surface nécessaire pour implanter plusieurs fois le MQ-Décodeur dans le FPGA. Dans notre cas, la reconfiguration dynamique n'augmente pas le temps de traitement de manière significative par rapport au cas statique.

Nous pouvons calculer les coûts, performances et rendements obtenus dans chaque cas. Nous prenons comme temps de référence pour nos calculs, le temps T = 40 ms. Le tableau 1 récapitule ces résultats.

TAB. 1 – Comparaison des résultats obtenus pour les cas de configurations statique et dynamique sur ARDOISE

	Configuration statique	Configuration dynamique	Gain dynamique/statique
Coût (CLBs)	4 733	3 626	23,39 %
Performance	44,66 ms	45,33 ms	-1,5 %
ρ_{Spat}	0,5136	0,5246	2,14 %
ρ_{Temp}	1,12	1,13	0,9 %
ρ_{Spat_temp}	0,5752	0,5928	3,06 %

6 Conclusion

Nous avons montré ici l'implantation d'une partie de l'algorithme de décodage JPEG 2000 en utilisant la reconfiguration dynamique des FPGAs et comparé les coûts et les performances de cette méthode par rapport à l'implantation statique. Dans ce cas précis, la reconfiguration dynamique nous permet uniquement d'obtenir un gain en surface. Nous avons pour le moment étudié uniquement la reconfiguration dynamique des FPGAs de la famille ATMEL. Il serait également intéressant d'étudier la reconfiguration dynamique des FPGAs de la famille XILINX.

Références

[1] E. Bourennane, C. Milan, M. Paindavoine, S. Bouchoux, *Real Time Image Rotation using Dynamic Reconfiguration*, Real Time Imaging Journal 8, pp. 277 - 289, 2002.

[2] M. Wirthlin, B. Hutchings, *Improving Functional Density using Run-Time Circuit Reconfiguration*, IEEE Transactions on Very Large Scale Circuit Integration (VLSI) Systems, Vol. 6, No. 2, pp. 247 - 256, juin 1998.

[3] JPEG Group, *JPEG 2000 Image Coding System*, JPEG 2000 final commitee draft version 1.0, mars 2000.

[4] D. Taubman, M. Marcellin, *JPEG 2000 Image Compression Fundamentals and Practice*, Kluwer Academic Publishers, ISBN : 079237519X, novembre 2001.

[5] D. Demigny, M. Paindavoine, S. Weber, *Architecture Reconfigurable Dynamiquement pour le Traitement Temps Réel des Images*, Revue Techniques et Sciences de l'Information, Vol. 18, No. 10, pp. 1087 - 1112, 1999.

[6] D. Nicholson, P. Martinez, M. Iregi, J. Corral *Evaluation de la COmplexité Algorithmique de JPEG 2000*, Proceedings CORESA, 2000.

[7] J. Fagès, *JPEG 2000 - Principes, Implémentation et Evaluation*, Mémoire d'ingénieur CNAM, septembre 2000.

[8] ATMEL Corporation, *Data Sheet : 5K - 50K Gates Coprocessor FPGA with FreeRAM*.

Optimisation sur calculateur d'un modèle d'architecture de couche de cellules rétiniennes

François DEVILLARD[1],Frédéric GRANDIDIER[2],Bernard HEIT[2]

[1]Centre de Recherche en Automatique de Nancy UMR CNRS 7039 (CRAN - CNRS)
11 rue de l'Université, F-88100 Saint-Dié-des-Vosges

[2]Centre de Recherche en Automatique de Nancy UMR CNRS 7039 (CRAN - CNRS)
2 rue Jean Lamour, F-54500 Vandoeuvre-lès-Nancy Cedex
francois.devillard@iutsd.uhp-nancy.fr, bernard.heit@esstin.uhp-nancy.fr

Résumé – Cet article traite de l'implémentation sur calculateur d'une modélisation des couches cellulaires de la rétine biologique envue de son application en vision artificielle. En vision artificielle ces traitements permettent d'extraire dans un flot continu les indices visuels avec une bonne robustesse. Le coût de calcul résultant nécessite par contre des algorithmes adaptés et des machines spécialisées pour une mise en oeuvre performante. La solution proposée utilise une adaptation de l'algorithme pour accélérer le calcul des couches rétiniennes tout en conservant leurs propriétés locales adaptatives.

Abstract – This article presents a work about computer implementation of a retinal layer modelisation that is dedicated to machine vision. In computer vision this kind of processing allows to extract visual indices in a continuous flow with a good robustness. To minimize its computing cost, the implementation require well suited algorithms and hardware systems to implement efficient applications. The proposed method uses an adaptated algorithm to accelerate the perform of the retinal layer modelisation while preserving its adaptive local properties.

1 Introduction

La rétine fovéale est constituée d'un *empilement* de couches de cellules traitant le signal lumineux pour en extraire des indices visuels (contrastes de chrominance statique et dynamique, luminance...). Dans l'objectif de modéliser son fonctionnement, une couche élémentaire de cellules nerveuses rétiniennes a été définie et sert de base à la conception d'une architecture de rétine utilisable en vision artificielle. Cette couche élémentaire est modélisée par des outils de traitement du signal et permet de reproduire une approche comportementale de la couche rétinienne biologique. Elle est composée d'un réseau de cellules de traitement paramétrables localement pour reproduire les mécanismes d'adaptation biologiques observables. L'assemblage des différentes couches de la rétine doit satisfaire à des contraintes d'implémentation temps réel dans le cadre d'applications de vision artificielle ; l'algorithme correspondant décrit la structure des couches dans la section 2. La section 3 présente une approche plus efficace sur calculateurs de la programmation des couches rétiniennes. La démarche des modifications apportées sur l'algorithme est présentée. Le gain observé sur le coût de calcul est évalué à chaque étape. La section 4 donne quelques résultats qualitatifs obtenus avec l'algorithme appliqué sur une séquence d'images d'entrée statique. Enfin la dernière section conclut et donne les perpectives envisageables à court terme pour nos travaux.

2 Architecture de la rétine

2.1 Modélisation de la PLE

Les mécanismes de la perception visuelle [4] réduits à la couche plexiforme externe (*PLE*) du système parvocellulaire (zone fovéale de la rétine) sont schématisés dans la figure 1. L'acquisition de la scène observée est réalisée par une matrice de photorécepteurs (cellules de type cône) qui transforment l'énergie lumineuse I en énergie électrochimique C. Deux autres couches de cellules effectuent à la suite une première analyse du message par le réhaussement des contrastes présents dans l'image. La couche des cellules horizontales est une couche de régularisation estimant localement un référentiel de luminance H de la scène. La couche des cellules bipolaires différencie la réponse des cellules cône et horizontale pour fournir une réponse *Bip* mettant en évidence les contrastes caractérisant les objets observés. Chaque couche est constituée d'un maillage plus ou moins diffus de cellules interagissant latéralement. Chaque cellule fournit une réponse, fonction d'une somme pondérée et non linéaire de ses entrées synaptiques. De plus, la réponse est une combinaison de filtrages passe-bas spatial et temporel. Par conséquent, chaque cellule peut être modélisée par un processeur ou filtre élémentaire à entrées multiples (les synapses cellulaires) et dont la sortie (l'activité cellulaire) est donnée par un filtre passe-bas spatio-temporel [1]. L'association de filtres élémentaires travaillant en maillage permet de construire une couche de traitement cellulaire aux comportements comparables à celui d'une couche de la rétine.

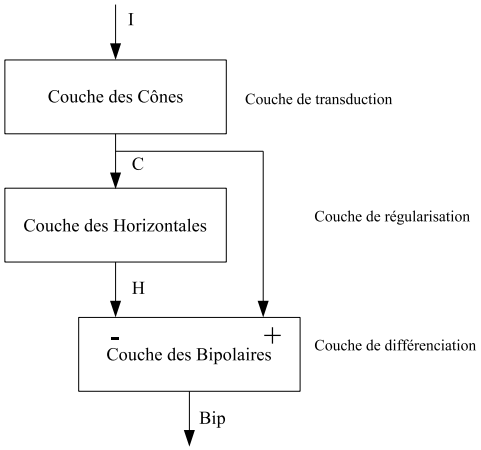


FIG. 1 – Schéma fonctionnel de l'architecture de la couche *PLE* de la rétine biologique photopique (vision diurne).

2.2 Modélisation des couches rétiniennes

La mise en oeuvre de notre solution originale repose sur un filtrage spatio-temporel numérique distribué et adaptatif. La fonction de sortie de la cellule $s_{x,k}$ est donnée dans l'équation 1 (pour une seule dimension spatiale x). Elle traduit l'activité cellulaire pour une entrée $e_{x,k}$ à la position x et à l'instant k. Les différentes couches de cellules présentées figure 1 utilisent la même équation avec des paramètrages différents. Les périodes d'échantillonnage spatiale et temporelle sont unitaires pour simplifier la formulation.

$$s_{x,k} = (1 - r_{x,k}) \cdot \left[b_{x,k} \sum_{i=-\lambda}^{\lambda} \frac{C_{i+\lambda}^{2\lambda}}{4^{\lambda}} \cdot e_{x+i,k} \right] + r_{x,k} \cdot s_{x,k-1} \quad (1)$$

Avec λ représentant $n_{x,k}$ et $0 \leq r_{x,k} < 1$.

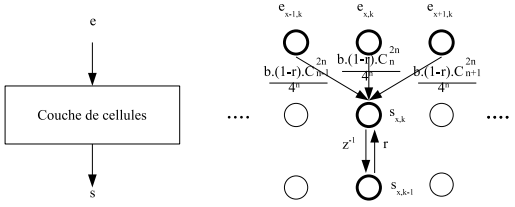


FIG. 2 – Graphe de fluence du traitement du processeur élémentaire équivalent à une cellule à synapses positionnée en x à l'instant k. Seules sont représentées les connexions avec le proche voisinage cellulaire. L'opérateur z^{-1} représente un retard pur et C la fonction statistique de combinaison.

Ce filtre linéaire et séparable est une association de structures de type *RIF* (Réponse Impulsionnelle Finie) et *RII* (Réponse Impulsionnelle Infinie) appliquées respectivement dans les domaines spatial et temporel. Les paramètres influençant la réponse spatiale de la cellule sont n et b (respectivement le couplage latéral entre cellules et le facteur de gain de cellule). Le coefficient b agit sur l'excitation de la cellule x comme un gain et le coefficient n traduit l'influence de la cellule x sur ses voisines. Le paramètre n d'étalement fixe le diamètre de l'arborescence dendritique à $2n + 1$.

Ces coefficients agissent différemment sur la réponse spatiale et temporelle de la couche. L'intensité de l'influence de n, en fonction de la distance, est obtenue à l'aide d'une fonction binomiale de manière à fournir une réponse impulsionnelle passe-bas, centrée et pseudo gaussienne [5]. Le paramètre influençant la réponse temporelle de la cellule est r. Il permet de reproduire l'effet intégrateur d'une constante de temps sur la sortie de la cellule (due à la résistance et à la capacité membranaire de la cellule) [6]. La figure 2 représente l'équation 1 graphiquement sous la forme d'un maillage. La couche de cellules cône prend en entrée la succession d'images formant la séquence vidéo (1 pixel = 1 cône). Les cônes dans notre version d'algorithme sont considérés achromatiques.

3 Mise en œuvre algorithmique

La mise en oeuvre de l'équation 1 nécessite une adaptation algorithmique et/ou architecturale pour obtenir une rétine minimisant les temps de calcul. En effet, la composante spatiale du filtrage est la principale consommatrice de ressources de calcul. Le deux paragraphes qui suivent décrivent la décomposition du filtrage *RIF* spatial en une cascade de filtres élémentaires. Cette décomposition est réalisée d'une part pour accélérer le calcul et d'autre part respecter les consignes locales $n_{x,k}$ des cellules de filtrage.

3.1 Décomposition en filtres élémentaires

L'équation 1 (pour $r_{x,k} = 0$) étendue aux deux dimensions spatiales conduit à une convolution binomiale locale équivalente à un masque de taille $(2n_{x,k} + 1) \times (2n_{x,k} + 1)$.

D'un point de vue algorithmique cette opération peut être facilement optimisée par une décomposition en cascade de convolutions élémentaires bidimensionnelles appliquées dans l'image (masque $b_{in}[3 \times 3]$ de l'équation 2). Par cette méthode, nous passons d'un coût en opérations *MAC* (Multiplication ACcumulation)de $(2n + 1)^2$ à $9n$.

D'un point de vue architectural, cette solution programmée sous forme d'itérations simples permet de mieux optimiser l'utilisation du code dans l'architecture des processeurs par un adressage rapide et simple des opérandes et d'exploiter pleinement le cache d'instructions, même limité, de certains de calculateur ([3] *DSP*, *CPU*...). Pour ce qui est de la composante temporelle de filtrage décrite par l'équation 1, elle est cascadée avec la composante spatiale.

3.2 Filtrage spatial

Si on applique les optimisations précédentes le temps de calcul, fonction de $n_{x,k}$, reste toujours important. Nous avons donc mis en œuvre une architecture de filtrage multi-échelle pour la couche illustrée figure 3.

3.2.1 Architecture du filtrage

Le filtrage multi-échelle effectue des traitements dans les bandes de fréquence adaptées de façon à réduire les

temps de calcul. Il est réalisé à 2 échelles de fréquence ; l'échelle 0 qui est le niveau de représentation de l'image originale et l'échelle i, celui d'une représentation de plus basse résolution de l'image dans un spectre de fréquences divisé par 2^i. Dans les conditions fixées par la section 3.2.3, le filtrage est effectué en 3 étapes :

1. Cette étape prépare les images au passage de l'échelle 0 à i. Pour cela, elle filtre les hautes fréquences pour éviter un recouvrement de spectre puis elle sous-échantillonne les images opérandes de la couche (images e, n et b) ;
2. Cette étape réalise le filtrage passe-bas dans l'échelle i ;
3. Pour finir cette dernière étape à l'inverse de l'étape 1 prépare l'image résultat de la couche au passage de l'échelle i à 0 ;

Le filtrage dans l'échelle de fréquence i est effectué par n_i itérations successives. Pour une échelle i, un pixel $e_{x,k}$ de l'image d'entrée est convolué avec le noyau b_{in} à la condition où le pixel $n_{x,k}$ est supérieur strictement à 0. Dans la condition inverse la convolution n'est pas faite.

$$\texttt{Avec } b_{in} = \frac{1}{16} \cdot \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \tag{2}$$

b_{in} est l'opérateur binomial élémentaire de convolution.

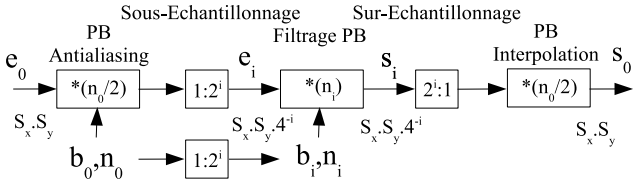


FIG. 3 – Architecture du filtre spatial multi-échelle. La notation $*$ est l'opérateur de convolution. e_i et s_i sont respectivement l'image d'entrée et l'image résultat provisoire de la couche dans l'échelle i.

3.2.2 Fonction *coût* calcul

L'équation 3 donne le calcul du nombre d'itérations n_{eq} du filtrage équivalent fait à pleine résolution avec les conditions sur n_0 pour l'antialiasing. L'ordre de grandeur de la fonction *coût* s'exprime à partir de deux paramètres de filtrage n_0 et n_i pour une image de taille $S_x \times S_y$. Les temps de filtrage d'antialiasing et d'interpolation sont pris en compte. Le coefficient α entier et supérieur à 2^i [2] et représente le nombre de filtrages dans les étapes de sous ou bien sur-échantillonnage.

$$\begin{aligned} n_{eq} &= n_0 + n_i \cdot 2^i \quad \texttt{avec } n_0 = 2 \cdot \alpha \\ \textit{Coût} &= S_x \cdot S_y \cdot ((n_0 + 3) + (n_i + 1) \cdot 4^{-i}) \end{aligned} \tag{3}$$

On constate dans la figure 4 qu'avec une telle architecture de filtre qu'il existe un temps minimal de traitement correspondant à une échelle i_{opt}. Tout filtrage de couche en fonction du paramétrage sera effectué à une échelle adaptée i_{opt} afin de minimiser son temps global de calcul.

3.2.3 Choix de l'échelle de fréquence

Les paramètres n_0 et n_i doivent être choisis pour respecter à la fois les conditions de Shannon et obtenir un temps de calcul minimisé. L'algorithme 4 donne la méthode utilisée. La valeur minimale de n restreint éventuellement le choix de l'échelle i_{opt} détecté à l'aide de la fonction *Coût*. La valeur β est le minimum souhaité d'itérations de filtrage dans l'échelle i.

if $(Min(n) - 2 \cdot \alpha) > (\beta \cdot 2^i)$ **then**
  $n_i = Partie\ Entière\ (2^{-i} \cdot (n - n_0))$;
else
  $n_i = 0$; $n_0 = n$;
end if

(4)

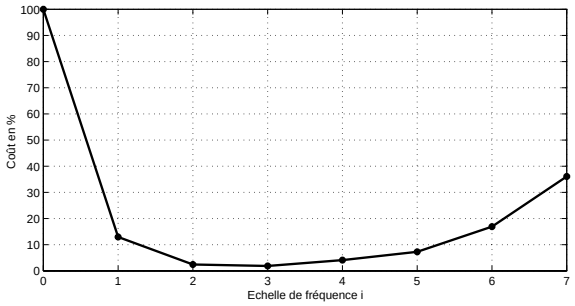


FIG. 4 – La fonction *Coût* en % par rapport au calcul en résolution 0 est représentée pour une image 1024×1024 filtrée spatialement de façon homogène avec $n = 1000$. Elle passe par une valeur minimale $i_{opt} = 3$ détectable et utilisée pour miniminiser le temps de calcul en fonction du contexte de la couche.

3.3 Filtrage temporel

Le filtrage est réalisé à la suite du spatial à pleine résolution. Le coût de calcul de ce filtre, bien moindre que le spatial, est de l'ordre au mieux de $2 \cdot S_x \cdot S_y$ traité dans un calculateur à jeu d'instructions arithmétiques de type MAC.

4 Résultats

Nous présentons ci-dessous un exemple de traitement obtenu sur une couche simulant la perception des cônes de la rétine. Les cônes sont disposés dans la rétine avec une densité variable en fonction de leur position [7]. Le centre de la rétine appelée fovéa est une zone très dense en cellules cônes pour la vision de précision ($150000/mm^2$). La périphérie dediée plutôt à la vision dynamiques est moins précise car elle possède une densité de cellules cônes vingt fois moindre. Nous simulons le résultat de perception des cônes figure 6 en réponse d'une image statique de luminance représentée figure 5. Notre maillage de cellule étant uniforme, nous agissons sur le couplage latéral entre les cellules pour créer cet effet. Ainsi on observe figure 6 une

zone centrale de nettetée (région fovéale) et une périphérie floue (macula).

La figure 7 présente un comparatif de réponses. La référence est un filtre passe-bas gaussien de variance fréquencielle ${\sigma_f}^2 = 2 \cdot 10^{-2} pixel^{-2}$ équivalent à un filtrage de $n = 100$ itérations du noyau b_{in}. Ce filtre gaussien est comparé figure 7.b déja avec ce dernier (résultats similaires). La figure 7.a présente l'écart de réponse créé par l'utilisation du multi-échelle pour $i = 2$ avec $n_0 = 16$ et $n_2 = 21$.

FIG. 5 – Image *château* de luminance originale.

FIG. 6 – Réponse de la couche cône.

5 Conclusions et perspectives

Nous avons présenté une solution à l'implémentation de couches rétiniennes constituées de cellules de filtrage. Une méthode de filtrage est proposée pour calculer l'algorithme caractérisant la fonction d'une couche rétinienne. Elle propose une organisation de la forme du filtre (décomposition en filtres élémentaires et le multi-échelle) pour accélérer les temps de calcul. L'utilisation de filtres élémentaires nous fait gagner des facteurs importants de temps (section 3.1). Le multi-échelle accentue ce gain illustré figure 4 sachant que l'inverse de la fonction *Coût* en % nous donne le facteur de temps gagné. Cependant le choix de l'échelle de filtrage dépend dans notre approche du paramètre $n_{x,k}$

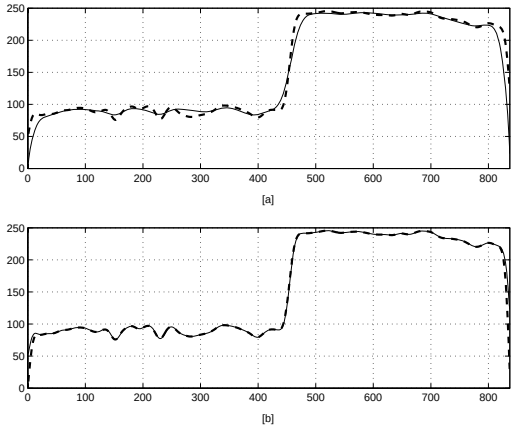


FIG. 7 – Réponses observées sur le profil de la ligne centrale de l'image *château* par trois méthodes de calcul. La courbe en pointillée est la réponse de référence (filtre gaussien avec ${\sigma_f}^2 = 2 \cdot 10^{-2} pixel^{-2}$).

des cellules. Il est important que ce paramètre ne varie pas dans des proportions importantes car le choix de i_{opt} est fonction de la taille de l'image et de la valeur minimale de $n_{x,k}$.

Les difficultés de faire évoluer la structure de filtre d'inspiration *traitement d'images*, nous ont incitées à nous orienter vers une architecture d'automates cellulaires [8]. Un première réalisation portant sur la modélisation de la couche *PLE* a déjà été menée [3]. A court terme nous envisageons, afin d'accèder à plus de puissance de calcul, d'utiliser les architectures parallèles présentes dans les processeurs graphiques (*GPU*).

Références

[1] W.H.A Beaudot. *Le traitement neuronal de l'information dans la rétine des vertébrés – Un creuset d'idées pour la vision artificielle.* PhD thesis, Institut National Polytechnique de Grenoble, France, 1994. 228 pages.

[2] A. Chehikian. Fast computation of optimal semi-octave pyramids. *7th S.C.I.A.*, August 1991.

[3] F. Devillard. Implémentation par automates cellulaires d'une modélisation architecturale de rétine biologique. *R.E.A.D. 2005*, Juin 2005.

[4] J.-M. Jolion. *Les systèmes de vision.* Traitement du Signal et de l'Image. HERMES Science Europe, Paris, 2001.

[5] David Marr. *Vision.* W. H. Freeman and Company, New York, 1982.

[6] J.-F. Risse. *Exploration de la fonction visuelle – Applications au domaine sensoriel de l'œil normal et en pathologie.* Masson - Société Française d'Ophtalmologie, 1999.

[7] W.F. Schreiber. *Fundamentals of Electronic Imaging Systems – Some Aspects of Image Processing.* Springer-Verlag, 3rd edition, 1993.

[8] S. Wolfram. *A new kind of science.* Wolfram Media Inc., 1st edition, 2002.

Optimisation de la largeur des opérateurs arithmétiques pour la synthèse de haut niveau

Nicolas HERVÉ , Daniel MENARD , Olivier SENTIEYS

Équipe de recherche R2D2
IRISA/ENSSAT
6, rue de kérampont
22300 Lannion FRANCE
eMail : *name*@enssat.fr

Résumé – Les applications embarquées de traitement numérique du signal requièrent l'utilisation de l'arithmétique virgule fixe pour satisfaire les contraintes de coût et de consommation d'énergie. Pour réduire le temps de développement des applications, des outils de haut niveau permettant d'automatiser la conversion en virgule fixe sont nécessaires. Dans cet article, une nouvelle méthodologie d'implantation d'algorithmes au sein d'une plate-forme matérielle utilisant l'arithmétique virgule fixe est proposée. L'objectif de cette méthode est de minimiser la surface de l'architecture synthétisée sous contrainte de précision des calculs. Notre approche permet de coupler, à travers un processus itératif, l'optimisation de la largeur des opérateurs et la synthèse d'architecture.

Abstract – Most of embedded signal processing applications uses fixed-point arithmetic to satisfy cost and power consumption constraints. To reduce the time-to-market, high-level tools enabling fixed-point conversion automatically, are required. In this article, a new methodology for hardware implementation using fixed-point arithmetic is proposed. The objective is to minimise the chip area under a computation accuracy constraint. Our approach links the operator word-length optimisation and the architecture synthesis through an iterative process.

1 Introduction

Les architectures reconfigurables telles que les FPGA représentent un bon compromis entre un niveau de performance élevé et une flexibilité de développement. Dans un contexte de systèmes embarqués où les contraintes de coût, de consommation et d'encombrement sont primordiales [2, 3], l'implantation des algorithmes de traitement numérique du signal nécessite l'utilisation de l'arithmétique virgule fixe. En effet, les opérateurs en virgule fixe ne travaillent que sur des entiers alors que les opérateurs en virgule flottante doivent manipuler la mantisse et l'exposant associés à chaque donnée. En conséquence, la largeur des opérateurs arithmétiques, des bus et des mémoires est plus faible dans les architectures virgule fixe, ce qui permet de diminuer le coût en surface, le temps d'exécution des opérations et la consommation d'énergie. En contrepartie, l'utilisation de l'arithmétique virgule fixe implique de déterminer le codage de chaque donnée de l'application. Cette tâche de codage est longue, fastidieuse et source d'erreurs. De plus, la réduction du temps de mise sur le marché des applications exige l'utilisation d'outils de développement de haut niveau, permettant d'automatiser certaines tâches.

L'implantation efficace d'un algorithme au sein d'une plate-forme matérielle (ASIC, FPGA) nécessite de minimiser la surface et la consommation d'énergie. Ainsi, dans le cadre d'une implantation matérielle en virgule fixe, la stratégie est de déterminer la largeur des opérateurs permettant de minimiser la surface de l'architecture tant que la contrainte de précision associée à l'application est satisfaite. Un problème se pose du fait que, d'une part, la synthèse d'architecture nécessite la connaissance de la largeur des opérations et que, d'autre part, l'optimisation de la largeur des opérateurs requiert la connaissance de l'assignation des opérations aux opérateurs. En conséquence, pour atteindre au mieux l'objectif, le processus d'optimisation doit être couplé avec la synthèse d'architecture.

Uniquement la méthode proposée dans [5] réalise un couplage entre les processus de synthèse d'architecture et d'optimisation de la largeur des données. De plus, de nombreuses méthodes utilisent une technique d'évaluation de la précision basée sur la simulation conduisant à des temps de simulation prohibitifs.

Dans cet article, une nouvelle méthodologie de synthèse d'architecture sous contrainte de précision est proposée. Un processus itératif d'optimisation de la largeur des données et de synthèse d'architecture est mis en œuvre afin de combiner les effets de ces deux taches. De plus, l'évaluation de la précision est réalisée par une approche analytique permettant d'obtenir des temps d'optimisation beaucoup plus raisonnables qu'avec les approches basées sur la simulation. Dans la section 2, les principales méthodes de synthèse d'architecture en virgule fixe sont présentées. Notre méthodologie et les différentes phases la constituant sont détaillées dans la section 3. Enfin, pour montrer l'intérêt de notre approche, les résultats d'expérimentations sont fournis dans la section 4.

2 État de l'art

L'approche classique utilisée pour optimiser la largeur des données consiste à représenter toutes les données dans un format unique [6]. Ceci réduit l'espace de recherche à une seule dimension et simplifie la synthèse car toutes les opérations s'exé-

cuteront sur des opérateurs de même largeur. Cependant, dans le but d'obtenir une implantation efficace d'un algorithme de traitement du signal en virgule fixe, tout en respectant la contrainte de précision de calcul imposée par l'environnement, il est nécessaire de considérer un format propre à chaque signal [1].

La méthodologie proposée par Constantinides et *al* [1] se compose de deux étapes principales. La première permet d'obtenir une spécification virgule fixe respectant une contrainte de précision. La seconde, correspond à la synthèse de l'architecture. Cette méthodologie réalise l'implantation d'une spécification virgule fixe de l'application conduisant à une précision des calculs souvent nettement supérieure à la contrainte de précision. En effet, la seconde phase affectant des opérations sur des opérateurs de largeur plus importante, cela se traduit par une augmentation de la précision des calculs.

Dans [5], les auteurs proposent une méthode pour laquelle, la synthèse d'architecture est effectuée entre deux phases d'optimisation de la largeur des données. Une première étape analyse le graphe flot de signal de l'application pour former des groupes de données. Ensuite, une étape détermine la largeur de données minimale requise pour chaque groupe. Cette combinaison de largeur est utilisée dans une troisième étape de synthèse pour ordonnancer et assigner une spécification virgule fixe. La dernière étape recherche la largeur des opérateurs optimale. Enfin, le processus de synthèse et d'optimisation de la largeur des données doit être itératif et se terminer par une synthèse permettant d'implanter exactement la spécification virgule fixe optimisée pour la contrainte de précision donnée.

3 Synthèse sous contrainte de précision

La méthodologie d'implantation automatique d'algorithmes spécifiés en virgule flottante au sein de FPGA utilisant l'arithmétique virgule fixe est détaillée à la figure 1. Le code C virgule flottante représentant l'application est tout d'abord transformé à l'aide de l'outil SUIF en une représentation intermédiaire correspondant à un graphe flot de données et de contrôle. Cette représentation intermédiaire est ensuite convertie en un graphe flot de signal (GFS). Les différentes phases de conversion en virgule fixe et de synthèse de l'architecture sont effectuées à partir de ce GFS.

La première phase du processus de conversion en virgule fixe correspond à la détermination de la position de la virgule permettant de définir le nombre de bits associés à la partie entière de chaque donnée.Ensuite, la largeur des différentes données est déterminée et le nombre de bits pour la partie fractionnaire en est déduit.L'objectif de cette seconde phase est d'optimiser la largeur des opérateurs pour une contrainte de précision donnée, de manière à obtenir le coût (e.g. surface) le plus faible. Un environnement permettant d'effectuer conjointement l'optimisation de la largeur des opérateurs et la synthèse d'architecture est en cours de développement (c.f. fig. 1).

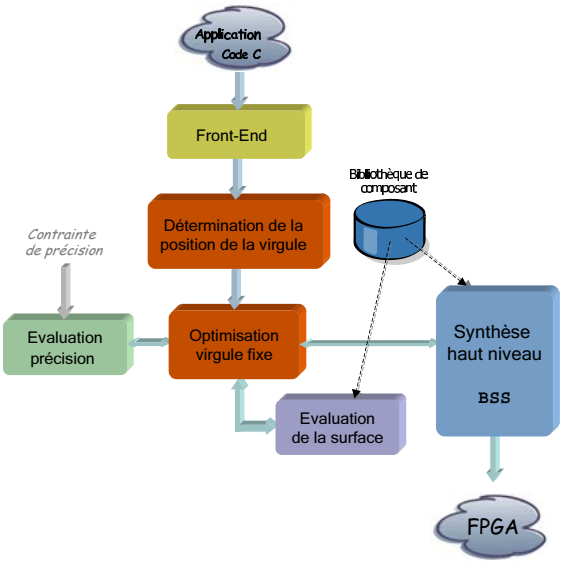


FIG. 1: *Environnement conjoint de synthèse d'architecture et d'optimisation des largeurs d'opérateurs*

3.1 Détermination de la position de la virgule

La première étape correspond à la détermination de la position de la virgule de chaque donnée. Le choix de cette position doit permettre de représenter toutes les valeurs prises par cette donnée (garantir l'absence de débordement) et minimiser le nombre de bits pour la partie entière de la donnée. Ce choix est réalisé à partir de la connaissance de la dynamique des données. Cette dernière est évaluée à l'aide d'une méthode analytique basée sur l'arithmétique d'intervalle [4] pour les systèmes non-récursifs [1] et la norme L1 [10] pour les systèmes linéaires invariants dans le temps.

Ces résultats sont directement utilisés pour définir la position de la virgule. Pour déterminer la position de la virgule des données en entrée et en sortie des opérateurs, des règles de propagation du format sont définies pour chaque type d'opération. Ensuite, ces règles sont appliquées à chaque opération lors d'un parcours du graphe flot de données (GFD) de l'application. Des opérations de recadrage sont insérées afin d'adapter le format de codage de la donnée à sa dynamique ou pour adapter le format des entrées d'un additionneur.

3.2 Synthèse d'architecture à largeurs multiples

La synthèse de l'architecture à largeurs multiples est effectuée après la détermination de la position de la virgule et va permettre de déterminer le nombre de bits nécessaires pour coder la partie fractionnaire de chaque donnée. L'objectif de cette phase est de minimiser la surface du circuit en réduisant la largeur des opérateurs. La contrainte de précision des calculs associée à l'application doit bien sûr être vérifiée. Ce processus itère tant que la contrainte de précision est vérifiée et doit évaluer la précision des calculs et la surface de la partie opérative à chaque itération.

Afin d'obtenir des temps d'optimisation du codage des don-

1. Dans une structure récursive, la sortie dépend des entrées et des échantillons précédents de la sortie. Pour une structure non-récursive, la sortie dépend uniquement des entrées.

nées raisonnables, une méthode analytique d'évaluation de la précision [8, 7] est utilisée. Cette approche détermine automatiquement l'expression analytique de la puissance du bruit de quantifi cation et de la puissance du signal pour des systèmes linéaires invariants dans le temps ou non-linéaires non-récursifs.

L'outil de synthèse utilisé [9] s'appuie sur une bibliothèque d'opérateurs arithmétiques virgule fi xe associée à une famille de FPGA donnée. Chaque opérateur est caractérisé en termes de surface, de temps d'execution et de consommation énergétique en fonction des différentes largeurs des entrées et de la sortie. La surface globale de la partie opérative est simplement la somme des surfaces de chaque opérateur.

3.2.1 Description du processus itératif

La méthodologie proposée pour la synthèse d'architecture à largeurs multiples est un processus itératif permettant de coupler effi cacement la synthèse d'architecture et l'optimisation de la spécifi cation en virgule fi xe. Une itération du processus réalise successivement, le regroupement des opérations, l'optimisation de la largeur des groupes et la synthèse de l'architecture.

Pour commencer, la largeur propre de chaque opération est déterminée. Celle-ci est obtenue en optimisant la largeur des opérateurs pour une implantation spatiale. Dans ce cas, un opérateur est affecté à chaque opération.

Constitution de groupes d'opérations L'objectif de la première étape du processus d'optimisation est de constituer des groupes d'opérations. Un groupe rassemble les opérations d'un même type qui auront la même largeur. Tout d'abord, le nombre de groupes utilisés pour chaque type d'opérateur arithmétique est déterminé. Pour la première itération, tous les opérateurs d'un même type seront mis à la même largeur. Ainsi, un seul groupe est associé à chaque type d'opération arithmétique. Pour les itérations suivantes, le nombre de groupes sera défi ni à partir de l'analyse des résultats de la synthèse d'architecture.

Ensuite, un algorithme de groupement est appliqué afi n d'affecter chaque opération à un groupe. Ce groupement des données est fait à partir de l'analyse des résultats de la synthèse d'architecture, de la largeur propre de chaque opération et des dépendances de données.

Notre technique de groupement se base sur la *mobilité* des opérations défi nie comme la différence des indices de cycle obtenus lors de deux ordonnancements par liste : un dans le sens directe et l'autre dans le sens inverse. Les opérations dont la mobilité est la plus faible sont traitées en premier et défi nissent les largeurs initiales associées à chaque groupe. La largeur d'un groupe donné correspond à la largeur maximale des opérations associées à ce groupe. Les opérations sont traitées ensuite par ordre décroissant de leur largeur propre. La mobilité de ces opérations est utilisée pour les déplacer en vue des les associer à un groupe disponible et de largeur la plus faible possible. L'objectif de cette approche est d'obtenir pour chaque groupe la largeur la plus faible.

Optimisation de la largeur des groupes d'opérations La troisième étape est l'optimisation de la largeur des différents groupes d'opérations. L'objectif est de minimiser la surface du circuit en respectant la contrainte de précision. Cette étape utilise les modules d'estimation de la précision et de la surface. L'algorithme d'optimisation se base sur la dérivée de la fonction de coût et de précision pour déterminer la direction à suivre.

Synthèse de l'architecture À partir de la spécifi cation virgule fi xe obtenue, une synthèse d'architecture est ensuite réalisée. Cette synthèse, effectuée à l'aide de l'outil BSS [9], va pouvoir remettre en cause le nombre d'opérateurs utilisés pour chaque type d'opération. En effet, par rapport à la synthèse précédente, la largeur de certains groupes d'opérations peut avoir diminué, entraînant également une réduction du temps d'exécution des opérations associées. Cette diminution peut conduire à une réduction du nombre d'opérateurs lors de la synthèse, ce qui justifi e l'utilisation d'un processus itératif. En pratique, le nombre d'itérations est relativement faible. Le nombre de groupes utilisés pour l'itération suivante correspond au nombre d'opérateurs obtenus lors de cette phase de synthèse.

4 Expérimentations

Pour montrer l'intérêt d'une technique d'optimisation de la largeur des opérateurs en fonction d'une contrainte de précision des calculs, un fi ltre à réponse impulsionnelle infi nie (IIR) d'ordre 8 a été testé. Ce fi ltre est implanté sous forme de 4 fi ltres cascadés d'ordre 2. Le graphe flot de signal de ce fi ltre récursif est présenté à la fi gure 3. Cette application est composée de 36 opérations correspondant à 20 multiplications et 16 additions.

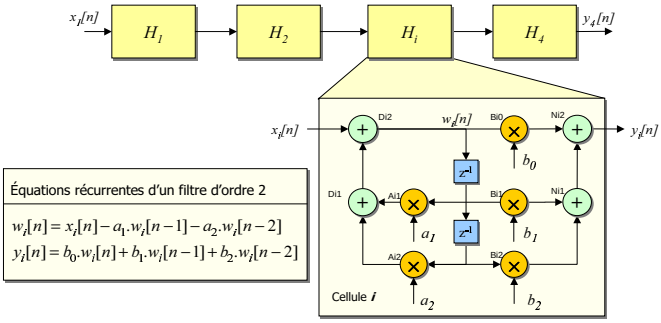


FIG. 3: *Synoptique du filtre IIR d'ordre 8 et graphe flot de signal d'une cellule d'ordre 2. Le filtre IIR d'ordre 8 est composé de 4 cellules d'ordre 2 cascadées.*

La méthodologie présentée dans cet article a été utilisée pour obtenir la dynamique et la position de la virgule de chaque donnée afi n d'obtenir une spécifi cation en virgule fi xe correcte. De plus, l'expression analytique du rapport signal à bruit de quantifi cation a été déterminée et la contrainte de précision a été fi xée à 60 dB.

La largeur propre associée à chaque opération est présentée à la fi gure 2.a. Cette largeur est utilisée comme référence lors du processus de regroupement des données. Pour la première itération du processus global d'optimisation, un groupe est associé à chaque type d'opérateur et la largeur de celui-ci est fi xée à la largeur maximale des opérations de ce groupe. Ainsi, les

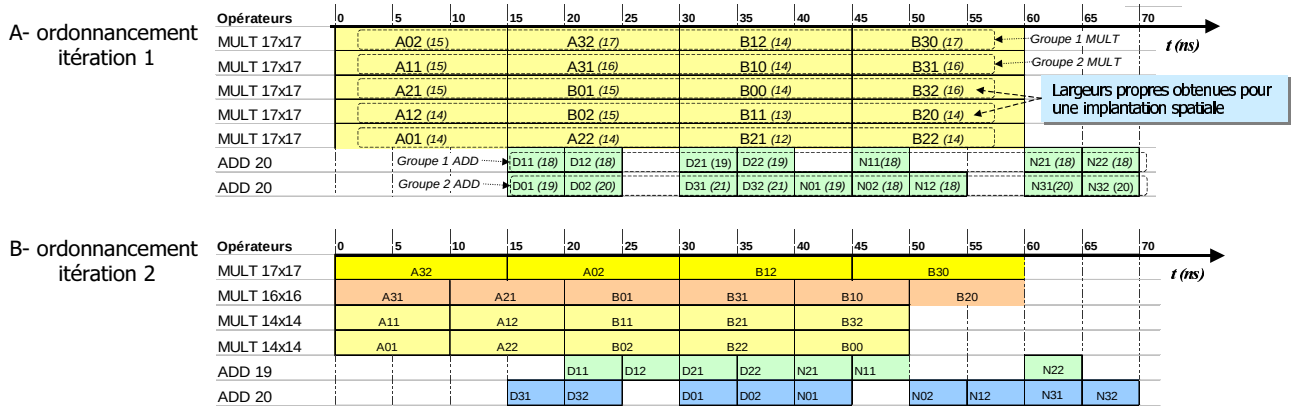


FIG. 2: *Ordonnancements et largeurs d'opérateurs obtenus pour le filtre IIR lors des deux premières itérations du processus.*

multiplications sont réalisées sur un multiplieur 17×17 bits et les additions par un additionneur 20 bits. La synthèse d'architecture pour cette spécification virgule fixe conduit à l'ordonnancement présenté à la figure 2.a. La fréquence minimale de l'horloge du système est fixée à 200 MHz. Ainsi, la latence de chaque opérateur est un multiple de $5\ ns$ Pour une contrainte de temps de $70\ ns$, 5 multiplieurs et 2 additionneurs sont nécessaires. Suite à cette synthèse, 5 nouveaux groupes sont définis pour les multiplications et 2 pour les additions. Ces groupes, représentés à la figure 2.a, sont formés en fonction de la largeur obtenue pour une implantation spatiale et de la mobilité des opérations.

Une optimisation de la largeur des groupes sous contrainte de précision est effectuée pour ces 7 groupes d'opérations. Cette optimisation permet d'obtenir des groupes de largeur plus faible. Les largeurs des 5 groupes pour les multiplications sont égales à 17, 16, 15, 14 et 14 bits et les largeurs pour les 2 groupes d'addition sont égales à 20 et 17 bits. La synthèse de l'architecture pour cette nouvelle spécification virgule fixe conduit à l'ordonnancement présenté à la figure 2.b. Les multiplieurs de largeur 14 à 16 bits ayant une latence plus faible ($10\ ns$), quatre multiplieurs seulement sont maintenant nécessaires. Ainsi, cette architecture utilise un multiplieur de moins que celle obtenue à la première itération. La réduction de la largeur de certains opérateurs combinée à la diminution du nombre d'opérateur, conduit à un gain au niveau de la surface des opérateurs de 35%.

La méthode proposée dans [6] réalise une architecture pour laquelle une largeur unique d'opérateur est utilisée. Pour cette application, 5 multiplieurs 19×19 et 2 additionneurs 19 bits sont donc utilisés. Par rapport à cette architecture, la surface des opérateurs obtenue avec notre approche est réduite de 47%. Ces résultats montrent l'intérêt d'utiliser des largeurs multiples pour les opérateurs et l'efficacité de notre approche.

5 Conclusion

Une nouvelle méthodologie de synthèse d'architecture pour FPGA sous contrainte de précision est décrite dans cet article. Pour obtenir une architecture optimisée en termes de largeur des opérateurs, un processus itératif d'optimisation de la largeur des données et de synthèse d'architecture est mis en œuvre afin de coupler ces deux processus. Pour obtenir des temps d'optimisation raisonnables l'évaluation de la précision est réalisée par une approche analytique.

Les premiers résultats obtenus avec cette approche montrent l'intérêt d'utiliser des largeurs multiples pour les opérateurs. Dans le cas du filtre IIR présenté, notre méthode permet tout en respectant la contrainte de temps et de précision de réduire la surface de la partie opérative de l'architecture de 47% par rapport à une approche classique pour laquelle les différents opérateurs possèdent la même largeur.

Ces travaux s'inscrivent dans le cadre du projet RNTL OSGAR (Outils de Synthèse Génériques d'Architectures Reconfigurables), en collaboration avec le CEA LIST à Saclay, l'UBO à Brest et la société TNI-Valiosys.

Références

[1] G. A. Constantinides, P. Y. K. Cheung, and W. Luk. *Synthesis and Optimization of DSP Algorithms*. Kluwer Academic, 2004.

[2] J. Eyre and J. Bier. The Evolution of DSP Processors. *IEEE Signal Processing Magazine*, 17(2):44–51, March 2000.

[3] G. Goossens and al. Embedded Software in Real-Time Signal Processing Systems: Design Technologies. *Proceedings of the IEEE*, 85(3):436–453, March 1997.

[4] R. Kearfott. Interval Computations: Introduction, Uses, and Resources. *Euromath Bulletin*, 2(1):95–112, 1996.

[5] K. Kum and W. Sung. Combined Word-Length Optimization and High-level Synthesis of Digital Signal Processing Systems. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 20:921–930, August 2001.

[6] E. Martin, J. Toureilles, and C. Nouet. Conception optimisée d'architectures en précision finie pour les applications de traitement du signal. *Traitement du Signal 2001*, 18(1):47–56, 2001.

[7] D. Menard, R. Rocher, P. Scalart, and O. Sentieys. SQNR determination in non-linear and non-recursive fixed-point systems. In *XII European Signal Processing Conference (EUSIPCO 2004)*, pages 1349–1352, Vienna, Austria, September 2004.

[8] D. Menard and O. Sentieys. A methodology for evaluating the precision of fixed-point systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2002)*, Orlando, May 2002.

[9] O.Sentieys, J.P.Diguet, and J.L.Philippe. GAUT: a High Level Synthesis Tool dedicated to real time signal processing application. In *University Booth, EURO-DAC*, Brighton, 18-22 september 1995.

[10] T.W. Parks and C.S. Burrus. *Digital Filter Design*. John Willey and Sons, 1987.

Réalisation d'une IP de scrutation et de codage entropique pour un codeur MPEG4

Khalil Hachicha[1], Patrick Garda[1]

[1]Universite Pierre et Marie Curie LISIF - SYEL - B.C. 252 4, place Jussieu - 75252 PARIS CEDEX 05

khalil.hachicha@lis.jussieu.fr

Résumé – Dans cet article, nous décrivons la conception et la réalisation d'une IP de scrutation et de codage entropique. Notre approche basée sur un traitement en parallèle a permis une réduction significative du chemin critique. La conception est entièrement synchrone et contient 4000 portes logiques. Le format de sortie est compatible avec celui du codage à longueur variable pour un codeur vidéo MPEG4 standard. L'architecture synthétisée par Précision RTL, est très performante pour la compression vidéo en temps réel.

Abstract *– In this paper, we describe the design and the implementation of the scan and the entropic coding IP. The realized architecture, synthesized by Precision RTL, offers high performances for real time video compression. A novel approach allowing the reduction of the critical path and based on parallel coding is presented. The design is fully synchronous and contains 5000 logic gates. The format of the output is compatible with the variable length coding for MPEG4 video coder.*

1. Introduction

MPEG4 part 9[1] est un projet international regroupant plusieurs laboratoires de recherche qui conçoivent des IP pour les différents blocs d'un codeur MPEG4 dans le but d'exécuter le codage de séquence grand format en temps réel. Cette bibliothèque servira de base à la normalisation pour l'ISO d'un codeur matériel pour MPEG4. La plateforme choisie est une carte WILDCARDII [2] contenant un FPGA VIRTEX de Xilinx et les protocoles de communication ont été développés et testés. Notre participation à ce projet porte sur la conception et la réalisation d'une IP de scrutation suivie du codage entropique.

La scrutation consiste à changer la disposition des différents coefficients d'un bloc de telle façon que ces derniers deviennent disposés dans un ordre allant des coefficients reflétant les basses fréquences jusqu'aux coefficients reflétant les hautes fréquences. Cette disposition plus compacte des données permet d'avoir une plus grande probabilité de valeurs non nulles suivies de valeurs plus ou moins nulles correspondant aux hautes fréquences. Le codage entropique est ensuite réalisé en deux étapes : une première consiste à mettre les données scannées sous forme de couples « run » et « level ». Le « run » correspond aux nombres de 0 qui se suivent et le « level » correspond aux nombres différents de zéro venant juste après cette succession de zéros. Il est clair que plus il y a de zéros qui se succèdent, moins on aura de couples à coder et la compression sera plus efficace. La deuxième étape consiste à retrouver pour chaque couple (run, level), à partir des différents tableaux de codage mis à disposition, un code correspondant qui va être écrit dans le bitstream. Pour cela, nous retrouvons plusieurs tableaux dont chacun est utilisé pour un cas bien spécifique. Ainsi, nous retrouvons des tableaux utilisés en cas de codage : codage Inter, codage Inter fin bloc, codage Intra et codage Intra fin bloc. La différenciation entre ces cas s'appuie sur le fait que les différentes probabilités de succession des données change considérablement dans ces 4 cas. Prenons le cas où nous sommes à la fin du bloc. Il y a plus de probabilité à retrouver beaucoup de zéros qui se succèdent. Ainsi attribuer un code plus petit à un couple (run, level) ayant un grand run est plus astucieux.

Les travaux précédents dans la conception de circuits pour le codage entropique se sont concentrés exclusivement sur la réalisation d'un débit binaire élevé à la sortie. Lei et al [3] ont proposé une architecture employant des opérations parallèles pour transformer les données d'entrée en codes, les enchaîner ensemble et les segmenter en mots binaires de 16 bits. Stephen Arthur [4] a adopté une approche de basse consommation. Il a prouvé que la recherche de codes dans les différentes tables est la source dominante de consommation. En conséquence, il a présenté une nouvelle approche de partition des tables basée sur la probabilité d'occurrence des symboles. L'inconvénient majeur de ces architectures est imputé au fait qu'elles gèrent uniquement un seul mode de codage qui se fait directement à partir de tableaux VLC. Ceci n'est pas suffisant pour assurer l'attribution de code pour chaque couple (run, level) possible. Nous nous sommes proposés ainsi, de concevoir une architecture d'un codeur complet tenant compte des trois modes de codage tels qu'ils sont proposés par la norme MPEG4[5] et permettant de gérer tous les cas possibles.

Dans cet article, une architecture d'encodeur parallèle complet sera présentée ; nous détaillerons les unités de scrutation, de prétraitement, de contrôle, de codage et finalement l'unité de paquetage. Nous donnerons le diagramme de fonctionnement de l'unité principale de codage et les différentes optimisations pour accélérer le traitement dans cette unité. Enfin nous présenterons les résultats et analyserons l'architecture proposée.

2. Description de l'IP

Pour concevoir l'IP du module de scrutation et de codage entropique, la première étape de notre travail a été de décomposer notre IP en différentes unités afin de faciliter le travail de test. La décomposition s'est faite d'une manière judicieuse en tenant compte de la façon dont l'algorithme est organisé (figure 1)

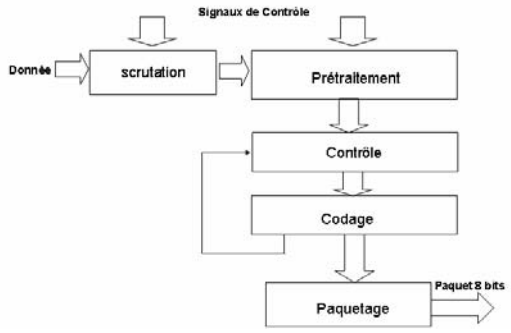


FIG. 1 Schéma des blocs

1.1 Unité de scrutation

Cette unité peut effectuer la scrutation de trois manières différentes : la scrutation verticale, la scrutation horizontale et la scrutation normale (Figure 2).

Le type à choisir dépend de l'utilisation ou non de la prédiction des coefficients Ac. Trois cas se présentent :

- pas de prédiction de coefficients Ac : mode de scrutation normal,
- prédiction des coefficients Ac à partir de la ligne 1 : mode de scrutation horizontal,
- prédiction des coefficients Ac à partir de la colonne 1 : mode de scrutation vertical.

Dans les 64 premiers coups d'horloge, nous chargeons le premier bloc dans une zone mémoire. Le chargement se fait en série (échantillon par échantillon). Aucune donnée n'est envoyée vers le deuxième bloc durant ce temps là. Le type de scrutation est décidé par l'intermédiaire de 2 signaux de contrôle qui reflètent le mode de prédiction choisi pour les coefficients Ac. Une fois ces 64 cycles écoulés, l'envoi des données scrutées, correspondant au premier bloc commence, en même temps que le deuxième bloc est chargé. Ce travail se poursuit jusqu'au chargement complet des 6 blocs.

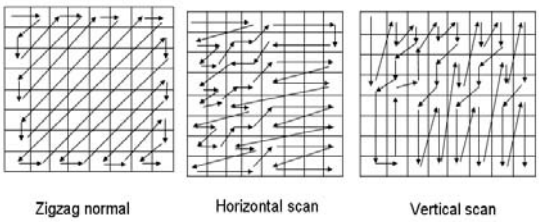


FIG. 2 : Les différents types de scrutation

1.2 Unité de prétraitement

Le bloc de prétraitement fournit à celui du codage tous les signaux nécessaires dans le but de choisir les tableaux de codage qui doivent être utilisés. Dans cette unité, la première étape consiste à déterminer le type de données que l'on vient de recevoir à savoir :

- est il Ac ou Dc ?
- correspond t il à la luminance ou à la chrominance ?
- est il le dernier échantillon dans le bloc ?

La connaissance de ces détails est possible puisque nous connaissons l'ordre d'arrivée des données. En effet, les quatre premiers blocs correspondent aux données de luminance, les deux derniers correspondent à ceux de chrominance. Dans chaque bloc, le premier échantillon correspond au coefficient Dc et ceux qui suivent à des coefficients Ac. Le 63ème échantillon correspond aux derniers échantillons dans le bloc. Nous avons ainsi créé des compteurs d'échantillons et de blocs qui permettent de décider le type d'échantillon que nous avons à traiter. Ces compteurs sont créés dans des processus à part. Un autre processus calcule, pour chaque échantillon, sa valeur absolue et un troisième processus s'intéresse aux calculs du run comptant le nombre de zéros qui se succèdent. Le dernier processus contient une machine à état interne qui change d'état suivant le type de données à traiter. Si le coefficient est un Dc, nous envoyons seulement le « level » de l'échantillon. Si le coefficient est un Ac, nous envoyons le run calculé, le complément A2 du « level » et le bit de signe du « level».

1.3 Unité de contrôle

Le bloc de contrôle est nécessaire pour deux raisons. La première est que le bloc de traitement envoie les données correspondant aux couples (run, level) d'une manière non continue. La deuxième est que le nombre de cycles nécessaires pour trouver le code correspondant aux couples envoyés varie selon le mode de codage : normal, escape mode 1, escape mode 2 ou escape mode 3. Ainsi, l'unité de contrôle reçoit des signaux du bloc de codage pour savoir si ce dernier peut recevoir de nouvelles données et permet la mémorisation de toutes les données que le bloc de codage ne peut pas traiter à un instant précis.

1.4 Unité de codage

Le troisième bloc, est celui du codage, dans lequel nous essayons de trouver les différents codes.

- Pour les coefficients Dc, nous écrivons dans la mémoire le code correspondant au nombre de bits sur lesquels ils sont définis, suivi de la valeur du coefficient DC,
- pour les coefficients Ac, plusieurs paramètres entrent en jeux dans le codage : est ce le dernier couple ? est ce que le codage est en mode Inter ou Intra ? Ces

informations sont récupérées par l'intermédiaire de différents signaux issus du bloc de prétraitement.

Quatre tableaux VLC pour quatre modes de codage : le mode « Inter », « Inter_last », « Intra » et « Intra_last » sont prévue pour ce fait. Nous avons stocké ces tableaux dans différents emplacements mémoires et pour chaque couple de données, nous essayons de voir si le couple possède un code correspondant dans la mémoire ou non.
Nous avons rencontré ici un problème dû au fait que les codes correspondant aux différents couples sont de longueur variable. C'est pour cette raison que pour chaque code, un mot binaire de 4 bits, indiquant la taille du code que nous devons récupérer, est ajouté au code lui même. Suite à la récupération du code, nous écrivons le résultat dans un tampon prévu à cet effet pour stocker le flux. Chaque fois qu'un code est écrit, un compteur s'incrémente du nombre de bits écrits.
Il est à noter que la recherche d'un code correspondant à un couple de données peut échouer. Ceci engendre de nouveaux calculs et traitements et par conséquent des retards qu'il faut absolument prendre en compte lors de la conception. Nous avons essayé de développer une architecture permettant d'avoir les meilleures performances. Le séquencement des traitements et l'optimisation de l'architecture sont détaillés dans le paragraphe suivant.

1.4.1 Explication détaillée

Cycle 1 :

- traiter le coefficient Dc,
- chercher le code correspondant au nombre de bits sur lesquels le coefficient Dc est écrit.

Cycle 2 :

Pour le coefficient Dc, nous écrivons le code trouvé :
- écrire le code de la dimension du coefficient Dc,
- écrire la valeur du coefficient,
- écrire le signe du coefficient.

Pour les coefficients Ac, nous :
- lisons le code correspondant au 1er couple (run, level),
- mémorisons le run et level dans deux tampons.

Cycle 3 :

Pour le coefficient Ac, deux situations sont possibles :

1er cas :

Si la longueur du code lu pendant le cycle 2 est nulle, cela signifie que le code correspondant au couple précédent n'a pas été trouvé. Dans ce cas, nous passons au mode « escape 1 ». Un signal « busy » est mis à 1 et envoyé à l'unité de prétraitement pour qu'elle arrête l'envoi de nouveaux couples.

Nous recalculons alors le couple (run, level) en modifiant la valeur du level. La nouvelle valeur du level est calculée à l'aide de sa valeur mémorisée pendant le cycle 1 et de la valeur « escape_level » déterminée en fonction de la valeur du run. La détermination de cette valeur se fait à partir de tableaux qui dépendent de la nature du codage (Inter ou Intra) et de la position du couple (run, level) dans le bloc à coder (dernier ou non).

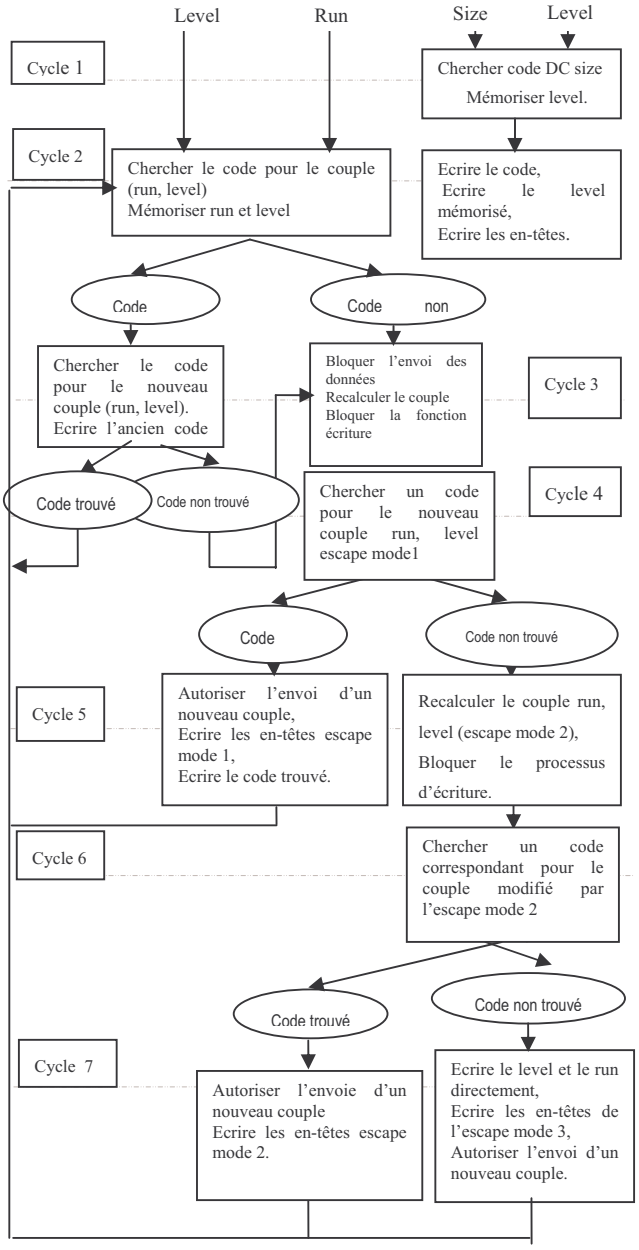


FIG. 3 Diagramme de l'unité de codage

2ème cas

Si la longueur du code lu au cours du cycle 2 est différente de 0, alors le code correspondant au 1er couple (run, level) est trouvé dans la mémoire. Dans ce cas, nous stockons le code correspondant dans un tampon et nous activons la fonction

écriture pour le cycle suivant. En même temps nous récupérons le nouveau couple (run, level).

Cycle 4 :

Si nous sommes dans le mode escape numéro 1, nous recherchons le code correspondant au couple (run, level) re-calculé au cours du cycle 3.

Cycle 5 :

– 1er cas :
Si le code correspondant au nouveau couple re-calculé n'a pas été trouvé, alors aucune opération de lecture ou d'écriture n'est effectuée lors du cycle suivant. Le drapeau «busy» reste à 1. Nous re-calculons alors le couple en changeant cette fois-ci la valeur du run à partir des tableaux produits à cet effet et nous passons ainsi à l'« escape mode 2 ».
– 2ème cas :
Si le code correspondant au couple recalculé est trouvé, nous repassons au mode normal et le signal busy est remis à 0. Dans le cycle suivant, nous récupérons le nouveau couple et nous activons la fonction écriture pour écrire dans le tampon, le code trouvé avec les différentes en-têtes correspondant à « l'escape mode 1 ».

Cycle 6 :

Si nous sommes dans « l'escape mode 2 », nous cherchons dans ce cycle le code correspondant au couple re-calculé une deuxième fois lors du cycle 5.

Cycle 7 :

Deux cas se présentent :
- si le code cherché au cours du cycle 6 est trouvé dans la mémoire, nous écrivons les différentes en-têtes correspondant l'« escape mode 2 » suivies du code en question. Egalement, nous autorisons l'envoi d'un nouveau couple en forçant le drapeau «busy» à 0,
- si le code cherché au cours du cycle 6 n'est pas trouvé, alors nous passons à l'« escape mode 3 ». Dans ce mode, aucune recherche de code n'est effectuée. A l'aide des différentes en-têtes correspondant à ce mode, nous écrivons le run et le level tels qu'ils viennent du bloc de prétraitement. En parallèle, nous autorisons l'envoi d'un nouveau couple de données.

1.5 Mise en forme du flux de sortie

Un problème est dû au fait qu'à la sortie, les codes ont des longueurs variables ce qui rend difficile la prédiction de la taille du flux à la sortie. Par conséquence, nous avons essayé de régulariser ce dernier en mettant les données sous forme de paquets de 8 bits. Nous avons créé un tampon dans lequel nous écrivons chaque nouveau code avec les en-têtes correspondantes. Un pointeur vers le dernier bit écrit est toujours mis à jour à chaque ajout d'un nouveau code. A chaque cycle, nous vérifions le nombre de bits existants dans le tampon. Si il est supérieur à 8, un signal « write » est activé et l'octet est envoyé à la sortie.

3. Implémentation

Ce travail a été effectué sous MENTOR. Nous avons utilisé le modèle « SIM5.0 » pour la simulation des résultats et « Précision » pour synthétiser l'architecture. Les différents résultats sont comme suit :
Cible : FPGA VIRTEX, 20 MHz (2V250fg256)
Fréquence maximale : *77* MHz

TAB. 1 : Utilisation des ressources

Ressources	utilisé	disponible	%
IOs	35	172	20.35 %
Global Buffers	1	16	6.25%
Functions Generators	2544	3072	82.85%
CLB Slices	1272	1536	82.81%
Dffs or Latches	610	3588	17.25%
Block RAMS	5	24	20.83%

4. Conclusion

Le module de codage entropique complet représente un bon exemple d'algorithme dont la complexité dépend des données. L'IP développée a été synthétisée et testée. Comparé aux autres architectures développées, notre design prend en compte de tous les « escape modes » comme cela a été spécifié dans la norme. Les résultats trouvés montrent de bonnes performances en vue d'un codage vidéo temps réel pour la TVHD. Nous exprimons nos remerciements aux laboratoires EPFL, LE2I pour leur aide et à Xilinx pour nous avoir fournit la WildcardII

Références

[1] ISO/IEC JTC1/SC29/WG11 AHG on MPEG-4 Part 9: Référence Hardware, Palma de Mallorca, Spain, October 2004
[2] "Virtex™-II X Platform FPGAs: Complete Data Sheet", DS110, Mar 5, 2004
[3] S.-M. Lei and M.-T. Sun, "An Entropy Coding System for Digital HDTV Application", IEEE Trans. On Circuit and Systems for Video Technology Vol.1 No.1, pp.147- 155, mars 1991.
[4] Stephen Arthur Molloy, Low Complexity, Low Power VLSI Architectures for Image Codec's, thesis University of California, pp 149-181, Los Angeles 1997.
[5] ISO/IEC 14496-5:2001.Coding of Audio-Visual Objects – Part 5: Reference Software, 2d Edition, 2001

Reconstruction 3D temps réel dans un VSIP

Tarik GRABA[1], Bertrand GRANADO[1], Olivier ROMAIN[1], Thomas EA[2], Andrea PINNA[1], Patrick GARDA[1]

[1]Laboratoir des Instruments et Systèmes d'Ile de France (LISIF),
Université Pierre et Marie Curie (UPMC)
4 Place Jussieu, BC 252, 75252 Paris Cedex 05, France

[2]Institut Supérieur d Electronique de Paris (ISEP)
28 Rue Notre Dame des Champs, 75006 Paris, France

tarik.graba@lisif.jussieu.fr, bertrand.granado@lisif.jussieu.fr,
olivier.romain@lisif.jussieu.fr
thomas.ea@isep.fr, andrea.pinna@lisif.jussieu.fr, garda@lisif.jussieu.fr

Résumé – Dans cet article sont décrits un système intégré pour la reconstruction du relief, les contraintes et les choix de conception, la structure du capteur de vision active et les méthodes et architectures utilisées pour obtenir une représentation 3D en temps réel.

Abstract – In this paper are described a integrated 3D system, the constraints and conception choices, the vision sensor structure and the algorithmes and architectures used to have a realtime 3D reconstruction

1 Introduction

Dans de nombreux domaines, l'information apportée par l'imagerie classique est insuffisante. Par exemple, avec l'endoscopie, l'image apporte des information sur la couleur et sur la forme des lésions possibles mais très peu sur leurs dimensions ou leurs rugosités, obligeant dans de nombreux cas à pratiquer des biopsies. Pour faciliter les diagnostics et limiter le recours à des méthodes plus invasives, une information supplémentaire sur le relief apporterait au praticien des données quantitatives, lui permettant ainsi d'affiner ses diagnostics sans avoir recours à la chirurgie.

Pour obtenir une information du relief d'un objet ou d'une scène de nombreuses solutions ont été envisagées. Les recherches menées dans ce domaine depuis de nombreuses années ont mis en avant des techniques permettant de restituer le relief qui est perdu avec la prise d'image classique [1, 2, 3].

L'utilisation des système de vision est la moins contraignante et la plus commune. Ces systèmes sont basés sur l'utilisation de capteurs d'images pour obtenir l'information nécessaire à la restitution du relief. Ces caméras sont utilisées seules dans le cas des capteurs passifs ou associées à une source d'éclairage supplémentaire pour faciliter la reconstruction du relief dans le cas des capteurs actifs.

Les algorithmes de restitution du relief ont beaucoup évolué et permettent maintenant de retrouver de manière réaliste le relief d'une scène ou d'un objet. Ces algorithmes assez complexes font appel à des méthodes de calcul lourdes qui, en général, ne permettent pas d'avoir un résultat en temps réel, surtout sur des architectures où les ressources de calcul sont réduites et la consommation énergétique limitée telles celles destinées à l'embarqué.

Ces dernières années de nouveau axes de recherches sont apparus dans la vision intégrée. Des systèmes ont été conçus, intégrant sur une même puce le capteur d'images et un système de traitement de l'information [4]. Ces VSOC (Vision System On Chip) permettent d'effectuer des traitements complexes (tels que la reconnaissance de forme, la classification, la compression vidéo...) tout en offrant un encombrement réduit et une consommation énergétique faible. Ces rétines intégrées peuvent ainsi répondre aux exigences des systèmes embarqués.

La miniaturisation et l'embarquabilité permettent d'envisager de nouvelles applications où les capteurs sont de très petite taille, autonomes et peuvent ainsi être dispersés en grand nombre dans des régions inaccessibles autrement, communiquer entre eux en des réseaux de capteurs et permettre la fusion de données. Le déploiement de capteurs de ce type, qui réalisent localement une reconstruction 3D puis communiquent leurs résultats, permettrait d'avoir rapidement une vue globale, sans zones d'ombres, de l'environnement.

Dans ce contexte, Nous proposons une nouvelle approche où la reconstruction 3D est envisagée non pas comme un problème isolé dans une chaîne de traitement mais dans le cadre plus global d'un système conçu dans sa totalité pour réaliser cette tâche de la manière la plus efficace possible. Ce système nous l'avons appelé "Cyclope", un système de vision en relief intégré basé sur le principe de stéréovision active. Il regroupe tous les organes nécessaires pour fonctionner en temps réel, c'est un système de type VSIP (Visual System In Package).

2 CYCLOPE

Cyclope est notre système de vision en relief intégré [5]. Nous avons opté pour un système de vision stéréoscopique actif. Les algorithmes de traitement pour la reconstruction 3D, dans les systèmes actifs, permettent de satisfaire les contraintes de temps réel et de précision dans la reconstruction [12]. De plus cela permet de limiter les ressources ce qui est est indispensable pour un système embarqué si l'on faible consomma-

tion.

Cyclope est composé de trois parties principales :

- un bloc d'instrumentation : pour l'acquisition de l'image et la génération du motif structuré composé :
  - d'une source de lumière cohérente (LASER) couplée à une optique de diffraction. La lumière LASER étant produite par une diode à cavité verticale VCSEL (Vertical Cavity Surface Emitting Laser) émettant dans le proche infra-rouge.
    Les principaux avantages de cette structure par rapport aux diodes lasers conventionnelles, résident dans un faible encombrement permettant une forte intégration, une faible consommation et une émission par la face.
  - d'une matrice de photocapteurs de type CMOS. Elle permet l'acquisition d'une image de texture dans le visible et une image du motif projeté dans le proche infrarouge.
- un bloc de traitement et de contrôle composé d'une partie figée (microprocesseur) et d'une partie reconfigurable (FPGA).

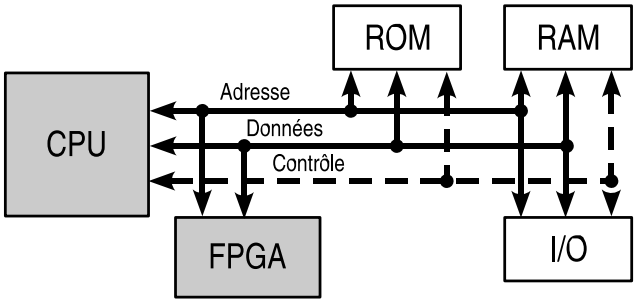


FIG. 1 – Schéma structural du bloc numérique

- un bloc de radio communication : pour la transmission des résultats et la reconfiguration dynamique du bloc numérique.

Ces trois parties sont réalisées grâce à des technologies différentes qui ne sont pas forcément compatibles entre elles, et l'intégration de celles-ci dans un circuit monopuce est de prime abord ardue. Le développement d'un boîtier intégré de type SIP-AMS (System In Package – Analog Mixed Signal) est actuellement la meilleure solution pour faire sauter les verrous technologiques qui se présentent tout en offrant des dimensions proches de celles d'un système monopuce. Cette intégration permet d'avoir des caractéristiques stables pour le couple caméra/projecteur et d'effectuer les traitements sur le circuit.

3 Vision active

Les système de vision stéréoscopique actifs peuvent être vus comme des systèmes de stéréovision classiques où une des deux caméras est remplacée par une source d'éclairage projetant un motif structuré [6, 7, 8, 9, 10]. La position des points caractéristiques du motif dans l'image permet de calculer les profondeurs par triangulation. L'information contenue dans la structure du motif permet de lever une partie de l'ambiguïté d'appariement.

Le principe de restitution de la dimension manquante reste le même qu'en vision passive [11], sauf que nous avons une connaissance à priori du motif projeté et donc des primitives à rechercher dans l'image. Cette connaissance à priori permet de réduire grandement la complexité du problème réduisant ainsi les temps de calcul nécessaires pour effectuer les traitements, et permettant de répondre à des contraintes temporelles fortes [12].

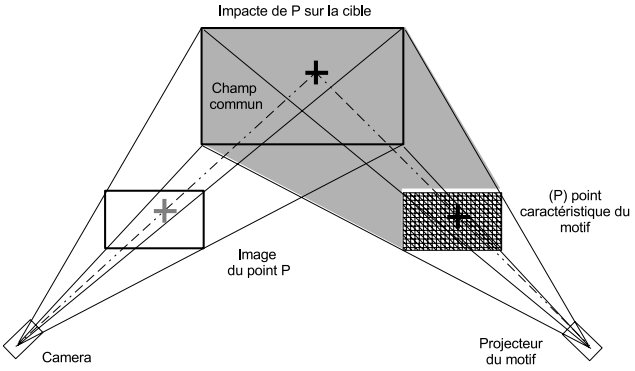


FIG. 2 – Principe de la stéréo-vision active

4 Le motif structuré

Le choix du motif projeté est très important [13, 14]. Celui-ci doit apporter suffisamment d'information pour éviter les problèmes d'appariement tout en étant assez simple à générer de manière intégrée. L'option choisie est d'utiliser une diode LASER de type VCSEL couplée à un réseau de diffraction, nous sommes limités à un motif monochrome de géométrie simple et régulière. L'utilisation d'un faisceau laser monomode dans le proche infrarouge permet d'être plus robuste aux variations d'albédo et de colorimétrie. Nous avons opté pour un maillage régulier de points, ce maillage a l'avantage d'être simple à générer. L'imageur utilisé [15] est multi-spectral, il nous permet d'extraire facilement les points de ce motif dans l'image infrarouge indépendamment de la scène et de sa texture.

5 Modélisation du couple stéréoscopique

Les systèmes de stéréovision classiques sont composé en général, de deux caméras regardant le même objet avec deux angles différents. Calibrer le système revient à trouver la relation entre les coordonnées d'un point de l'espace réel (repère monde) et ses projections dans les deux images prises par les caméras. Ceci est réalisé en trouvant le modèle de chacune des caméras. Dans notre système le problème se présente différemment, en effet, pour faciliter l'appariement nous avons opté pour un système stéréoscopique actif où l'une des caméras est remplacée par un projecteur de motif structuré, projetant une image fixe. Chaque point de la matrice projetée peut être considéré comme un point d'une image que nous connaissons à priori. La projection de ce point dans la seconde image (celle de la vraie caméra) appartient forcément à une droite, c'est la contrainte épipolaire [11]. La connaissance de ces droites permet de faciliter grandement le problème d'appariement.

En se basant sur ce principe et en tenant compte des travaux [16, 17] nous avons décidé de caractériser le couple stéréo-

scopique (projecteur-caméra) non pas par un système d'équations décrivant toutes les transformations d'un repère à l'autre mais seulement par les équations de ces droites épipolaires. Les équations de ces droites sont obtenues après une étape de calibration analogue à celle décrite dans [18]. Plusieurs prises

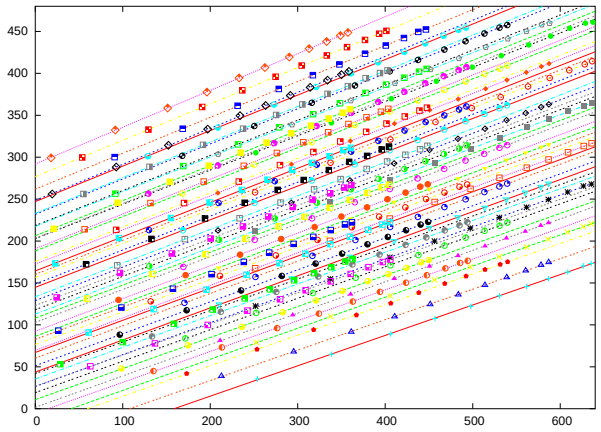


FIG. 3 – Lignes épipolaires

d'images ont été faite à différentes distances. Pour chaque image les coordonnées des impacts laser sont extraites. La trajectoire de chaque point en fonction de la profondeur donne une droite. Les paramètres de chaque droite sont obtenus par une simple approximation aux moindres carrés. La figure 3 montre les droites obtenues à partir de onze images d'un motif laser composé de 7×7 points.

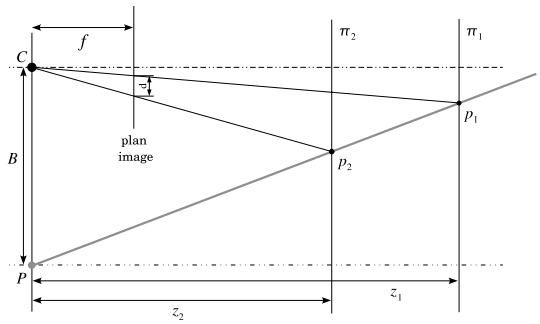


FIG. 4 – Déplacement d'un point en fonction de la distance

En plus du modèle des lignes épipolaires, nous avons dû modéliser la position de l'image des points du motif en fonction de la distance du couple stéréoscopique à l'impact de ces points dans la scène.

Si on considère un rayon laser projeté sur deux plans consécutifs se situant respectivement aux distances z_1 et z_2 (figure 4) alors le déplacement de l'image de ces impacts sur la ligne épipolaire est de la forme :

$$d = Bf\frac{(z_1 - z_2)}{z_1 z_2} \qquad (1)$$

Où B est la base du stéréoscope et f la distance focale de la caméra.

A partir de là on montre que l'on peut lier la distance à laquelle se trouve un impact LASER à l'une des coordonnées image (i, j) si l'on connaît ce point et la ligne épipolaire à laquelle il appartient, par un modèle de la forme :

$$z = \frac{1}{\alpha i + \beta} \qquad (2)$$

La même démarche expérimentale que celle utilisée pour la détermination des droites épipolaires nous a permis d'obtenir un modèle par rayon laser projeté. La figure 5 montre les courbes donnant z en fonction de i obtenues après approximation aux moindres carrés des paramètres du modèle.

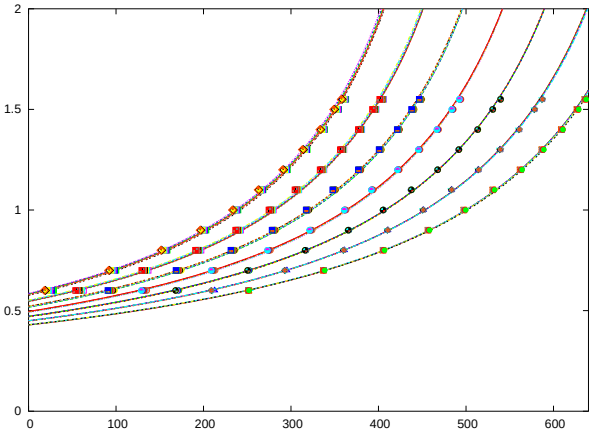


FIG. 5 – Distance en fonction de d'abscisse dans l'image

6 Architecture de reconstitution

Nous proposons (figure 6) une architecture massivement parallèle , qui à partir des coordonnées des points laser, trouve la droite épipolaire la plus proche. Cette dernière conditionne le choix du modèle de profondeur utilisé pour le calcul de la troisième dimension z .

Nous avons estimé le nombre d'opérations nécessaires pour effectuer ces opérations pour un motif de dimension N et grâce à une architecture de traitement séquentielle (un micro processeur standard).

Pour une image et un motif de N points projeté, $5N^2 + 2N$ opérations (en considérant les opérations d'addition, de multiplication, de division et de comparaison comme équivalentes) sont nécessaires.

Ce qui donne pour un motif de quatre cents points (ce qui correspond à un impact laser pour un carré de 8x8 pixel pour une image QCIF) à une cadence de $25 images/s$ une puissance de calcul nécessaire de $20 Mops$ par seconde.

7 Conclusion

Nous avons présenté dans cet article "Cyclope" un système temps réel, intégré et communicant pour la vision en relief. Nous avons présenté aussi une architecture pour la reconstitution du relief et nous avons estimé le nombre d'opérations

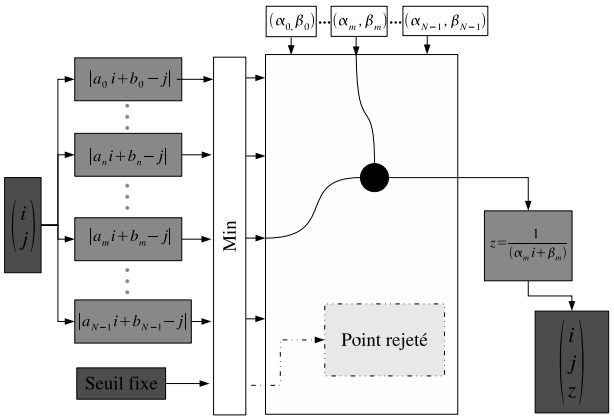


FIG. 6 – Architecture de traitement

nécessaire à la restitution du relief en temps réel et montré que ceci était compatible avec les contraintes d'exécution temps réel dans un système embarqué.

Références

[1] K. Ng & al. "An integrated multi-sensory system for photo-realistic 3d scene reconstruction", in Proceedings ISPRS Comm. V, pp. 356- 363, Hakodate, Japan, 1998

[2] V. Sequeira & al. "3D environment modeling using laser range sensing", in Robotics and Autonomous Systems, vol. 16, pp. 81-91, 1995

[3] O. Romain & al. "A multi spectral sensor dedicated to 3D spherical reconstruction", in Proceedings of IEEE ICIP 2001, Oct. 2001

[4] A. Pinna & al. "Connectionnist Retina : a neural network system integrated into an electronic retina", in the proceeding of the 25 th international congress on high speed photography and photonics, pp 218-223, 2002

[5] T. Graba & al. "Cyclope : An Integrated Real-time 3D Image Sensor", in Proc. DCIS 2004

[6] E. Cortizo & al. "Application of the structured illumination method to study the topography of the sole of the foot during a walk", in Optics and Laser in Engineering, Vol. 40

[7] F. Foster & al. "The hiscore camera – Areal time three dimensional color camera", in proceedings of IEEE ICIP 2001, Oct. 2001

[8] C. Guan & al. "Composite structured light pattern for three-dimensional video", in Optic express, Vol. 11, N. 5, pp. 406-417, 2003

[9] D. Scharstein & al. "High-accuracy stereo depth maps using structured light", in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 1, pp. 195-202, 2003

[10] T. Ea & al. "Spherical stereoscopic sensor for 3D color imaging", in Proc. SPIE Vol. 4297, pp. 91-97, jun 2001

[11] J. Battle & al. "Recent progress in coded structured light as a technique to solve the correspondence problem : a survey", in Pattern Recognition, Vol. 31, No. 7, pp. 963-982, 1998

[12] S. Boverie & al "Comparison of structured light and stereovision sensors for new airbag generation", in Control Engineering Practice, vol 11, pp. 1413-1421, 2003

[13] J. Salvi & al. "A robust-coded pattern projection for dynamic 3D scene measurement", in Pattern Recognition Letters, Vol. 19,pp. 1055-1065, 1998

[14] J. Salvi & al. Pattern "codification strategies in structured light systems", in Pattern Recognition, 2004

[15] A. Pinna & al. "Design of an integrated silicon connecsionnist retina",in the proceedings of the fifteenth annual IEEE international ASIC/SOC conference, pp 133-137, 2002

[16] A. Dipanda & al. "3-Dshape reconstruction in an active stereo vision system using genetic algorithms", in Pattern Recognition, vol. 36 pp. 2143-2159, 2003

[17] S. Woo & al. "Determination of an Optimal Configuration for a Direct Correspondence in an Active Stereovision System", in Proc. on 2nd IASTED INT. Conf. VIIP, Malaga, Espagne, pp. 596-601, September 2002

[18] F. Marzani & al. "Calibration of a three-dimensional reconstruction system using a structured light source", in Optical Engineering, Vol. 41 No. 2, pp. 484-492, February 2002

Séquenceur Mémoire pour Applications Multimédias Temps Réel Gérant les Séquences d'Accès Indéterministes

Bertrand Le Gal, Emmanuel Casseau, Sylvain Huet, Caaliph Andriamisaina, Eric Martin

Laboratoire LESTER CNRS-FRE 2734
Université de Bretagne Sud, 56000, Lorient, FRANCE
Bertrand.Legal@univ-ubs.fr, Emmanuel.Casseau@univ-ubs.fr, Sylvain.Huet@univ-ubs.fr,
Caaliph.Andriamisaina@univ-ubs.fr, Eric.Martin@univ-ubs.fr,

Résumé – Dans le domaine du traitement du signal et de l'image, les applications multimédias sont souvent caractérisées par un grand nombre d'accès aux données. Pour la plupart de ces applications, les accès aux données structurées (tableaux, vecteurs) sont réguliers et périodiques. Dans ces conditions, il est possible et efficace de générer des contrôleurs pipeline d'accès à la mémoire. Cette technique est utilisée afin d'améliorer les accès en mode pipeline autorisés par les mémoires actuelles. On utilise pour cela des composants matériels dédiés pour générer les adresses et pour packer/dépacker les données. Dans cet article nous présentons l'architecture d'un séquenceur mémoire qui permet de prendre en compte de manière efficace les accès prédictibles aussi bien que les séquences d'accès non prédictibles (calculs d'adresses dynamiques) de manière pipeline.

Abstract – Multimedia applications such as video and image processing are often characterized by a large number of data accesses. In many digital signal-processing applications, the array access patterns are regular and periodic. In these cases, it becomes feasible and efficient to generate optimized Pipelined Memory Access Controllers. This technique is used to improve the pipeline access mode to actual RAM by creating specialized hardware components for generating addresses and packing and unpacking data items. In this paper we focus on the design of a memory sequencer which can efficiently handle predictable address patterns as well as unpredictable (dynamic address computations) in a pipeline way.

1 Introduction

Les applications multimédias de type traitement du signal et de l'image (TDSI) sont caractérisées par un nombre important d'accès aux données. Dans de telles applications, les accès à la mémoire sont des facteurs limitant de la vitesse d'exécution des processeurs qui leurs sont dédiés. L'architecture mémoire (la hiérarchie, le nombre de bancs, le placement des données, etc ...) impacte sur les performances du système mais aussi sur la consommation d'énergie qui est critique pour les applications embarquées. Par ailleurs temps, le temps attribué à la conception personnalisée de circuits diminue face à la pression du "time to market".

D'un coté, les recherches actuelles dans le domaine des applications multimédias visent à réduire la complexité calculatoire des algorithmes en utilisant des solutions ad-hoc reposant sur des calculs conditionnels (par exemple, la transformation de la recherche exhaustive en recherche en 3 étapes dans le domaine de l'estimation de mouvement [5]). Ces transformations introduisent des aléas d'exécution dus aux calculs conditionnels (sélection du meilleur vecteur dans la recherche en 3 étapes).

D'un autre coté, les recherches effectuées sur les architectures d'implémentation de ces algorithmes sous contrainte temps réel, tentent d'exploiter le parallélisme entre les opérations. Dans ce cas, les architectures optimales sont obtenues pour des algorithmes réguliers, sans aléas d'exécution. Cela est aussi vrai dans le cadre des architectures mémoires qui sont optimales lorsque leurs séquences d'accès sont prédictibles.

Malheureusement dans une certaine partie des applications multimédias actuelles, les séquences d'accès à la mémoire ne sont pas entièrement connues a priori. Les aléas présents dans ces dernières empêchent les concepteurs et les outils de synthèse de haut niveau de pouvoir prendre en compte efficacement les séquences d'accès répétitives de l'application.

Dans cet article, nous apportons les contributions suivantes : nous présentons une nouvelle architecture de séquenceur mémoire qui permet de réaliser des accès pipelinés à la mémoire dans le cadre de séquences d'accès à la mémoire statiques mais aussi dynamiques. Cette technique sera ensuite étendue en transférant des calculs d'adresses du chemin de données vers des unités spécialisées au sein du séquenceur mémoire afin d'augmenter les performances et diminuer la consommation (le nombre de transferts entre ces deux entités est alors réduit). Les résultats présentés dans ce papier montrent qu'il est possible d'exploiter les informations spécifiques à l'application afin de générer un séquenceur d'accès à la mémoire capable de réduire les surcoûts associés aux accès aléatoires à la mémoire.

2 Travaux Connexes

La majorité des recherches ciblant des applications orientées contrôle intensif sont basées sur des accès directs à des RAM (mémoire à accès aléatoires). Ces applications uti-

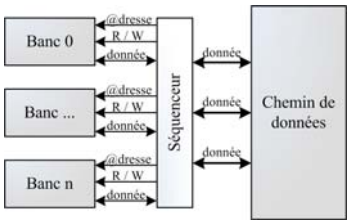


FIG. 1 – Architecture à Base de Séquenceur Mémoire

lisent une architecture mémoire générique qui prend une adresse binaire en entrée et la décode en ligne et colonne afin de réaliser l'accès souhaité. Cette solution est adaptée aux applications pour lesquelles les accès mémoire sont non prédictibles a priori.

Une bonne partie des applications TDSI les séquences d'accès aux éléments d'un vecteur ou d'un tableau sont régulières et périodiques. Dans ces dernières, il est possible de mémoriser/programmer directement au sein du séquenceur les séquences d'adresses des données manipulées par l'application (utilisant des compteurs dédiés [2], des registres à décalage, une machine à états finis ...). Le séquenceur peut ordonnancer de manière optimale l'ordre des accès afin d'exploiter les spécificités de la mémoire réalisant si besoin des accès décalés (lecture préemptives, écriture tardives).

Les circuits basés sur une architecture à séquenceur mémoire permettent au concepteur (ou à l'outil de synthèse) de synthétiser et optimiser de manière decorélée le chemin de données et l'architecture de la mémoire. Comme cela est montré par la figure 1, seules les données produites et consommées par le chemin de données sont transférées d'une unité à l'autre (le transfert des adresses n'étant pas nécessaire dans le cadre de séquences d'accès connues à priori). Ces approches permettent de réduire la consommation de l'architecture finale en réduisant le nombre de transferts inter-unités. Des recherches ont démontré que les séquenceurs d'accès à la mémoire générés à partir des séquences d'accès connues a priori peuvent être optimisés afin d'obtenir des architectures mémoire optimales [3] [1]. Des optimisations visant à augmenter les performances du circuit peuvent être aussi employées. Dans [6], l'auteur présente l'impact sur la surface et les performances (latence) de la simplification de l'architecture du séquenceur et de la mémoire.

L'utilisation d'un séquenceur permet aussi au concepteur de séparer les problèmes d'interfaçage de la mémoire de celui de l'ordonnancement des calculs du chemin de données [4]. Cela permet une meilleure exploitation a priori des techniques d'accès pipeline à la mémoire en tamponnant via des FIFOs les données en cours de transfert.

Dans les approches traditionnelles, le séquenceur d'accès à la mémoire *ne peut être crée que dans le cas d'accès à la mémoire purement déterministes (nommés aussi statiques)*. C'est pourquoi nous proposons une nouvelle architecture de séquenceur permettant la décorrélation de la partie traitement de données et de la partie mémorisation, prenant en compte les accès non déterministes nommés aussi dynamiques(adresses calculées à l'exécution et accès réalisés au sein des branches conditionnelles).

3 Architectures d'Implémentation

Notre objectif est de générer un séquenceur d'accès à la mémoire optimal supportant les adressages dynamiques pour les séquences d'accès *principalement déterministes*. Nous présentons dans un premier temps une architecture autorisant les adressages dynamiques, puis dans un deuxième temps une architecture étendue permettant de réaliser des calculs d'adresses dans le séquenceur mémoire afin de réduire encore le nombre de transferts d'adresses entre les différentes unités. Cela permet de réduire de manière significative le nombre de bus et induit une réduction du coût en surface tout en réduisant le nombre de commutations (réduction de la consommation).

3.1 Séquenceur avec Adressage Dynamique

Dans cette première architecture, nous supposons que tous les calculs d'adresses dynamiques sont réalisés au sein de l'unité de traitement et que les calculs résultats de ces calculs sont ensuite sont ensuite transférés au sein du séquenceur par l'intermédiaire des bus de données.

L'architecture du séquenceur présentée dans la figure 2 est composée de 4 unités différentes : un ordonnanceur d'accès à la mémoire, un contrôleur pour les accès dynamiques, un générateur d'adresses et une table de translation d'adresses.

Les bus reliant le chemin de données sont reliés à la mémoire par un *crossbar* qui les multiplexe. Ce *crossbar* est piloté par l'intermédiaire du séquenceur d'accès à la mémoire, qui connaît a priori les séquences d'accès à la mémoire.

L'ordonnanceur contrôle la progression du générateur d'adresses de manière synchrone par rapport à la progression des calculs dans le chemin de données. Dans le cas d'accès dynamiques à la mémoire, l'adresse utilisera la table de translation d'adresses afin de convertir l'adresse logique en un couple (banc mémoire, adresse physique). Cela a pour effet de permettre au concepteur d'assigner différents éléments d'un même vecteur au sein de différents bancs mémoire et cela de manière non contiguë afin d'exploiter au mieux le parallélisme de l'application.

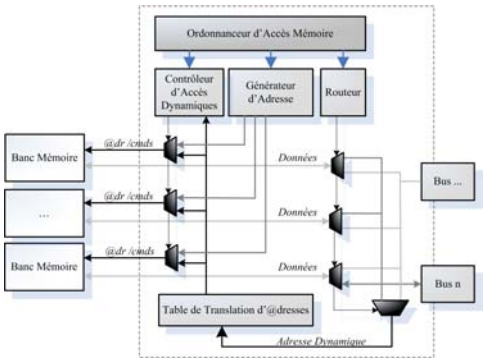


FIG. 2 – Séquenceur pour des Accès Dynamiques

Dans le cas d'un accès dynamique à la mémoire, en fonction du couple (banc mémoire, adresse physique), le contrôleur d'accès dynamique va générer les signaux de commande permettant d'accéder au banc mémoire ciblé. Durant cet accès dynamique, l'ensemble des bancs mémoire potentiellement accessibles sera bloqué.

Cette architecture de séquenceur permet de traiter des accès dynamiques à la mémoire dans une approche de type "séquence d'accès connue à priori" en s'assurant du respect des contraintes de latence/cadence.

3.2 Transfert des Calculs d'Adresse Dynamiques au Sein du Séquenceur

Nous avons précédemment présenté une architecture permettant de réaliser des accès dynamiques au sein de la mémoire à l'aide d'un séquenceur. Nous étendons maintenant notre approche en transférant les calculs d'adresses dynamiques de l'unité de traitement vers le séquenceur mémoire. Afin de permettre cette évolution, nous ajoutons au sein de l'architecture du séquenceur des unités de calculs dédiées. Le chemin de données interne au séquenceur est spécialisé dans les calculs d'adresses dynamiques. Ces unités sont donc dimensionnées de manière optimale par rapport à la largeur (nombre de bits) des adresses. Un autre intérêt de cette méthode est la localisation en mémoire des variables et des constantes exclusivement réservées aux calculs d'adresses. Leur utilisation dans les calculs d'adresses dynamique ne nécessite donc plus de transfert vers le chemin de données.

La nouvelle architecture du séquenceur est présentée en figure 3; un chemin de données dédié, composée d'opérateurs et de registres a été ajouté à coté de la table de translation d'adresses.

Cette évolution permet d'obtenir des gains intéressant en terme de performance dans le cas de circuits pipelines où le transfert des données d'une unité à l'autre peut prendre plusieurs cycles d'horloge (mémoires "lentes", bus pipelinés, etc.).

Localiser les calculs d'adresses dynamiques au sein du séquenceur fournit alors un gain en terme de latence. La réduction des transferts entre l'unité de calculs et le séquenceur entraîne également une réduction de la commutation des fils composant les bus (i.e. réduction de la consommation globale du circuit). De plus la réduction du trafic due à la déportation des calculs d'adresses entraîne une réduction des besoins filaire entre les unités (bus) ainsi que des mémorisations nécessaires pour ces transferts.

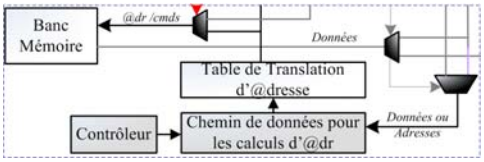


FIG. 3 – Évolution de l'Architecture du Séquenceur

4 Flot de Conception

Afin d'implémenter une application possédant des calculs d'adresses dynamiques sur l'architecture définie précédemment, nous avons élaboré un flot de conception basé sur un modèle de graphe qui permet de prendre en compte les contraintes temporelles dues aux transferts de données entre les unités ainsi que le temps d'accès aux RAM.

Le point d'entré de notre flot de conception (figure 4) est une description algorithmique de l'application accompagnée du mapping mémoire si le concepteur l'a déjà réalisé. Notre méthode peut être utilisée avec ou sans mapping à priori. La première étape consiste à annoter le graphe modélisant l'application avec les contraintes temporelles dues aux transferts de données et aux adressages dynamiques. Une partie des informations utilisées dans cette étape est fournie par le mapping mémoire. Le graphe de type flot de signaux (SDF) est ensuite annoté de manière à prendre en compte le transfert des données et des adresses entre les diverses unités composant le circuit.

Au départ, les calculs d'adresse dynamiques sont supposés être localisés au sein de l'unité de traitement(figure 5a). Ensuite en fonction du mapping mémoire et des informations sur le type des données, des noeuds modélisant les transferts de données et les adressages dynamiques sont insérés dans le graphe. Une fois cette étape réalisée, on obtient le graphe présenté en figure 5b. Ensuite à l'aide d'une métrique de décision, nous déportons les calculs dynamiques d'adresses de l'unité de traitement vers le séquenceur mémoire (figure 5c).

La métrique de décision prend en considération plusieurs critères : le nombre de transferts nécessaires entre les unités, l'allongement ou la réduction des chemins critiques, la réduction de la largeur des opérateurs dans le séquenceur mémoire, etc. La méthode de transfert des calculs d'adresses dynamiques de l'unité de traitement vers le séquenceur est appliquée statiquement sur le graphe. Cela génère un graphe optimal modélisant les calculs d'adresses et leur lieu d'implémentation privilégié.

Le concepteur va ainsi pouvoir synthétiser manuellement (handed coding) ou automatiquement (high-level syn-

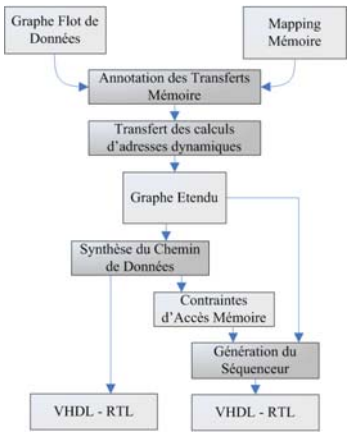


FIG. 4 – Flot de Conception

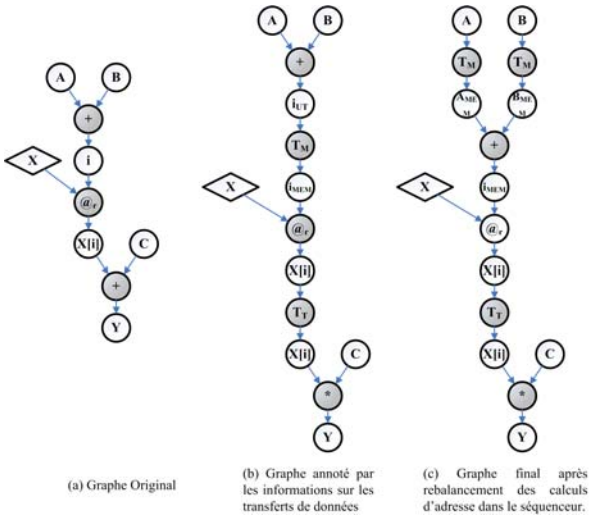


FIG. 5 – Exemple de Graphe Modélisant l'Application et les Transferts (algo : $y = X[i] + c$)

thesis tool) le chemin de données sans se préoccuper de l'implémentation du séquenceur mémoire car les contraintes temporelles sont directement intégrées dans le graphe modélisant l'application.

A l'issue de la synthèse des chemins de données, le concepteur doit fournir les contraintes liées au séquenceur mémoire (date des accès, type des accès, calculs à réaliser, etc.) afin de permettre la génération du séquenceur adéquate.

5 Expériences

Nous avons appliqué notre approche de conception sur une estimation de mouvement de type *Three Step Search* [5], qui possède une séquence d'accès à la mémoire indéterministe. Dans le cadre de l'architecture de référence (contrôle), nous considérons l'ensemble des transferts d'adresses (bloc de référence et macro-blocs dynamiquement sélectionnés). Dans la première solution d'architecture présentée dans cet article, le transfert des 5 premiers macro-blocs est connus de manière statique *à priori* de la même manière que le bloc de référence. Pour la seconde architecture proposée (avec unités de calculs dédiées dans le séquenceur), les seuls transferts qui sont réalisés sont les adresses de base des macro-blocs dynamiquement sélectionnés, les autres calculs étant réalisés au sein du séquenceur.

Les résultats présentés dans la figure 6 montre que les approches à base de séquenceur permettent de réduire le nombre de transferts entre le chemin de données et la mémoire. Il faut tout de même noter que le nombre d'accès à la mémoire est identique mais ces derniers peuvent être optimisés dans le cadre d'une architecture à base de séquenceur. Ces résultats montrent aussi l'intérêt de la méthode dans le cadre d'architecture pipelines où les communications entre les unités prennent plusieurs cycles d'horloge. Cela permet comme le souligne Park [4] d'utiliser au mieux les techniques d'accès à la mémoire afin de réaliser le cas échéant des lectures spéculatives.

	Macro-bloc	*Fenêtre de recherche*	*Transferts d'adresses*
Architecture Contrôle	*8x8*	*16x16*	*1664*
Architecture avec le 1er Séquenceur			*1024*
Architecture avec le 2nd Séquenceur			*16*
Architecture Contrôle	*16x16*	*48x48*	*6656*
Architecture avec le 1er Séquenceur			*4096*
Architecture avec le 2nd Séquenceur			*16*

FIG. 6 – Nombre de transferts nécessaires

6 Conclusion

Dans ce papier nous avons présenté une nouvelle architecture de séquenceur mémoire pour les applications TDSI possédant des séquences d'accès à la mémoire *en partie indéterministes*. Nous montrons de plus que le flot de synthèse proposé permet au concepteur d'optimiser librement chaque unité de son architecture (unité de traitement, unité de mémorisation). Notre travail futur consistera en l'intégration du processus de création du séquenceur dans un Flot de Synthèse de Haut Niveau afin d'automatiser toutes les étapes de création de l'architecture.

Références

[1] T. Kim C. Lyuh and K. Kim. Coupling-aware high-level interconnect synthesis. *IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems*, 23(1) :157–164, January 2004.

[2] P.B Denyer D. Grant and I. Finlay. Synthesis of address generators. In *Proceedings of ICCAD 89*, pages 116–119, 1989.

[3] A. Dasgupta and R. Karri. High-reliability, low-energy microarchitecture synthesis. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 17(12) :1273–1280, 1998.

[4] J. Park and P.C. Diniz. Synthesis of pipelined memory access controllers for streamed data applications on fpga-based computing engines. In *Proceedings of ISSS '01*, pages 221–226. ACM Press, 2001.

[5] B.Zeng R.Li and M.L.Liou. A new three-step search algorithm for block motion estimation994. *IEEE Trans. on Circuits and Systems for Video Technology*, 4(4) :438–442, August 1994.

[6] P. Cheung S. Hettiaratchi and T. Clarke. Performance-area trade-off of address generators for address decoder-decoupled memory. In *Proceedings of DATE 2002*, page 902. IEEE Computer Society, 2002.

Etude Spectrale de l'Échantillonnage par Traversée de Niveaux

F. AESCHLIMANN, E. ALLIER, L. FESQUET, M. RENAUDIN

TIMA, Institut National Polytechnique de Grenoble, 46 avenue Félix Viallet, 38031 Grenoble Cedex, France.

{fabien.aeschlimann, emmanuel.allier, laurent.fesquet, marc.renaudin}@imag.fr

Résumé – La tendance actuelle en microélectronique est d'intégrer des systèmes complexes sur une seule puce (« SoC » pour Sytems on Chip) : capteur, convertisseur analogique-numérique (CAN), traitement numérique du signal, transmission de données. Ces systèmes doivent être peu coûteux, de taille réduite, faible bruit et surtout peu consommateur d'énergie parce qu'il sont souvent alimentés par batteries. Parmi les étages cités, les CAN sont les plus difficiles à concevoir dans le respect du cahier des charges. Dans cette perspective, une nouvelle classe de convertisseur a été développée à partir de l'association de la technologie asynchrone et de l'échantillonnage non uniforme par traversée de niveaux. Le schéma d'échantillonnage est analysé pour les signaux périodiques mettant en évidence un repliement de spectre. Cependant l'étude de la reconstruction montre qu'un signal peut être reconstruit grâce à l'information portée par les intervalles de temps et ce malgré le repliement.

Abstract *– New microelectronic trends is to integrate complex systems on a single chip ("SoC" for System on Chip): sensor, analog-to-digital converter (ADC), digital signal processing, data transmission. These systems must be low-cost, reduced-sized, low-noise and especially low-power because they are often powered by batteries. Among all the existing blocks, ADCs are the most critical to design, respecting all these constraints. In this way, a new class of ADC has been developed combining asynchronous technology and non-uniform level-crossing sampling. The sampling scheme is analysed for periodic signal highlighting alias effects. However, the study of reconstruction showed that a signal can be recovered in spite of aliasing thanks to the information included in time intervals.*

1. Introduction

Notre travail contribue à un changement drastique des chaînes classiques de traitement numérique du signal. L'objectif principal est en effet de réduire leur consommation électrique par un, voire deux ordres de grandeur en repensant complètement leurs architectures et la théorie du signal associée. Dans cette perspective, il a été prouvé que la conception de circuits intégrés en technologie asynchrone [1], c'est-à-dire sans utilisation d'une horloge globale cadençant système, a permis d'améliorer les performances de convertisseurs analogique-numérique (CAN) classiques en les rendant plus sûrs vis-à-vis de la métastabilité et moins sensibles aux interférences électromagnétiques, en augmentant leur vitesse et en diminuant leur consommation électrique [2][3]. De plus, bien que la plupart des systèmes utilisant des CAN traite des signaux dont les propriétés statistiques sont intéressantes, aucune architecture de traitement du signal n'en tient vraiment compte. En réalité, ces signaux (comme la température, la pression, les électrocardiogrammes, la parole…) sont la plupart du temps constants mais peuvent varier fortement pendant une courte période. Ainsi, l'échantillonnage classique à pas constant, que l'on appellera *échantillonnage régulier*, et le système de conversion sont contraints par le théorème de Shannon : assurer que la fréquence d'échantillonnage soit au moins égale au double de la fréquence maximale du signal ; il s'agit là d'un pire cas d'échantillonnage.

De cette manière, notre approche du traitement du signal est de combiner la conception de circuits asynchrones avec des processus commandés par le signal lui-même afin de réduire l'activité dynamique et donc la consommation. Nous souhaitons adapter dynamiquement l'échantillonnage à l'information présente dans le signal. Ainsi une nouvelle classe de CAN appelée CANA (pour *Asynchrone*) a été développée par le groupe CIS du laboratoire TIMA. Cette classe est fondée parallèlement sur une technique d'échantillonnage non uniforme par traversée de niveaux et sur une implémentation matérielle du circuit en technologie asynchrone.

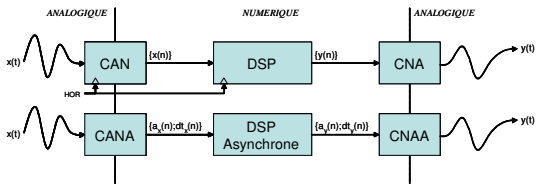


FIG. 1 : Approches parallèles d'une chaîne de traitement numérique du signal.

Par ailleurs, il est connu qu'échantillonner un signal et ce quelle que soit la technique utilisée induit des modifications spectrale pouvant parfois amener une perte d'information telle qu'il est impossible de retrouver les caractéristiques du signal. Dans cet article, nous nous focalisons donc sur le schéma d'échantillonnage par traversée de niveaux afin d'étudier ces modifications. Comme l'échantillonnage dépend du signal, et de ses variations, seuls seront considérés les signaux périodiques (et plus particulièrement les signaux sinusoïdaux) car leurs dérivées possèdent la propriété intéressante d'être elles-mêmes périodiques.

2. Convertission analogique-numérique asynchrone

Nous présentons succinctement dans cette partie la convertission asynchrone ainsi que la technique d'échantillonnage par traversée de niveaux.

En considérant un convertisseur dont la résolution est de M bits, $L = 2^M-1$ niveaux de quantification sont uniformément répartis le long de la dynamique E du signal d'entrée avec un pas de quantification (appelé *quantum*) :

$$q = \frac{E}{2^M - 1} \tag{1}$$

Un échantillon est alors prélevé lorsque le signal à temps continu $x(t)$ traverse l'un des niveaux. Ainsi contrairement à l'échantillonnage régulier, les échantillons ne sont plus uniformément espacés dans le temps parce qu'ils dépendent des variations du signal (cf figure 2). Par conséquent, en plus de la valeur de l'échantillon ax_n, le temps dtx_n écoulé depuis l'échantillon précédent ax_{n-1} doit être mesuré. Un timer de période T_C est donc dédié à cette tâche. L'amplitude de l'échantillon est alors précise mais l'intervalle de temps entre deux échantillons est quantifié en fonction de la précision T_C du timer. Le rapport signal sur bruit (RSB) ne dépend plus du nombre de niveaux de quantification mais uniquement de la valeur de T_C. Ainsi, pour une implémentation donnée du convertisseur, c'est-à-dire pour un nombre de niveaux fixé, le RSB peut être commandé en variant la valeur de la période du timer.

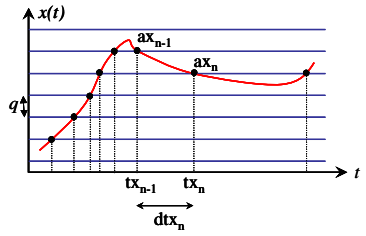


FIG. 2 : Echantillonnage non uniforme par traversée de niveaux.

Bien qu'il ne soit jamais disponible, nous supposons qu'à partir d'un temps initial t_0, l'instant d'échantillonnage du n^{ième} point noté tx_n peut être reconstruit par la relation :

$$tx_n = tx_{n-1} + dtx_n \tag{2}$$

3. Echantillonnage par traversée de niveaux d'un signal périodique

Dans cet article nous considérons un signal périodique $x(t)$ définie par sa période fondamentale $T_0 = 1/f_0$.

3.1 Etude des intervalles de temps

L'échantillonnage par traversée introduit, par rapport à l'échantillonnage régulier classique, une nouvelle donnée : les intervalles de temps séparant les points prélevés. Bien qu'ils ne soient pas explicitement calculables car dépendant à la fois des paramètres du signal et du convertisseur, ils peuvent être approché au 1er ordre par la relation suivante :

$$dtx_n \approx \frac{q}{\left|\frac{dx(t)}{dt}\right|_{t=tx_n}} \tag{3}$$

Les intervalles de temps sont liés à la valeur absolue de la dérivée du signal. Or la dérivée d'un signal périodique possède deux propriétés intéressantes :

- Elle est périodique de période T_0 ;
- Elle est proportionnelle à la fréquence fondamentale du signal f_0.

De cette manière, les intervalles de temps sont également périodiques de période T_0 et proportionnel à f_0.

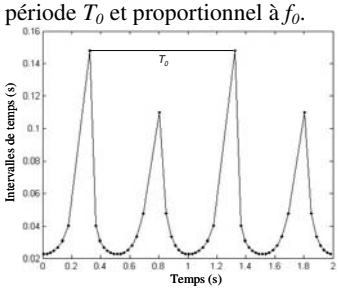


FIG. 3 : Intervalles de temps d'un signal sinusoïdal de fréquence 1Hz. Les intervalles ont une période de 1s.

3.2 Schéma d'échantillonnage

Dans le domaine temporel, il est connu que l'échantillonnage est équivalent à une multiplication entre le signal analogique et un peigne de Dirac, c'est-à-dire un train d'impulsions placées aux instants d'échantillonnage [5]. Dans le cadre de l'échantillonnage par traversée de niveaux, le peigne de Dirac noté $S_F(t)$ dépend du signal :

$$S_F(t) = \sum_{n=-\infty}^{+\infty} \delta(t - tx_n) \tag{4}$$

Ainsi le signal échantillonné s'écrit :

$$x_E(t) = \sum_{n=-\infty}^{+\infty} ax_n \delta(t - tx_n) \tag{5}$$

En revanche dans le domaine fréquentiel, le spectre du signal analogique $X(f)$ est convolué par la transformée de Fourier du peigne de Dirac notée $S_E(f)$ et appelée *Schéma d'échantillonnage*. Comme les intervalles de temps sont périodiques de période T_0, le peigne est également périodique de période T_0. Par conséquent, $S_E(f)$ peut être calculée par série de Fourier : il s'agit alors d'une somme d'harmoniques de fondamental f_0 dont les amplitudes $\hat{h}_n$ dépendent des caractéristiques du signal :

$$S_E(f) = \sum_{n=-\infty}^{+\infty} \hat{h}_n \delta(f - nf_0) \tag{6}$$

Le symbole ^ est utilisé pour représenter la quantification du temps. En effet, comme les coefficients dépendent des

instants d'échantillonnage qui varient en fonction de la résolution du timer, il permet de généraliser la notation. De plus il a été démontré [4] que si tous les intervalles de temps étaient multiples d'une valeur (en l'occurrence la résolution du timer T_C), le schéma d'échantillonnage était périodique de période l'inverse de cette valeur i.e $1/T_C$.

Au final $S_E(f)$ est un spectre d'harmoniques de fréquence fondamentale f_0 et de période $1/T_C$.

Le spectre du signal analogique étant lui-même un spectre d'harmoniques de fréquence fondamentale f_0. Le spectre du signal échantillonné $X_E(f)$ est alors un spectre d'harmoniques repliées de fréquence fondamentale f_0 et de période $1/T_C$:

$$\begin{aligned} X_E(f) &= X(f) * S_E(f) \\ &= \sum_{n=-\infty}^{+\infty} a_n \delta(f - nf_0) * \sum_{m=-\infty}^{+\infty} \hat{h}_n \delta(f - mf_0) \\ &= \sum_{n=-\infty}^{+\infty} \sum_{m=-\infty}^{+\infty} a_n \hat{h}_m \delta(f - (n+m) f_0) \end{aligned} \quad (7)$$

L'équation 7 montre effectivement que l'amplitude de chaque harmonique du spectre du signal échantillonné est une combinaison linéaire de toutes les amplitudes du signal analogique. Le spectre est replié : il y a donc perte d'information.

3.3 Transformée de Fourier discrète Généralisée

Nous proposons de retrouver le résultat décrit par l'équation 7 grâce à la Transformée de Fourier Généralisée [4]. En effet, dérivée de la traditionnelle Transformée de Fourier Discrète, elle permet de calculer le spectre d'un signal échantillonné non uniformément à partir d'une séquence de N points consécutifs :

$$\tilde{X}_E(f) = \frac{1}{N} \sum_{n=0}^{N} ax_n \, e^{-j2\pi f t x_n} \quad (8)$$

A titre d'exemple, un signal sinusoïdal de fréquence 1Hz est échantillonné par un convertisseur 4 bits (15 niveaux) utilisant un timer de résolution 100Hz. Le fenêtrage est effectué par une fonction porte dont la largeur est un multiple entier de la période afin de limiter les effets de fuite d'énergie. Le schéma d'échantillonnage calculé en simulation est conforme à celui obtenu en théorie puisqu'il est de fréquence fondamentale 1Hz et périodique de période 100Hz.

Parallèlement, les simulations montrent que le spectre du signal échantillonné est systématiquement replié. Ainsi d'une part l'erreur introduite sur les harmoniques du signal empêche de retrouver ses caractéristiques et d'autre part les amplitudes élevées des harmoniques parasites induisent une indétermination sur le nombre exacte de composantes présentes dans le signal. Dans notre exemple, l'erreur sur le fondamental est de 4,2dB tandis que la première harmonique parasite est 1,5dB plus grande que le fondamental.

Ceci prouve que la Transformée de Fourier Généralisée ne peut pas être utilisée pour analyser le spectre d'un signal à partir de ses échantillons obtenus par traversée de niveaux. En effet, cette technique attribue implicitement la même valeur informative à chaque point sans tenir compte des intervalles de temps pourtant porteurs d'information (cf équation 7).

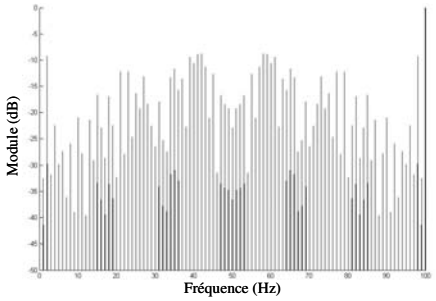


FIG. 4 : Schéma d'échantillonnage d'un signal sinusoïdal de 1Hz échantillonné par un convertisseur 4 bits utilisant un timer à 100Hz

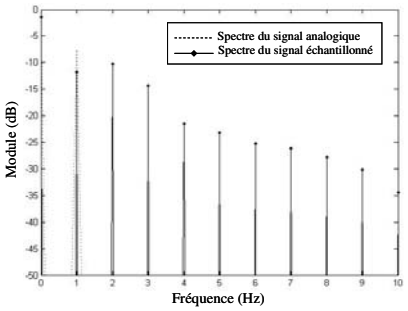


FIG. 5 : Transformée de Fourier Généralisée d'un signal sinusoïdal de 1Hz échantillonné par un convertisseur 4 bits.

4. Analyse spectrale du signal reconstruit

Pour que l'analyse spectrale d'un signal analogique à partir de ses échantillons soit efficace, il faut que l'information portée dans les intervalles de temps soit utilisée. L'idée consiste donc à pondérer, dans la Transformée de Fourier Généralisée, chaque échantillon ax_n par une fonction particulière dont l'unique paramètre est l'intervalle de temps dtx_n. Dans ce but, il suffit de considérer l'opération de reconstruction car le signal étant échantillonné non uniformément, elle doit obligatoirement être adaptée aux intervalles de temps. Dans cet article seul le bloqueur d'ordre 0 est analysé car il représente la technique la plus utilisée en convertission numérique-analogique.

4.1 Reconstruction par bloqueur d'ordre 0

Avec un bloqueur d'ordre 0, le signal reconstruit est obtenu par convolution entre le signal échantillonné $x_E(t)$ et des fonctions rectangulaires dont les largeurs dépendent des intervalles de temps :

$$x_0(t) = x_E(t) * \Pi_\theta(t) \quad (9)$$

Une fonction rectangulaire est définie pour n'importe quel réel θ par :

$$\Pi_{\theta}(t) = \begin{cases} 1 & \text{pour } t \in \left[-\frac{\theta}{2}; \frac{\theta}{2}\right[\\ 0 & \text{ailleurs} \end{cases} \quad (10)$$

En centrant chaque fonction porte entre les points ax_n et ax_{n-1}, le signal reconstruit peut alors s'écrire :

$$x_0(t) = \sum_{n=-\infty}^{+\infty} ax_n \Pi_{dtx_n}\left(t - tx_n + \frac{dtx_n}{2}\right) \quad (11)$$

Puis grâce aux propriétés de linéarité et de translation de la transformée de Fourier, le spectre du signal reconstruit est déduit de l'équation 11 :

$$X_0(f) = \sum_{n=-\infty}^{+\infty} ax_n W_n^0(f) e^{-j2\pi f tx_n} \text{ pour tout } f \in \Re \quad (12)$$

où $W_n^0(f)$ représente la pondération à l'ordre 0 du n^{ième} échantillon :

$$W_n^0(f) = dtx_n e^{-j\pi f dtx_n} sinc(\pi f dtx_n) \quad (13)$$

En simulation, le signal échantillonné est tronqué par une fenêtre rectangulaire de largeur T permettant le traitement d'une séquence de N points consécutifs :

$$\tilde{X}_0(f) = \frac{1}{T} \sum_{n=0}^{N-1} ax_n W_n^0(f) e^{-j2\pi f tx_n} \quad (14)$$

4.2 Résultats et discussion

Le même exemple traitant un signal sinusoïdal est analysé. Les résultats montrent que l'erreur sur l'amplitude du fondamental passe de 4,2dB à 0.2dB tandis que l'harmonique 2 est rejetée à -39,7dB au lieu de +1,5dB avec la Transformée de Fourier Généralisée.

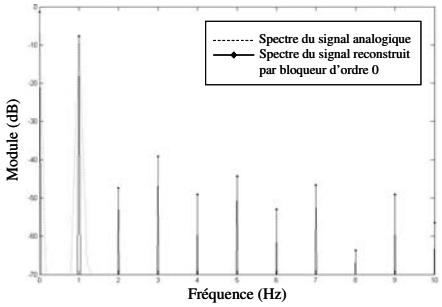


FIG. 6 : Transformée de Fourier d'un signal sinusoïdal de 1Hz échantillonné par un convertisseur 4 bits puis reconstruit.

Le spectre du signal reconstruit permet donc de retrouver toutes les caractéristiques du signal (amplitudes, fréquence) avec une précision acceptable grâce à la pondération. La perte d'information générée lors l'échantillonnage par traversée de niveaux n'est en fait qu'apparente car le processus de reconstruction est non linéaire. En effet, dans le cas classique de l'échantillonnage régulier, la reconstruction est un système linéaire et invariant dans le temps [5]. Il y a donc une réponse impulsionnelle unique (une fonction rectangulaire de largeur la période d'échantillonnage par exemple pour le bloqueur d'ordre 0) et donc une seule réponse en fréquence filtrant le spectre du signal échantillonné. Dans cette situation, le repliement est préjudiciable. En revanche, avec l'échantillonnage par traversée de niveaux, la reconstruction utilise plusieurs réponses impulsionnelles. En fréquence, elle n'est donc plus assimilable à une opération de filtrage ; l'étude préalable du repliement devient alors secondaire. En revanche, la précision de la reconstruction devient une étude prioritaire : en effet, la précision fluctue en fonction des intervalles de temps, et donc des variations du signal. Or les variations dépendent des composantes du signal et de leurs amplitudes. Le choix du quantum détermine donc la précision. Nous pensons sur ce point qu'il existe un critère théorique, reliant le quantum au signal, permettant de borner l'erreur de reconstruction.

Ainsi sous couvert de respecter ce critère, il est possible de concevoir une chaîne de traitement du signal basé sur l'échantillonnage non uniforme par traversée de niveaux pour échantillonner un signal, le traiter, le reconstruire avec une précision garantie.

5. Conclusion

Dans cet article, nous présentons une nouvelle classe de convertisseurs analogique-numérique basée sur un principe d'échantillonnage non uniforme appelé traversée de niveaux. Nous prouvons que le spectre de signaux périodiques échantillonnés est replié. Cependant en étudiant la reconstruction, nous montrons que la perte d'information peut être compensée par l'utilisation des intervalles de temps mesurés lors de l'échantillonnage. Une chaîne complète est en cours de réalisation pour montrer que notre approche conduit à une réduction significative de la consommation électrique, rendant cette technologie très attractive pour le domaine de la conception des systèmes sur puces.

Références

[1] M. Renaudin, "Asynchronous Circuits and Systems: a Promising Design Alternative", *Journal of Microelectronic Engineering*, Vol. 54, pp. 133-149, 2000.

[2] E. Allier, L. Fesquet, G. Sicard, M. Renaudin, "Low Power Asynchronous A/D Conversion", *Proceedings of the 12th International Workshop on Power and Timing, Modeling, Optimization and Simulation (PATMOS'02)*, September 11-13 2002, Sevilla, Spain.

[3] E. Allier, G. Sicard, L. Fesquet, M. Renaudin, "A New Class of Asynchronous A/D Converters Based on Time Quantization", *ASYNC Proceedings*, pp. 197-205, May 12-16 2003, Vancouver, Canada.

[4] P.C. Bagshaw, M. Sarhadi, "Analysis of samples of wideband signals taken at irregular sub-Nyquist intervals", *IEE Electronic Letters*, Vol. 27, n°14, pp. 1228-1230, July 1991.

[5] A.V. Oppenheim, A.S. Willsky, I.T. Young, "Signals and Systems", *Prentice-Hall Signal Processing Series*, 1995.

Séquences de canalisation chaotiques pour les systèmes CDMA Multicodes

Mourad Khanfouci, Sylvie Marcos
Laboratoire des signaux et systèmes
SUPELEC plateau de moulon
91192, Gif sur Yvette CEDEX (FRANCE)
Email : khanfouci, marcos@lss.supelec.fr

***Résumé*- Nous proposons dans cette communication une méthode simple et systématique pour la construction/ le choix de séquences d'étalement pour un système d'accès multiple par répartition des codes, basée sur la technique d'étalement du spectre par séquence directe (DS-CDMA) et employant plusieurs séquences par utilisateur. La technique de construction des séquences d'étalement proposée minimise la statistique d'ordre 2 de l'interférence et les séquences d'étalement optimales peuvent être très simplement implémentées par les séquences chaotiques $(10, 2)-$Markoviennes [5], [7], [8], [20].**

I. Introduction

Les systèmes accès multiple par répartition des codes, employant plusieurs séquences d'étalement par utilisateur, i.e. systèmes CDMA multicodes, ont suscité l' intérêt d'une communauté de plus en plus importante de chercheurs en traitement du signal et en communications numériques [4], [11], [14]–[16], [19].

En effet, les systèmes CDMA multicodes permettent d'augmenter le débit de transmission des utilisateurs du système en gardant les avantages des systèmes CDMA conventionnels, i.e. résistance à une interférence bande-étroite, faible probabilité d'interception, résistance aux multi- trajets, etc.

Néanmoins, l'emploi de plusieurs séquences d'étalement par utilisateur, réduit de façon dramatique les ressources disponibles dans le domaine des codes [14], ce qui induit une interférence inter- utilisateurs nettement plus importante que pour les systèmes CDMA conventionnels et détériore les performances du système, évaluées en termes de taux d'erreurs binaires (TEB) [18].

Ce problème a donné lieu à tout un ensemble de travaux sur *l'optimisation* et le partage des ressources du système dans le domaine des codes [2], [3], [10], [14], [16], [21].

Dans [2], [3], [14], [21] les auteurs ont proposé, des procédures d'allocation dynamique des séquences d'étalement, appelées encore codes de canalisation. Cette procédure se base sur le *partage* d'une ou d'un groupe de séquences d'étalement pour un groupe d'utilisateurs actifs dans le système. Le contrôle du débit des utilisateurs peut se faire en choisissant, en plus du nombre des séquences d'étalement, une constellation de modulation différente [14]. Les séquences d'étalement considérées sont des séquences orthogonales, obtenues à partir des séquences de Hadamard [16], ou encore des *séquences composites* obtenues par un mélange de séquences de Hadamard avec des séquences pseudo- aléatoires conventionnelles [13].

Dans cette contribution, on propose une autre solution au problème de la saturation des performances du système CDMA multicode. Cette solution se base sur une construction de séquences de canalisation, minimisant les statistiques d'ordre 2 de l'interférence inter- utilisateurs du système. Les séquences optimales ainsi définies peuvent être alors implémentées en utilisant les séquences chaotiques $(10, 2)-$Markoviennes [5], [7], [8], [20].

Dans la section II nous allons détailler un modèle analytique du système CDMA multicodes, employant des séquences d'étalement BPSK très générales, considéré en lien montant et en situation de contrôle parfait des puissances. On décrira les interférences associées au système multicode ainsi que les performances du système en approximation Gaussienne ainsi que le profil de corrélation apériodique optimal associé [12].

Nous décrirons dans la section III le formalisme associé aux séquences d'étalement chaotiques Markoviennes en montrant que l'on peut générer les séquences de canalisation optimales du système en modifiant correctement le générateur chaotique associé [20].

Dans la section IV on présentera une validation des séquences proposées en considérant différents scénarii de simulation, en montrant que les performances des séquences proposées sont nettement supérieures à celles des séquences de canalisation orthogonales et mélangées.

On finira par une conclusion et quelques remarques dans la section V.

II. Modèle du système et formulation du problème

Considérons un modèle très général du système CDMA multicodes, en lien montant et en situation de contrôle parfait des puissances où chaque utilisateur emploie L_k séquences d'étalement, prises dans un ensemble de séquences $\mathcal{A}_k = \{a_{k,0}(t), \cdots, a_{k,L_k-1}(t)\}$. Le signal mélange $r(t)$ à la

sortie du système s'exprime comme

$$r(t) = \sum_{k=1}^{K} \sum_{m=-\infty}^{+\infty} \sum_{l=0}^{L_k-1} b_k\left(mL_k + l\right) a_{k,l}(t - mT_s - \tau_k) + n(t) \tag{1}$$

où $\{b_k(n)\}_{n=0}^{L_k-1}$ est un bloc de L_k symboles de l'utilisateur k, transmis sur les L_k canaux parallèles associés aux séquences $\mathcal{A}_k$, T_s est la période symbole, τ_k est le retard d'accès de l'utilisateur k, K est le nombre des utilisateurs actifs dans le système et $n(t)$ est un bruit additif Gaussien et réel de densité spectrale de puissance $\sigma_n^2 = N_0/2$. La statistique de décision associée à la voie p d'un récepteur conventionnel [11] associé à l'ensemble des séquences $\mathcal{A}_1$ pour l'utilisateur 1 et le symbole $m = 0$, peut se mettre sous la forme

$$\begin{aligned} z_{1,p}(0) = b_1(p)T_s + \sum_{l=0, l\neq p}^{L_1-1} b_1(l)\rho_{p,l}^{(1)} \\ + \sum_{k=2}^{K} \int_0^{T_s} S^{(k)}(t-\tau_k) a_{1,p}(t)dt + \eta_{1,p} \end{aligned} \tag{2}$$

où $S^{(k)}(t) = \sum_{m=-\infty}^{+\infty} \sum_{l=0}^{L_k-1} b_k\left(mL_k + l\right) a_{k,l}(t - mT_s)$, est le signal transmis par l'utilisateur k, $\rho_{p,l}^{(1)} = \int_0^{T_s} a_{1,l}(t) a_{1,p}(t)dt$ est le coefficient de corrélation entre les séquences $a_{1,l}(t)$ et $a_{1,p}(t)$ de l'utilisateur 1 et $\eta_{1,p}$ est un bruit filtré par la séquence $a_{1,p}(t)$. L'expression (2) montre la présence de deux interférences limitant les performances du système CDMA multicodes. Une interférence *inter- voies*, appelée encore interférence intra-utilisateur et déterminée par $I_{1,p} = \sum_{l=0, l\neq p}^{L_1-1} b_1(l)\rho_{p,l}^{(1)}$ et une interférence inter- utilisateurs, déterminée par $M_{1,p} = \sum_{k=2}^{K} \int_0^{T_s} S^{(k)}(t-\tau_k) a_{1,p}(t)dt$. L'approximation Gaussienne [11] de l'interférence $\mathcal{I}_{1,p}$, donnée par $\mathcal{I}_{1,p} = I_{1,p} + M_{1,p}$, permet d'évaluer très simplement la probabilité d'erreur du système par

$$P_e \approx Q\left(\sqrt{\frac{1}{\sigma_{\mathcal{I}_{1,p}}^2 + N\dfrac{N_0}{2}}}\right) \tag{3}$$

où $Q(x) = \frac{1}{\sqrt{(2\pi)}} \int_x^{\infty} \exp\frac{-y^2}{2} dy$, N est le facteur d'étalement et $\sigma_{\mathcal{I}_{1,p}}^2$ est la statistique d'ordre 2 de l'interférence totale du système $\mathcal{I}_{1,p}$. L'optimisation des performances du système CDMA multicode asynchrone est équivalente dans le contexte de l'approximation Gaussienne au choix/construction des séquences d'étalement qui minimisent $\sigma_{\mathcal{I}_{1,p}}^2$, la statistique d'ordre 2 de l'interférence $\mathcal{I}_{1,p}$. Le moment d'ordre 2 de l'interférence peut être réécrit comme

$$\sigma_{\mathcal{I}_{1,p}}^2 = \frac{(L_1 - 1)}{N^2} + \frac{T_s^2}{3N^3} \sum_{k=2}^{K} \sum_{l=0}^{L_k-1} \left\{2\mu_{k,p,1,l}(0) + \mu_{k,p,1,l}(1)\right\} \tag{4}$$

où $\mu_{k,p,1,l}(n) = \sum_{\nu=1-N}^{N-1} C_{k,p,1,l}(\nu) C_{k,p,1,l}(\nu + n)$ et $C_{k,p,1,l}(\nu)$ est la fonction de corrélation apériodique de Pursley [12], donnée par

$$C_{k,p,1,l}(\nu) = \begin{cases} \sum_{m=0}^{N-n-1} a_{p,m}^{(k)} a_{l,m+\nu}^{(1)} & \text{si } 0 \leq n \leq N-1 \\ \sum_{m=0}^{N+n-1} a_{p,m}^{(k)} a_{l,m-\nu}^{(1)} & \text{si } 1-N \leq n < 0 \\ 0 & \text{ailleurs} \end{cases} \tag{5}$$

En considérant l'hypothèse asymptotique des facteurs d'étalement très grands, i.e. $N \gg$. $\sigma_{\mathcal{I}_{1,p}}^2$ est déterminé par l'interférence inter- utilisateurs $M_{1,p}$. Le profil de corrélation apériodique $C_{\text{opt}}(\nu)$ minimisant $\sigma_{\mathcal{I}_{1,p}}^2$ est identique au profil optimal pour les systèmes CDMA conventionnels. Il est donné par

$$C_{\text{opt}}(\nu) = \left(N - |\nu|\right)(-r)^{\nu} \tag{6}$$

avec $r = 2 - \sqrt{3}$. Nous proposons d'implémenter les séquences optimales ainsi construites par le biais de séquences chaotiques $(10, 2)-$Markoviennes. Ces séquences ont été récemment proposées pour l'optimisation des systèmes CDMA, munis d'une détection multi- utilisateurs linéaires [20].

III. Séquences chaotiques Markoviennes

Les séquences chaotiques Markoviennes peuvent être décrites par les équations de transitions d'état et d'observation suivantes [1].

$$\begin{aligned} x_{n+1} &= f(x_n) + \epsilon\, w_k \\ \tilde{a}_n &= \tilde{Q}(x_n) \end{aligned} \tag{7}$$

avec $n \in \{0, \cdots, N-1\}$, x_0 est la condition initiale, appelée encore état initial du système , $f(.)$ est le générateur chaotique du système, et $\tilde{Q}(.)$ est une fonction de quantification, w_n est un bruit d'état, associé à l'évolution du système $\{x_k\}_{k=0}^{N-1}$ et ϵ une constante réelle tel que $\epsilon \ll 1$. Dans ce travail, on s'intéressera à des générateurs chaotiques très simples, où $f(x)$ est une fonction linéaire par parties, affine et Markovienne [5], [8], [20]. On définira le générateur $f(x)$ sur une partition $\mathfrak{C}\left(N_p\right)$ de l'intervalle réel $I = [0, 1]$ et de cardinalité N_p. La distribution du bruit d'état $f_{w_n}(w_n)$ est la distribution invariante $\bar{f}(x)$ du système [1], [9].Dans notre cas, cette distribution invariante est la distribution uniforme $\mathcal{U}_I(x)$ [1].

La séquence binaire, $\{\tilde{a}_n\}_{n=0}^{N-1}$, obtenue par la quantification du système dynamique (7) est utilisée pour construire la séquence d'étalement BPSK $a_k(t)$ de la manière suivante.

On génère, tout d'abord la condition initiale, $x_0^{(l)}$ de manière i.i.d. à partir de la distribution invariante du système $\bar{f}(x) = \mathcal{U}_I(x)$, ensuite, la séquence $\{\tilde{a}_n\}_{n=0}^{N-1}$ est construite à partir de la condition initiale $x_0^{(l)}$ et les équations de transition et d'observation (7).La séquence d'étalement $a_k(t)$ est alors construite par modulation BPSK de la séquence $\{\tilde{a}_n\}_{n=0}^{N-1}$. [1].

Les propriétés de corrélation apériodique des séquences Markoviennes peuvent être très simplement évaluées en

[1] Le n-ième chip de la séquence $a_k(t)$ est donné par $a_{k,n} = (-1)^{\tilde{a}_n}$

utilisant la matrice de transition $\mathcal{K}$ du système chaotique [20].

Les entrées $\mathcal{K}_{i,j}, i, j = 0, \cdots, N_p - 1$ représentent la proportion de l'élément $I_i \in \mathfrak{C}$, transformée en I_j par les compositions successives du générateur $f(.)$. On peut alors écrire les éléments de la matrice $\mathcal{K}$ comme :

$$\mathcal{K}_{i,j} = \frac{\mu\left(I_i \bigcap f^{-1}\left(I_j\right)\right)}{\mu\left(I_i\right)} \quad i, j \in \{0, \cdots N_p - 1\} \qquad (8)$$

où $\mu(.)$ est la mesure standard de Lebesgue associée à la distribution invariante du système [2].

La fonction d'auto-corrélation apériodique moyenne des séquences peut être exprimée par la forme quadratique suivante

$$C(n) = \frac{1}{N_p} \sum_{i=1}^{N_p} \sum_{j=1}^{N_p} \tilde{Q}\left(I_i\right) \mathcal{K}_{i,j}^{(n)} \tilde{Q}\left(I_j\right) \qquad (9)$$

où la matrice $\mathcal{K}^{(n)}$ est la $n-$ième puissance de la matrice de transition $\mathcal{K}$. Par convention $\mathcal{K}^{(0)}$ est la matrice identité $N_p \times N_p$.

En rajoutant une queue de t_p décalages au générateur N_p-Markovien, il est possible de construire très simplement une famille de séquences d'étalement chaotiques, ayant des profils d'auto-corrélation apériodiques à décroissance exponentielle.

Un exemple typique de ces séquences $(N_p, t_p)-$Markoviennes est donné par le générateur $f(x)$ de la figure (1). En utilisant (8) et (9)

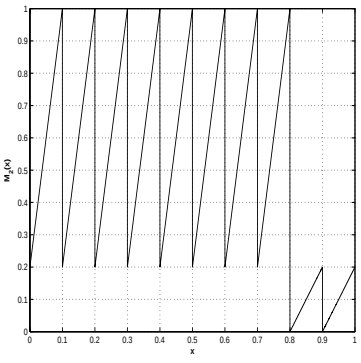


Fig. 1. Générateur $(10, 2)-$Markovien

il est possible, là encore d'évaluer la fonction d'auto- corrélation apériodique moyenne des séquences$(N_p, t_p)-$Markoviennes. On peut montrer que cette fonction s'exprime comme

$$C(n) = \begin{cases} N & \text{si } n = 0 \\ (N - |n|)h^n & \text{sinon} \end{cases} \qquad (10)$$

Le paramètre h est donné en fonction des paramètres du générateur comme $h = \frac{-t_p}{N_p - t_p}$. Il est donc possible de générer une variété de séquences d'étalement BPSK ayant des profils d'auto-corrélation à décroissance exponentielle par un réglage des paramètres N_p et t_p du générateur chaotique $f(.)$.

Pour le choix particulier de $N_p = 10$ et $t_p = 2$, on obtient une famille de séquences avec $h = -0.25$ ce qui correspond au profil d'auto- corrélation apériodique optimal $r = 2 - \sqrt{3}$, décrit dans la section II. Nous proposons donc d'utiliser les séquences $(10, 2)-$Markoviennes comme des séquences de canalisation dans le système CDMA multicodes décrit précédemment. Nous allons évaluer dans la section (IV) les performances des séquences proposées.

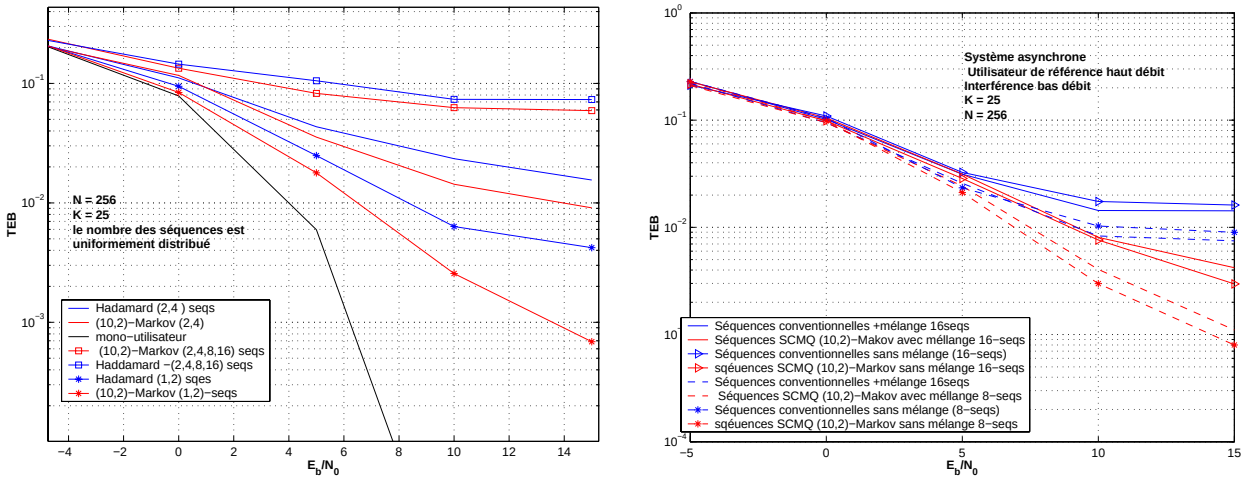
IV. Résultats de simulation

Dans la figure (2), nous avons comparé les performances des séquences de canalisation $(10, 2)-$Markoviennes et celles de schémas de canalisation classiques, basés sur l'emploi de séquences binaires orthogonales, i.e. séquences de Hadamard [6], en considérant des séquences de mélange de Gold de période $N_m = 511$. Le facteur d'étalement est fixé à $N = 256$. On considère des systèmes CDMA multicodes asynchrones en situation de contrôle parfait des puissances avec $K = 25$ utilisateurs. Afin d'illustrer les performances des séquences proposées dans des situations de débit non-homogène des utilisateurs, nous avons considéré que le nombre des séquences par utilisateur L_k est défini comme $L_k \in \{1, 2, \cdots, L_{\max}^{(k)}\}$. Dans la figure (2(b)), nous avons augmenté le débit de l'utilisateur de référence, i.e. l'utilisateur 1, en gardant la même configuration de l'interférence inter- utilisateurs que dans les systèmes CDMA conventionnels, i.e. $\{L_{\max}^{(k)} = 1\}_{k=2}^{K}$. Nous appellerons ce cas " interférence bas- débit". Dans la figure (2(a)), nous avons fixé le débit de l'utilisateur de référence à $R_1 = 1/N$, i.e. $L_{\max}^{(1)} = 1$, le nombre de séquences L_k des utilisateurs contribuant à l'interférence est alors uniformément choisi dans $\{1, 2, \cdots, L_{\max}^{(k)}\}$, avec $L_{\max}^{(k)} = 2, 4, 8, 16$ séquences respectivement. Dans les deux résultats proposés, on constate la supériorité des performances des séquences $(10, 2)-$Markoviennes par rapport aux séquences de Hadamard et séquences de Hadamard mélangées.

V. Conclusion

Nous avons proposé dans cette communication une nouvelle optimisation des séquences d'étalement pour le système CDMA multicodes et asynchrone, considéré en situation de contrôle parfait des puissances et employant des séquences BPSK ainsi qu'une mise en forme rectangulaire des chips. L'optimisation proposée dans ce travail consiste à construire des séquences d'étalement à partir de séquences chaotiques Markoviennes pour minimiser les statistiques d'ordre 2 de l'interférence à accès multiple du système. Nous avons évalué les performances des séquences d'étalement ainsi construites, et comparé leurs performances par rapport aux séquences classiques de Hadamard, mélangées par des séquences PN linéaires [17]. Nos résultats montrent que l'on peut améliorer d'au moins

[2]Pour tout élément I_i de la partition $\mathfrak{C}$, on a $\mu(I_i) = \int_{I_i} \mathcal{U}_I(x)dx$



(a) Interférence bas- débit

(b) Interférence haut- débit

Fig. 2. Performances des séquences $(10, 2)-$Markoviennes

2 dB les performances, en termes de TEB du système par l'emploi des séquences $(10, 2)-$Markoviennes.

References

[1] M.C.MacKey A.Lasota. *Chaos fractals and noise*. Springer-Verlag, 1994.

[2] R.D.Gitlin C.Lin. Multi-Code CDMA Wireless Personal communications Networks. *Proceedings of the ICC'95*, pages 1060 –1064, 1995.

[3] R.D.Gitlin C.Lin, G.P.Pollini. Performance of Multi-Code CDMA. *Proceedings of the IEEE Golbal communication conference, GLOBECOME'95*, pages 907–911, 1995.

[4] V.K.Bhargava D.I.Kim. Performance of Multidimensional Multicode DS-CDMA using code diversity and error detection. *IEEE transactions on communications*, 49(5) :875–887, May 2001.

[5] R. Rovatti G. Mazzini. Interference minimisation based on correlation shaping in asynchronous DS-CDMA. *Electronics Letters*, 35(13) :1054–1055, 1998.

[6] G.E.Bottomley. Signauture sequence selection in a CDMA system with orthogonal coding. *IEEE Transctions on Vehicular technology*, 42(1) :62–68, February 1993.

[7] R.Rovatti G.Mazzini, G.Setti. Chaotic complex spreading sequences for asynchronous DS-CDMA part : II, some theoretical performance bounds. *IEEE transactions on circuits and systems*, 45(4) :496–506, April 1998.

[8] G.Mazzini,G.Setti, R.Rovatti. Chaotic complex spreading sequences for asynchronous DS-CDMA Part : I, System modeling and results. *IEEE transactions on circuits and systems*, 44(10) :937–947, October 1997.

[9] G.Setti,G.Mazzini,R.Rovatti,S.Callegari. Statistical modeling of discrete-time chaotic process-Basic finite dimensional tools and applications. *Proceedings of the IEEE*, 90(5) :662–690, May 2002.

[10] J.Zou ,V.K.Bhagrava. Design issues in a CDMA cellular system with heterogenous traffic types. *IEEE Transactions on vehicular technology*, 47(3) :871–883, August 1998.

[11] K.B.Letaief and al. Multicode high speed transmission of wireless mobile communications. *IEEE Global communication conference, GLOBECOM'95*, 3 :1835–1839, november 1995.

[12] M.B.Pursley. Performance evaluation of phase coded SSMA part I : System analysis. *IEEE transactions on communications*, 25(8) :795–799, August 1977.

[13] Q.Wang Mo-Han.Fan, V.K.Bhargava. Concatenated orthogonal/PN spreading sequences and their application to cellular DS-CDMA Systems with integrated traffic. *IEEE Journal on selected areas in communications*, 14(3) :547–558, April 1996.

[14] L.B.Milstein N.Guo. On rate-variable multidimensional DS/SSMA with sequence sharing. *IEEE Journal on setelected areas in communications*, 17(5) :902–916, May 1999.

[15] F.Gourge P.Agin. Comparison between Multicode with fixed spreading and single with variable spreading operation in UTRA/TDD. *IEEE Workshope in signal processing advances in Wireless communications*, pages 325–328, 1999.

[16] D.V.Sarwate P.K.Enge. Spread-Spectrum Multiple-Access of orthogonal Codes : Linear Receivers. *IEEE Transctions on communications*, COM-35(12) :1309–1319, December 1987.

[17] R.Padovani. Reverse link performance of IS-95 based cellular systems. *IEEE Transactions on personal Communications*, pages 28–34, 1994.

[18] R.Zhang and al. BER performance comparison of single code and multicode DS/CDMA channelization schemes for high rate data transmission. *IEEE Communication letters*, 5(2) :67–69, February 2001.

[19] H.W.Lee et D.K.sung S.J.Lee. Capacities of single code and multicode DS-CDMA system accomodating multiclasse services. *IEEE Transactions on vehicular technology*, 48 :376–384, Mars 1999.

[20] M.Khanfouci S.Marcos. Performances des séquences Markoviennes linéaires par parties (PLM) dans un système AMRC asynchrone. *GRETSI'03*, 2 :43–48, September 2003.

[21] H.Kikuchi H.Watanabe J.Zhu S.Sasaki. Performances evaluation of parallel combinatory SSMA in Rayleigh fading channels. *Proceedings of ISSTA'94*, pages 198–202, 1994.

Contrôle de puissance et détection multicode conjointe dans une liaison descendante à haut débit de l'UMTS

Bessem Sayadi, Stefan Ataman, Inbar Fijalkow

Equipes Traitement des Images et du Signal, ETIS-CNRS-UCP
6 avenue du Ponceau, 95014, Cergy Pontoise, France
{sayadi, ataman, fijalkow}@ensea.fr

Résumé – Dans ce papier, nous nous intéressons à l'optimisation conjointe des paramètres des récepteurs linéaires joints multicodes, proposés récemment dans [4], et du contrôle de puissance, afin d'améliorer la qualité du lien descendant haut-débit de l'UMTS. L'approche proposée nous permet de minimiser la puissance transmise par la station de base et d'obtenir la répartition optimale des puissances des codes au niveau de la station de base.

Abstract – In this paper, the benefits of combining a downlink power control and a joint multicode detection, proposed recently in [4], is analyzed. It is demonstrated that by applying the proposed scheme, we have, first, the optimal code power allocation, and second, the minimization of the base station (BS) transmitted power. The integration of the power control and multicode receiver optimization is done by adapting the filter coefficients to suppress the self interference caused by the multicode transmission.

1 Introduction

La convergence des technologies de l'Internet et des réseaux mobiles qui se fera dans le cadre de la troisième génération des réseaux mobiles et au-delà et l'émergence des nouvelles applications temps-réel multimedias, telles que la vidéo et la visioconférence, nécessite la prise en compte de services à débit variables avec des contraintes de qualité de service diversifiées. Pour cela, deux schémas d'étalement ont été proposés. Le premier se base sur l'emploi des séquences d'étalement à longueur variable. Cependant, l'emploi des codes courts augmente le niveau de l'Interférence Inter-Symboles (IIS) en présence d'un canal à multi-trajets. Afin de remédier à ce problème, une transmission en multicodes a été proposée. Le flot de données haut débit est partitionné en plusieurs sous-flots à plus faible débit étalés par des séquences d'étalement courtes. Tous les sous-flots sont ensuite transmis d'une façon synchrone. Cette technique est déjà adoptée dans le nouveau lien UMTS : le lien HSDPA [1] qui peut atteindre un débit de l'ordre de 14 Mbits/s. Cependant, la présence des multi-trajets dans les canaux de propagation détruit l'orthogonalité des codes et crée l'Interférence Inter-Code (IIC) qui augmente avec le nombre des codes parallèles employés dans le schéma de transmission en multicodes [3].
Afin d'améliorer la qualité du lien descendant définie en terme de RSI (Rapport Signal à Interférence), une réception jointe multicodes récemment a été proposée, dans [4], sous l'hypothèse que les différents codes sont transmis avec la même puissance. En se basant sur la description du signal reçu en présence d'un canal sélectif en fréquence où plusieurs débits de transmission sont considérés, nous avons montré dans [4] que le problème de décodage de l'utilisateur haut débit multicodes peut être structuré comme un problème de détection multi-utilisateurs [2].

Indépendamment, le contrôle de puissance suppose une structure fixe du récepteur et optimise la communication entre la station de base et les mobiles en contrôlant la puissance transmise par les différents utilisateurs. Cependant, un nombre faible de travaux s'est intéressé au contrôle de puissance dans un lien descendant [6] ou sur la manière dont la station de base doit répartir la puissance sur les différents codes (du schéma multicodes) afin de minimiser l'IIC et la puissance transmise par la station de base.
Dans ce papier, nous proposons d'optimiser conjointement, au niveau du mobile, les paramètres du récepteur joint multicodes et du contrôle de puissance. Cette approche nous permet d'améliorer les performances du lien en augmentant sa capacité et de réduire l'IIC. La station de base adaptera ses puissances suivant les conditions de la propagation. Afin de mettre en oeuvre l'algorithme, les paramètres du récepteur joint multicodes et le rapport RSI à la sortie du récepteur sont calculés. La puissance de chaque code est alors mise à jour d'une façon récursive à partir des RSI. Le mobile transmet alors à la station de base la nouvelle répartition des puissances sur les codes par le biais d'une voie de retour. Cette voie est prévue, par exemple dans le lien HSDPA. L'algorithme converge si on atteint un RSI *cible* sur chaque code. Nous rappelons que notre mesure de la qualité de service est le RSI *cible* qui est fixé par les couches hautes du système. L'implémentation de cette approche nécessite la mesure du rapport RSI au niveau de chaque code. Dans cette étude, nous analysons les détecteurs MMSE et ZF (Zero Forcing) en comparant leurs performances à celles obtenues par un banc de récepteurs RAKE.

2 Modèle algébrique du lien UMTS haut débit

Supposons la présence de K utilisateurs dans une cellule transmettant chacun à un débit différent. Chaque utilisateur reçoit une trame de N_{chip} chips. La station de base attribue, en se basant sur la qualité de service requise par l'utilisateur k, un facteur d'étalement noté G_k et M_k codes d'étalement sous la condition que $N_{chip} = G_k N_{bit}^{(k)}$ où $N_{bit}^{(k)}$ représente le nombre des symboles transmis par chaque utilisateur k. Sous la contrainte que le débit chip est constant, $1/T_c$, la période symbole, notée par $T_{s,k} = G_k T_c$, varie en fonction de débit requis par l'utilisateur k.
L'atténuation entre la BS (Base-Station) et le $k^{\text{ème}}$ utilisateur est modélisée en présence d'un effet de *shadowing* par [5] :

$$z_k = d_k^{-\beta} 10^{\frac{\xi_k}{10}}, \quad \xi_k \sim \mathcal{N}(0, \sigma_\xi^2) \tag{1}$$

où d_k représente la distance entre l'utilisateur et la BS. β représente le coefficient d'atténuation. $\mathcal{N}(0, \sigma_\xi^2)$ représente une densité Gaussienne de moyenne 0 et de variance σ_ξ^2.
Nous supposons la présence d'un canal multi-trajets caractérisé par L trajets, noté $\mathbf{h} = [\alpha_0, \alpha_1, \ldots, \alpha_{L-1}]^T$. Ainsi le canal peut être modélisé comme suit :

$$\mathbf{h}_k = z_k \mathbf{h} \tag{2}$$

La puissance attribuée au $m^{\text{ème}}$ code de l'utilisateur k est notée par $p_m^{(k)}$. Le signal transmis par l'utilisateur $k^{\text{ème}}$ s'écrit,

$$y_k(t) = \sum_{n=0}^{N_{bit,k}-1} \sum_{m=1}^{M_k} \sqrt{p_m^{(k)}} b_m^{(k)}(n) c_m^{(k)}(t - nT_{s,k}), \tag{3}$$

$$\text{où } c_m^{(k)}(t) = \sum_{q=0}^{G_k-1} c_m^{(k),(q)} \psi(t - qT_c)$$

$b_m^{(k)}(n)$ représente le symbole de l'utilisateur $k$véhiculé sur $m^{\text{ème}}$ code transmis à l'instant n . ψ est la fenêtre de mise en forme de durée T_c. Le signal en bande de base s'écrit, ainsi,

$$\begin{aligned} r(t) &= \sum_{k=1}^{K} z_k \sum_{l=0}^{L-1} \alpha_l \sum_{n=0}^{N_{bit,k}-1} \\ &\quad \sum_{m=1}^{M_k} \sqrt{p_m^{(k)}} b_m^{(k)}(n) c_m^{(k)}(t - nT_{s,k} - \tau_l) + n(t) \end{aligned} \tag{4}$$

où $n(t)$ est un bruit blanc gaussien.
Le signal reçu discrétisé au rythme $1/T_c$, s'exprime comme suit :

$$r_l = r(lT_c) = \sum_{k=1}^{K} z_k \sum_{l=0}^{L-1} \alpha_l$$

$$\sum_{n=0}^{N_{bit,k}-1} \sum_{m=1}^{M_k} \sqrt{p_m^{(k)}} b_m^{(k)}(n) c_m^{(k)}((l - nG_k - t_l)\, T_c) + n(lT_c) \tag{5}$$

où t_l désigne le retard de trajet discrétisé.
La transmission de la séquence peut s'exprimer vectoriellement par $\mathbf{r} = \left[r_1, ..., r_{N_{chip}+L-1}\right]^T$:

$$\begin{aligned} \mathbf{r} &= \sum_{k=1}^{K} \mathbf{C}^{(k)} \widetilde{\mathbf{H}}^{(k)} \mathbf{W}^{(k)} \mathbf{b}^{(k)} + \mathbf{n} \qquad (6) \\ \mathbf{b}^{(k)} &= \left[b_1^{(k)}(0), ..., b_{M_k}^{(k)}(0), ..., b_{M_k}^{(k)}(N_{bit}^{(k)} - 1)\right]^T \end{aligned}$$

où $\widetilde{\mathbf{H}}^{(k)} = diag(\mathbf{h}_k, ..., \mathbf{h}_k)$ de dimension $(N_{bit}^{(k)} M_k L, N_{bit}^{(k)} M_k)$ et $\mathbf{W}^{(k)} = diag(\mathbf{P}^{(k)}, \mathbf{P}^{(k)}, \ldots \mathbf{P}^{(k)})$ de dimension $N_{bit}^{(k)} M_k$ où $\mathbf{P}^{(k)} = diag(\sqrt{p_1^{(k)}}, \sqrt{p_2^{(k)}}, \ldots \sqrt{p_{M_k}^{(k)}})$ et $diag(\mathbf{X})$ représente la matrice diagonale composée des éléments diagonaux de la matrice $\mathbf{X}$.
La matrice $\mathbf{C}^{(k)}$ représente la matrice des codes de dimension $((N_{chip} + L - 1), N_{bit}^{(k)} M_k L)$ construite comme suit [4],

$$\begin{aligned} \mathbf{C}^{(k)} &= \left[\mathbf{v}_{0,0,0}^k, ..., \mathbf{v}_{N_{bit,k}-1, M_k-1, L-1}^k\right] \\ \mathbf{v}_{n,m,l}^k &= \left[\mathbf{0}_{nG_k}^T, \mathbf{u}_{m,l}^{k^T}, \mathbf{0}_{(N_{bit,k}-n-1)G_k}^T\right]^T \\ \mathbf{u}_{m,l}^k &= \left[\mathbf{0}_{t_l}^T, \mathbf{c}_m^{k^T}, \mathbf{0}_{L-t_l-1}^T\right]^T \\ \mathbf{c}_m^{k^T} &= \left[c_m^k(1), ..., c_m^k(G_k)\right]^T \end{aligned}$$

$n = 0, ..., N_{bit,k} - 1$, $m = 0, ..., M_k - 1$ et $l = 0, ..., L - 1$. $\mathbf{0}_n$ désigne le vecteur nul de dimension n. Le vecteur $\mathbf{n}$, de longueur $N_{chip} + L - 1$, représente le vecteur de bruit ayant N_0 comme densité spectrale de puissance.

3 Structures des détecteurs joints multi-codes

Dans cette section, nous dérivons les rapports RSI à la sortie des récepteurs : RAKE, ZF et MMSE [5, 4]. Notons que les expressions des RSI seront intégrés dans notre algorithme présenté dans la section 4.
L'utilisateur 1 est considéré comme l'utilisateur d'intérêt. En posant $\mathbf{A}^{(k)} = \mathbf{C}^{(k)} \tilde{\mathbf{H}}^{(k)}$, le signal reçu s'exprime comme suit :

$$\begin{aligned} \mathbf{r} &= \mathbf{A}^{(1)} \mathbf{W}^{(1)} \mathbf{b}^{(1)} + \sum_{k=2}^{K} \mathbf{A}^{(k)} \mathbf{W}^{(k)} \mathbf{b}^{(k)} + \mathbf{n} \qquad (7) \\ &= \underbrace{\mathbf{AWb}}_{\text{signal utile}} + \underbrace{\qquad}_{\text{MAI+ISI+bruit}} \qquad (8) \end{aligned}$$

Ainsi, à la sortie du récepteur linéaire, le rapport RSI s'écrit en général,

$$RSI(m,n) = \frac{E\left(\left|\widehat{b}_m^{(1)}(n)\right|\right)^2}{E\left(|ICI + ISI + MAI + bruit|^2\right)} \tag{9}$$

pour $m = 1...M_1$ et $n = 1...N_{bit,1}$.

3.1 Récepteur RAKE

La sortie d'un banc de récepteur RAKE s'écrit :

$$\mathbf{y}_{RAKE} = \mathbf{A}^H \mathbf{r} = \mathbf{MWb} + \mathbf{A}^H \tag{10}$$

où $\mathbf{M} = \mathbf{A}^H \mathbf{A}$.
Nous séparons le terme utile, l'interférence ICI générée par la transmission en multi-codes et le terme MAI+ISI+bruit généré par le bruit et les autres utilisateurs actifs,

$$\mathbf{y}_{RAKE} = diag\{\mathbf{MW}\}\mathbf{b} + \overline{diag}\{\mathbf{MW}\}\mathbf{b} + \mathbf{A}^H \tag{11}$$

où $\overline{diag}(\mathbf{X}) = \mathbf{X} - diag(\mathbf{X})$ représente la matrice $\mathbf{X}$ ayant une diagonale nulle.

Le rapport RSI à la sortie du récepteur relatif au n^{me} symbole transmis sur le m^{me} code s'exprime par l'équation (12). $\mathbf{M'} = \mathbf{MW}$ et $\mathbf{R} = E\left[^{T}\right]$ représente la matrice de covariance du MAI, ISI et de bruit [5].

$$RSI_{RAKE}(m,n) = \frac{([\mathbf{M'}]_{j,j})^2}{[(\mathbf{M'})^2]_{j,j} - [(\mathbf{M'})_{j,j}]^2 + [\mathbf{M'RM'}]_{j,j}} \quad (12)$$

pour $j = m+(n-1)M_1$ où $m = 1...M_1$ et $n = 1...N_{bit,1}$. $[\mathbf{X}]_{j,j}$ désigne l'élément du j^{th} ligne et du j^{th} colonne de la matrice.

3.2 Récepteur ZF

La sortie du récepteur ZF s'exprime comme suit :

$$\mathbf{y}_{ZF} = \left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{y}_{RAKE} = \mathbf{Wb} + \left(\mathbf{A}^H\mathbf{A}\right)^{-1}\mathbf{A}^H \quad (13)$$

Le rapport RSI relatif au n^{me} symbole transmis sur le m^{me} code s'exprime comme suit :

$$RSI_{ZF}(m,n) = \frac{[\mathbf{W}]^2_{j,j}}{[\mathbf{M}^{-1}\mathbf{A}^H\mathbf{RAM}^{-H}]_{j,j}} \quad (14)$$

pour $j = m+(n-1)M_1$ où $m = 1...M_1$ et $n = 1...N_{bit,1}$.

3.3 Récepteur MMSE

Le récepteur MMSE minimise l'erreur quadratique $E\left\{\|\mathbf{Fy}_{RAKE} - \mathbf{Wb}\|^2\right\}$. La solution optimale est délivrée par l'expression suivante :

$$\mathbf{F} = \mathbf{W}^2\mathbf{M}^H\left[\mathbf{MW}^2\mathbf{M}^H + \mathbf{A}^H\mathbf{RA}\right]^{-1} \quad (15)$$

La sortie du récepteur MMSE s'écrit, ainsi, en notant $\mathbf{W}_0 = \mathbf{FM}$:

$$\mathbf{y}_{MMSE} = \mathbf{Fy}_{RAKE} = \mathbf{W}_0\mathbf{Wb} + \mathbf{W}_0\mathbf{M}^{-1}\mathbf{A}^H \quad (16)$$

En séparant, le terme utile des interférences et du bruit, nous avons :

$$\mathbf{y}_{MMSE} = diag\{\mathbf{W}_0\mathbf{W}\}\mathbf{b} + \overline{diag}\{\mathbf{W}_0\mathbf{W}\}\mathbf{b} + \mathbf{W}_0\mathbf{M}^{-1}\mathbf{A}^H\mathbf{A}^H \quad (17)$$

Le rapport RSI à la sortie du récepteur MMSE relatif au n^{me} symbole transmis sur le m^{me} code s'exprime en notant $\mathbf{W'} = \mathbf{W}_0\mathbf{W}$,

$$RSI_{MMSE}(m,n) = \frac{([\mathbf{W'}]_{j,j})^2}{[\mathbf{W'W'}^H]_{j,j} - ([\mathbf{W'}]_{j,j})^2 + [\mathbf{W}_0\mathbf{M}^{-1}\mathbf{A}^H\mathbf{RAM}^{-1}\mathbf{W}_0^H]_{j,j}} \quad (18)$$

pour $j = m+(n-1)M_1$ où $m = 1...M_1$ et $n = 1...N_{bit,1}$.

4 Contrôle de puissance et détection jointe multicodes combiné

Dans cette section, nous proposons la combinaison du contrôle de puissance descendant et l'adaptation du détecteur joint multicodes. L'objectif de l'algorithme est d'assurer un rapport RSI, à la sortie du récepteur, égale au rapport RSI cible, RSI_{cible} pour chaque code du schéma de transmission en multi-code. Pour cet effet, nous exploitons la relation linéaire entre le rapport RSI et la puissance de transmission de chaque code comme il est montré dans les équations ((12), (14) et (18)). L'algorithme proposé suit les étapes suivantes :

1) $t = 0$, initialisation des puissances de chaque code $p_0^{(1)}, ..., p_{M_1}^{(1)}$.
2) Calcul des paramètres du récepteur et le rapport RSI,
3) Mise à jour des puissances des codes : $p_m^{(1)}(t+1) = \frac{RSI_{cible}^{(1)}}{E_n[RSI(m,n)]}p_m^{(1)}(t)$, pour $m = 1...M_1$.
4) $[\mathbf{W}(t+1)]_{j,j} = \sqrt{p_m^{(1)}(t+1)}$, avec $j = m+(n-1)M_1$ où $m = 1...M_1$ et $n = 1...N_{bit,1}$.
5) $t = t+1$, stop si la convergence est atteinte ; sinon, aller à l'étape 2.

L'indice (1) désigne l'utilisateur multicodes dans la cellule.
Si on suppose une limitation de puissance de transmission p_m^{max},pour $m = 1...M_1$, l'étape 3) de l'algorithme proposé est modifiée par :

$$p_m^{(1)}(t+1) = min\left\{\frac{RSI_{cible}^{(1)}}{E_n[RSI(m,n)]}p_m^{(1)}(t), p_m^{max}\right\} \quad (19)$$

La répartition des puissances des codes obtenue au niveau de l'étape 3 est transmise via le canal de retour à la station de base.

5 Résultats de simulations

Les résultats de simulation analysent les performances de l'algorithme proposé en comparant la puissance totale transmise et la rapport RSI obtenu par chaque détecteur au niveau du mobile. Les utilisateurs sont répartis aléatoirement dans une cellule hexagonale de rayon $R = 1000m$. L'atténuation est choisi égale à $\beta = 4$. Le RSI cible est fixé à $RSI_{cible} = 4$. Nous considérons K=20 utilisateurs, mono-code, et un utilisateur multi-code (utilisateur d'intérêt). Le facteur d'étalement des utilisateurs mono-codes est choisi égale à $G_k = 128$ pour chaque $k = 2..K$. L'utilisateur multi-code est caractérisé par un facteur d'étalement égale à $G_1 = 64$. Nous faisons varier le nombre des codes alloués à l'utilisateur multi-code entre $M_1 = 4$ et $M_1 = 64$. Nous supposons, aussi, la présence d'un canal à 6 trajets.
La figure 1 présente la moyenne du RSI à la sortie des trois détecteurs. Nous remarquons une convergence en une itération du détecteur ZF, une convergence rapide du récepteur MMSE et une convergence lente du récepteur RAKE.
La figure 2 présente la variation de la puissance de la station de base : p_{ZF}/p_{RAKE} et p_{MMSE}/p_{RAKE} en fonction des itérations, dans le cas de $M_1 = 64$ codes. Nous remarquons une minimisation de $\approx$: 35% de la puissance transmise par la station de base. Notons que le récepteur RAKE n'atteint pas le rapport RSI cible, comme le montre la figure 2. Cependant, les récepteurs (ZF et MMSE) présente des meilleures performances.
Le détecteur MMSE présente les meilleures performances en terme de minimisation de la puissance transmise et la maximisation de la capacité du système. Le détecteur ZF présente des performances moindres dans le cas d'une charge du système assez importante. Cependant, notons que dans le cas d'une charge

faible (cas de $M_1 = 4$), les performances des trois détecteurs étudiés sont comparables.
Sur la figure 4, nous montrons qu'à la convergence nous n'obtenons pas une puissance égale sur chaque code puisque chaque code ne voit pas la même interférence.

6 Conclusion

Dans ce papier, nous avons étudier l'optimisation conjointe des paramètres des récepteurs linéaires joints multicodes et du contrôle de puissance. L'algorithme proposé nous permet de minimiser la puissance transmise par la station de base et d'obtenir la répartition optimale des puissances des codes au niveau de la station de base. Nous avons montré, à cet effet, que le récepteur joint de type MMSE diminue la puissance transmise de l'ordre de 35% par rapport à un récepteur en râteau.

Références

[1] 3GPP TR 25.858 V5.0.0 (2002-03), "High Speed Downlink Packet Access : Physical Layer Aspects, (Release 5)".

[2] S. Moshavi, "Multi-user Detection for DS-CDMA Communications", *IEEE Communications Magazine*, vol. 34, issue : 10, pp. 124-136, Oct. 1996.

[3] H. Holma and A. Toskala, Eds., "WCDMA for UMTS-Radio Access for Third Generation Mobile communications", New-York : Wiley, 2000.

[4] B. Sayadi and I. Fijalkow, "Joint Detection for Multicode Transmission in Downlink High Speed UMTS", *IEEE Vehicular Technology Conference*, VTC 2004, Los Angeles, September 2004.

[5] S. Ataman, "Power Control and Joint Detection in CDMA Systems", PhD thesis, Université Paris-Sud, july 2004.

[6] L. Song and J. M. Holtzman, "CDMA Dynamic Downlink Power Control", *IEEE Vehicular Technology Conference,* pp. 1101-1105, VTC 1998.

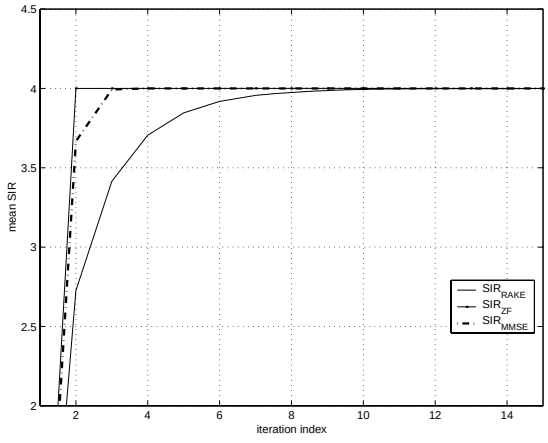


FIG. 1 – La convergence du RSI pour les détecteurs : RAKE, ZF et MMSE, dans le cas $M_1 = 4$.

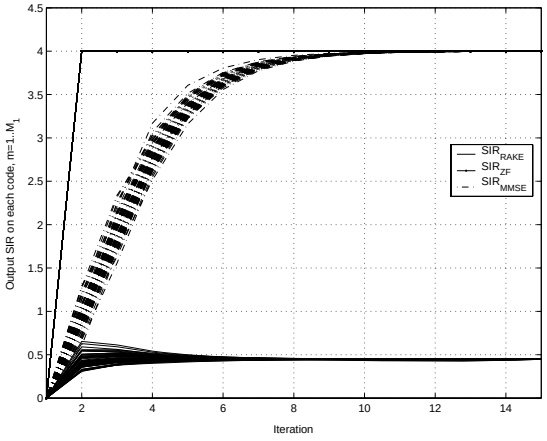


FIG. 2 – La convergence du rapport RSI à la sortie des récepteurs RAKE, ZF et MMSE dans le cas de $M_1 = 64$.

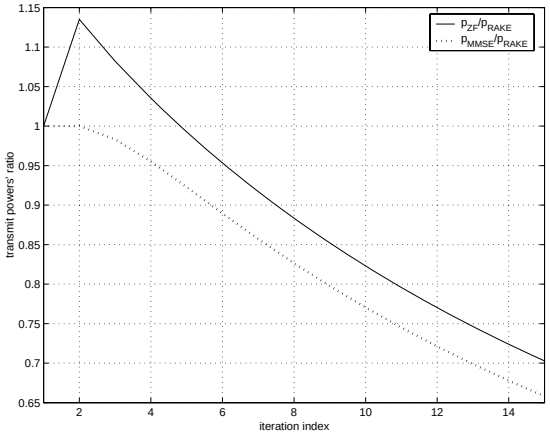


FIG. 3 – Le rapport des moyennes des puissances transmises p_{ZF}/p_{RAKE} et p_{MMSE}/p_{RAKE} en fonction des itérations pour un schéma multicodes à 64 codes.

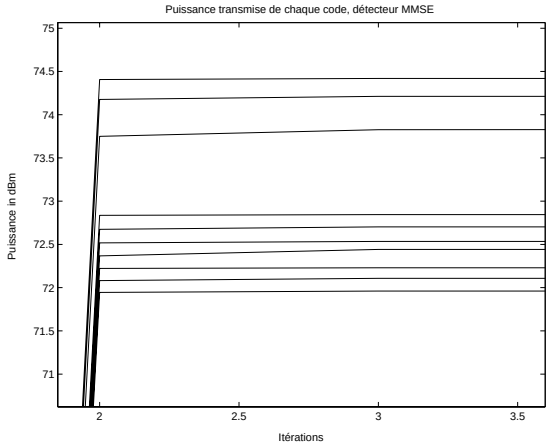


FIG. 4 – Répartition optimale des puissances sur les différents codes : cas d'un schéma à 10 codes.

Annulation d'interférence dans les systèmes optiques à accès multiple par répartition de codes (O-CDMA)

C. GOURSAUD, A. JULIEN-VERGONJANNE., C. AUPETIT-BERTHELEMOT, J.P. CANCES, et J.M. DUMAS.

UMOP/GESTE CNRS FRE 2701, Université de Limoges, Parc ESTER, B.P. 6804, 87068 Limoges

goursaud@ensil.unilim.fr

Résumé – Dans cet article, nous présentons l'étude d'un récepteur à annulation parallèle d'interférence précédé d'un limiteur optique, pour un système DS-CDMA optique. Le but de ce récepteur est de réduire l'Interférence d'Accès Multiple (IAM) due aux codes optiques unipolaires. Nous montrons que l'utilisation du récepteur présenté permet d'améliorer le système, au niveau de la longueur des codes utilisés, ou au niveau de la gestion des séquences allouées aux utilisateurs actifs.

Abstract *– In this paper, we study the parallel interference cancellation receiver, with hard limiter, for optical Direct Sequence Code Division Multiple Access (DS-OCDMA). The aim of such a receiver, is to mitigate the Multiple Access Interference (MAI) linked to the codes unipolarity. We show that by using this receiver, we can improve the system, either by reducing the code length, or giving flexibility for the codes assignment.*

1. Introduction

L'application des techniques CDMA, couramment utilisées en radiofréquences, est envisagée pour les réseaux optiques, en particulier dans le contexte des réseaux locaux haut débit [1-3].

Les systèmes optiques basés sur une modulation non-cohérente de la lumière imposent l'unipolarité des codes d'étalement, puisque l'intensité lumineuse est une quantité positive ou nulle. Dans ce cas, et contrairement aux systèmes CDMA radiofréquences, les codes ne peuvent pas être strictement orthogonaux. Les performances des systèmes O-CDMA sont donc limitées par l'Interférence d'Accès Multiple (IAM). Dans cette étude, nous présentons une méthode d'annulation parallèle (appelée PIC) développée pour les systèmes radiofréquences, appliquée à un système Optique CDMA à Séquence Directe (DS-OCDMA), les codes d'étalement étant des séquences unipolaires quasi orthogonales appelées OOC [1]. Le récepteur étudié est constitué d'un dispositif limiteur optique placé devant une structure PIC.

L'expression théorique de la probabilité d'erreur a été établie dans le cas d'un système optique DS-OCDMA incohérent synchrone et utilisant des codes OOC. En comparant l'efficacité de ce nouveau récepteur (appelé LO+PIC) avec les récepteurs à détection par corrélation et à annulation d'interférence sans limiteur, nous en déduisons la supériorité de la structure étudiée non seulement en terme de performances mais aussi en ce qui concerne les possibilités de réalisation. En effet, il résulte de l'analyse que le récepteur à annulation parallèle d'interférence avec un limiteur nécessite un facteur d'étalement plus faible que dans les autres cas, pour une performance et un débit donnés.

2. Description d'un système DS-OCDMA

Dans un système DS-OCDMA, les utilisateurs transmettent des données binaires équiprobables et indépendantes sur une fibre optique. La différenciation des utilisateurs se fait par multiplication des données par un code. Ce code doit être spécifique à chaque utilisateur, de façon à pouvoir extraire les données par comparaison du signal reçu avec le code de l'utilisateur désiré.

Les codes étudiés dans cet article sont des codes quasi-orthogonaux OOC [1] définis par $(F,W,\lambda_a,\lambda_c)$, où F est la longueur du code (c'est à dire le nombre total d'impulsions appelées chips dans un temps bit), W le poids (c'est à dire le nombre de chips ayant la valeur 1), et λ_a et λ_c les contraintes d'auto et d'inter-corrélation, que nous avons fixées à 1. Dans ce cas, le nombre maximal d'utilisateurs possibles dans la famille est :

$$N = \lfloor (F-1)/W(W-1) \rfloor \qquad (1)$$

Au niveau du récepteur, le dispositif le plus simple est le Détecteur Conventionnel par Corrélation (DCC)[3]. Grâce à une multiplication du signal reçu par le code de l'utilisateur désiré, il estime la puissance totale reçue aux positions des chips à 1, pour la comparer à un seuil S_T. Avec un tel récepteur et dans un système optique idéal, seuls les 0 peuvent être mal détectés, ceci lorsque les contributions des autres utilisateurs augmentent suffisamment la puissance reçue aux positions des chips à 1. Il a été montré que la probabilité d'erreur d'un tel récepteur dans le cas synchrone [1] est :

$$P_{EDCC} = \frac{1}{2}\sum_{i=S_T}^{N-1} C_{N-1}^{i} \left(W^2/2F\right)^i \left(1-W^2/2F\right)^{N-1-i} \qquad (2)$$

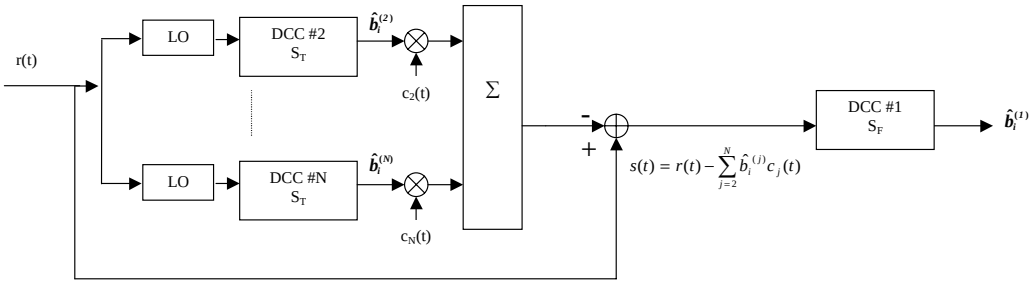


FIG 1 : La structure à annulation parallèle d'interférence avec limiteur : LO+PIC

Pour améliorer les performances de ce récepteur, un Limiteur Optique (LO)[4] peut être utilisé. Ce composant délivre une puissance unitaire si la puissance reçue est supérieure ou égale à 1, et une puissance nulle sinon. Ainsi, lorsque $S_T=W$, tous les chips à 1 du code doivent subir une interférence pour qu'il y ait erreur. Il a été montré [1] que la probabilité d'erreur d'un tel récepteur LO+DCC dans le cas synchrone est :

$$P_{ELO+DCC} = 1/2\, C_W^S \prod_{i=0}^{S-1} \left(1-\left(1-W^2/2F\right)^{N-1-i}\right) \quad (3)$$

3. Annulation parallèle d'interférence

3.1 Principe

Une structure à annulation parallèle d'interférence (PIC) [5-6] estime les contributions des utilisateurs non-désirés en détectant leurs données $\hat{b}_i^{(j)}$ par des récepteurs DCC, pour ensuite soustraire les contributions estimées du signal reçu. Puis, la donnée de l'utilisateur désiré est évaluée par un récepteur DCC. Afin d'améliorer les performances d'une telle structure, la détection des utilisateurs non désirés peut se faire avec un récepteur LO+DCC. Nous appellerons ce récepteur LO+PIC. Grâce aux limiteurs placés devant les récepteurs des utilisateurs non désirés, les données sont mieux estimées qu'avec le PIC, donc l'interférence est mieux évaluée (Fig 1).

Afin d'étudier les performances d'une telle structure, nous développons l'expression théorique de la probabilité d'erreur de ce récepteur.

3.2 Analyse théorique

Pour établir l'expression de la probabilité d'erreur $P_{ELO+PIC}$ nous nous plaçons dans le cas synchrone, avec des codes OOC *(F,W)*, et nous considérons que les seuils des récepteurs sont S_T *(*$0<S_T\leq W$*)* pour les utilisateurs non désirés et S_F *(*$0<S_F\leq W$*)* pour l'utilisateur désiré. Nous considèrerons que l'utilisateur désiré est l'utilisateur #1.

Le développement de l'expression théorique de la probabilité d'erreur du LO+PIC est basé sur des travaux publiés concernant le PIC [6], dans lesquels nous avons montré que, contrairement au DCC, le PIC ne peut faire une erreur que lorsque la donnée envoyée est un 1. De plus, nous avons exprimé la probabilité d'erreur du PIC :

$$P_{EPIC} = (1/2)^N . \sum_{N_1=S_T-1}^{N-1} \sum_{N_2=W+1-S_F}^{N-1-N_1} C_{N-1}^{N_1} C_{N-1-N_1}^{N_2} (P_{IPIC})(1-P_{IPIC})^{N-1-N_1-N_2} \quad (4)$$

où

$$P_{IPIC} = W^2/F . \sum_{n_1=S_T-1}^{N_1} C_{N_1}^{n_1} \left(W^2/F\right)^{n_1} \left(1-W^2/F\right)^{N_1-n_1} \quad (5)$$

L'expression (4), basée sur le fait que le récepteur final utilisé est un DCC, fait apparaître le terme P_{IPIC} qui est la probabilité pour un utilisateur non désiré d'être un interférent, c'est-à-dire un utilisateur ayant envoyé un 0 détecté comme un 1, et dont le code a un chip de valeur 1 en commun avec le code de l'utilisateur #1. Comme nous utilisons un DCC dans le LO+PIC pour détecter l'utilisateur #1 à partir du signal "nettoyé", nous aurons la même expression, mais avec un terme différent pour P_{IPIC} puisque les utilisateurs non désirés sont détectés différemment.

Nous devons donc exprimer la probabilité $P_{ILO+PIC}$ qu'un utilisateur non désiré soit un interférent.

3.3 Expression théorique de $P_{ILO+PIC}$

A partir de maintenant, nous considérons que l'utilisateur #1 a émis un 1, et que N_1 utilisateurs non désirés ont émis un 1. En conséquence, $N-1-N_1$ utilisateurs non désirés ont émis un 0.

Un utilisateur non désiré #j ayant émis un 0 sera un interférent si il a un chip en commun avec l'utilisateur #1, et si sa donnée est détectée avec erreur comme étant un 1. Donc, $P_{ILO+PIC}$ peut s'écrire :

$$P_{ILO+PIC} = P\left(\int_0^T c_1(t).c_j(t).dt = 1\right) \times P\left(Z_i^{(j)} \geq S_T \,/\, \left(\int_0^T c_1(t).c_j(t).dt = 1 \cap b_i^{(j)} = 0 \cap b_i^{(1)} = 1\right)\right) \quad (6)$$

où $Z_i^{(j)}$ est la variable décisionnelle de #j et s'écrit :

$$Z_i^{(j)} = W.b_i^{(j)} + \sum_{k=1,k\neq j}^{N} b_i^{(k)}.\int_0^T c_j(t)c_k(t).dt \qquad (7)$$

La probabilité que 2 codes aient un chip en commun est :

$$P(\int_0^T c_u(t)c_v(t).dt = 1) = W^2/F \qquad (8)$$

Or, si cette condition est réalisée entre #1 et #j, alors cela signifie que #1 (qui a émis un 1) génère une interférence de valeur +1 sur la variable décisionnelle de #j. En conséquence, la contribution des autres utilisateurs doit être supérieure à S_T –1 pour créer une erreur. Ainsi, nous obtenons :

$$P_{ILO+PIC} = \frac{W^2}{F}.P(\sum_{k=2,k\neq j}^{N} b_i^{(k)}.\int_0^T c_j(t)c_k(t).dt \geq S_T - 1) \qquad (9)$$

Le LO implique que les seuls motifs d'interférence entraînant une erreur sont ceux pour lesquels au moins S_T *-1* chips différents de #j sont en commun avec des utilisateurs non désirés ayant émis un 1, en plus de celui en commun avec #1. $C_{W-1}^{S_T-1}$ combinaisons sont possibles pour avoir S_T *-1* recouvrements parmi les *W-1* chips restants.

Or la probabilité *p* pour un chip d'être en commun avec au moins 1 utilisateur parmi *L*, est le complémentaire de la probabilité *q* de n'être recouvert par aucun des *L* utilisateurs. Comme :

$$q = (1 - W/F)^L \qquad (10)$$

alors

$$p = 1 - (1 - W/F)^L \qquad (11)$$

De plus, le premier chip des *W-1* restants peut être recouvert par n'importe lequel des N_1 utilisateurs ayant envoyé un 1. Cependant, pour le deuxième chip, il reste *(N_1 – 1)* utilisateurs disponibles lorsque le premier chip est recouvert... Donc, la probabilité que #j ait des recouvrements sur S_T*–1* chips différents parmi les *W-1* restants est :

$$P(\sum_{k=2,k\neq j}^{N} b_i^{(k)}.\int_0^T c_j(t)c_k(t).dt \geq S_T - 1) = C_{W-1}^{S_T-1}\prod_{k=0}^{S_T-2}\left(1-(1-W/F)^{N_1-k}\right) \qquad (12)$$

Nous obtenons finalement :

$$P_{ILO+PIC} = W^2/F \times C_{W-1}^{S_T-1}\prod_{k=0}^{S_T-2}\left(1-(1-W/F)^{N_1-k}\right) \qquad (13)$$

D'où

$$P_{ELO+PIC} = (1/2)^N.\sum_{N_1=S_T-1}^{N-1}\sum_{N_2=W+1-S_F}^{N-1-N_1} C_{N-1}^{N_1}C_{N-1-N_1}^{N_2}(P_{ILO+PIC})(1-P_{ILO+PIC})^{N-1-N_1-N_2} \qquad (14)$$

Nous pouvons remarquer que si

$$W+1-S_F > N-S_T \qquad (15)$$

alors il n'y a aucun terme dans les sommations (4) et (14). Pour ces combinaisons de paramètres, la probabilité d'erreur est nulle. Dans ces conditions, le PIC et le LO+PIC permettent de supprimer toutes les erreurs.

De plus, nous pouvons remarquer que la probabilité d'erreur sera minimale pour S_T maximal et S_F minimal, c'est-à-dire $S_T = W$ et $S_F = 1$.

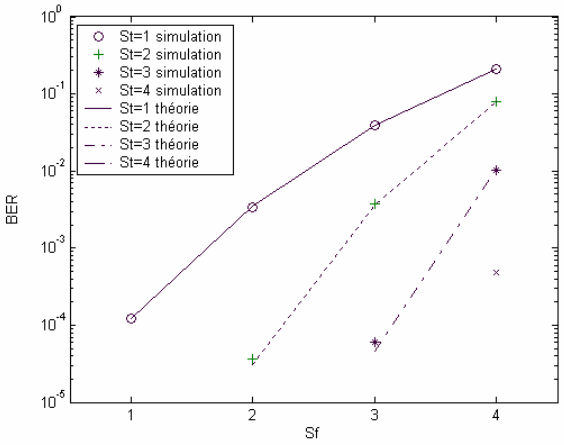


FIG 2 Comparaison du de la théorie et de la simulation pour le LO+PIC pour un OOC (64,4) avec N=5

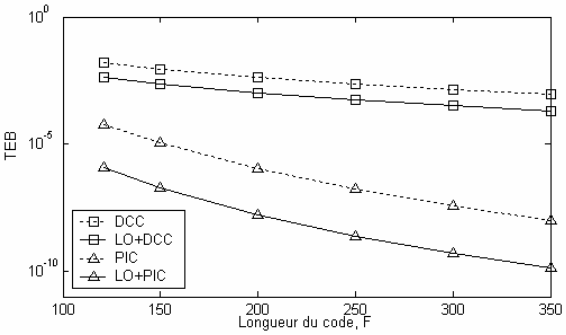


FIG 3 : Performances du LO+PIC pour un OOC(F,3) avec N=20

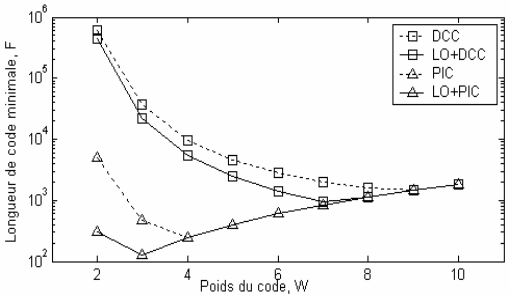


FIG 4 : Longueur minimale nécessaire pour avoir un TEB<10^{-9} pour N=20 en fonction du poids W

4. Analyse des résultats

Tout d'abord, nous avons tracé sur la figure 2, les taux d'erreurs binaires obtenus d'une part par simulation, et d'autre part grâce à l'expression théorique (14). Nous pouvons constater que les résultats théoriques correspondent à ceux de la simulation. Nous pouvons donc considérer que l'équation (14) décrit correctement les performances du récepteur LO+PIC dans un cas synchrone.

De plus, nous pouvons observer qu'il manque des points sur la courbe. Ces points, pour lesquels les seuils vérifient l'inégalité (15), correspondent à une probabilité d'erreur nulle. Afin de vérifier cette inégalité, nous avons simulé une chaîne de transmission avec 10^{13} bits envoyés pour des seuils $S_T = 2$ et $S_F = 1$ (c'est-à-dire pour un point frontière de

l'inégalité), et nous n'avons décelé aucune erreur. En comparaison avec les valeurs des probabilités non nulles, nous pouvons considérer que l'inégalité est vérifiée.

De plus, nous pouvons extrapoler à partir de la courbe que les seuils optimaux sont : $S_T = W$ et $S_F = 1$.

Après avoir validé l'expression théorique (14) de $P_{ELO+PIC}$, grâce aux résultats de simulation, nous avons réalisé une étude comparative des performances des différents récepteurs DCC, LO+DCC, PIC, LO+PIC en fonction des paramètres du code et du nombre d'utilisateurs. Pour chaque récepteur, nous avons considéré les seuils optimaux, c'est-à-dire $S = W$ pour le DCC et le LO+DCC et $S_T = W$ et $S_F = 1$ pour le PIC et LO+PIC. D'une part, nous avons constaté, comme par exemple sur la figure 3, que l'évolution des performances était la même pour tous les récepteurs, mais que le LO+PIC est plus performant pour un code donné. De plus, nous avons comparé les performances de notre système à ceux d'une étude récemment publiée sur un système de réception DS-OCDMA avec des limiteurs optiques [7], et nous obtenons de meilleurs résultats. Par exemple, pour un code *(43,3)*, pour *N=7* la structure dans [7] donne un TEB d'environ 10^{-4}, alors que le LO+PIC donne un TEB de 3.10^{-8}.

D'autre part, la supériorité du LO+PIC permet de réduire la contrainte sur la longueur du code. En effet, lorsque la longueur du code (c'est-à-dire le facteur d'étalement) augmente à un débit donné, la largeur des impulsions à générer est de plus en plus petite, nécessitant des sources optiques irréalisables actuellement. Nous avons tracé sur la figure 4, l'évolution de la longueur *F* minimale requise pour que le TEB soit inférieur à 10^{-9}, ceci en faisant varier le poids *W*, pour *N=20* utilisateurs actifs dans le réseau. Nous pouvons observer que dans un premier temps, les courbes décroissent.

Ceci est dû au fait que lorsque le poids *W* du code augmente pour une même longueur *F*, la probabilité d'erreur diminue. Ainsi, lorsque *W* augmente, il faut que *F* diminue pour obtenir un TEB de 10^{-9}. Dans un second temps, les courbes croissent. Ceci est dû au fait que, lorsque *W* augmente, il faut que la longueur *F* du code augmente pour que l'on puisse avoir 20 utilisateurs. Le changement de pente des courbes induit un code optimal pour lequel la longueur *F* est minimale. Ainsi, ces codes optimaux sont pour le DCC : *(1441,9)*, le LO+DCC : *(950,7)*, le PIC : *(241,4)*, et pour le LO+PIC : *(126,3)*. Dans un système réaliste, le débit des chips est limité par la bande passante des composants. Ainsi, le LO+PIC, en réduisant la longueur du code donc le débit des chips, pour les mêmes performances est intéressant dans le cas d'une implémentation réelle.

De plus, nous pouvons remarquer que pour une longueur de *F=241*, le LO+PIC obtient des performances de 10^{-9} avec un poids de *W=2*, alors que le PIC l'obtient avec un poids de *W=4*. Or avec un code *(241,2)*, il peut y avoir 120 utilisateurs, alors que pour *(241,4)* on peut en avoir 20. Ainsi, dans le cas où nous utilisons le LO+PIC avec un code *(241,2)*, les 120 séquences peuvent être assignées à 120 utilisateurs, et seulement 20 d'entre eux sont autorisés à communiquer simultanément. A l'inverse, avec le PIC, les 20 séquences doivent être en permanence réassignées parmi les utilisateurs.

En conclusion, le LO+PIC peut permettre de réduire la longueur du code ou d'amener de la flexibilité dans le système.

5. Conclusion

Nous avons étudié une technique d'annulation d'interférence pour un système O-CDMA, basé sur un récepteur à annulation parallèle d'interférence couplé à un limiteur, que nous avons appelé LO+PIC. Nous avons établi l'expression théorique de la probabilité d'erreur de cette structure, dans le cas de codes OOC.

Nous avons montré qu'un tel récepteur est plus performant que les récepteurs conventionnels, et que les récepteurs à annulation d'interférence récemment publiés. De plus, nous avons prouvé que, grâce à son efficacité, le LO+PIC permet d'améliorer les performances, de réduire la longueur du code pour une implémentation plus réaliste ou de permettre une flexibilité dans l'assignation des séquences parmi les utilisateurs actifs.

Références

[1] J.A. Salehi "Code Division Multiple-Access Techniques in Optical Fiber Networks –Part I and Part II", in IEEE Trans. Commun.,vol 37, n°8, pp 824-842, august 1989

[2] A.Stok, E.Sargent "Lighting the local area : Optical Code-Division Multiple Access and Quality of Service Provisioning" in IEEE Network ,pp 42-46, Nov/dec 2000.

[3] S. Zahedi, J.A. Salehi "Performance Analysis for various Fiber-Optic CDMA receivers structures", J.Lightwave Techn., vol 18, n°12, pp1718-1727, Dec 2000.

[4] J.J.Chen, G-C Yang "CDMA Fiber-Optic systems with optical hard-limiter" in J.Lightwave Techn.,vol 19, n°7, pp 950-958, july 2001.

[5] C. Goursaud, Naufal M.Saad, Y.Zouine, A.Julien-Vergonjanne, C. Aupetit-Berthelemot, J.P. Cances, J.M. Dumas "Parallel Multiple Access Interference cancellation in Optical DS-CDMA systems", in Annals of Telecommunications, vol 9-10, october 2004, pp 1053-1068.

[6] S.Verdù "Multiuser detection", in Cambridge University Press, New-York, USA,1998.

[7] K. Wakafuji, T. Ohtsuki. "Direct-Detection Optical CDMA Receiver With Interference Estimation and Double Optical Limiters" in. IEEE J.Lightwave Techn (2003), 21, n°10, pp. 2182-2188.

Méthode hybride de suppression d'interférence basée sur l'estimation par paquet des codes inconnus appliquée au HSDPA multicode

Ahmed Saadani, Jean-Marie Chaufray

France Telecom Division R&D
38-40 rue du Général Leclerc, 92794 Issy les Moulineaux Cedex 9-France
ahmed.saadani, jeanmarie.chaufray@rd.francetelecom.com

Résumé – Cet article traite la suppression des interférences causées par l'utilisation des multicodes du HSDPA et ceux des autres utilisateurs lorsque le canal est sélectif en fréquence. La taille fixe des codes d'étalement pour tout les utilisateurs HSDPA rend réalisable l'estimation des codes inconnus au récepteur. Un récepteur hydbride SIC/PIC utilisant deux étapes différentes d'annulation souple d'interference est proposé.

Abstract – This paper deals with the cancellation of the interference caused by the HSDPA multicode and multiuser access on a frequency selective channel. Since HSDPA spreading codes are constant on one frame, the unknown codes can be estimated. An hybrid PIC/SIC receiver using two different soft interference cancellation methods is proposed.

1 Introduction

La norme HSDPA (High Speed Data Packet Access [1]) est une évolution de la norme UMTS, compatible avec cette dernière, et destinée à améliorer les transmissions en mode paquet. Elle permet d'atteindre des débits instantanés allant jusqu'à 10 MBps en liaison descendante. Ceci est rendu possible par l'utilisation sur le canal physique HS-DSCH (High Speed Downlink Shared Channel) des modulations QPSK et 16-QAM, d'un faible facteur d'étalement égal à 16 et par l'allocation de plusieurs codes à un seul utilisateur [2][3].

Le passage du signal transmis dans un canal à trajets multiples entraîne une perte d'orthogonalité entre les codes à la réception. Cette perte se manifeste à la sortie d'un récepteur en râteau (Rake) par la présence d'interférences entre les symboles, les codes et les trajets. La puissance de ces interférences n'est pas négligeable et elle peut affecter significativement les performances du système. En effet, la modulation 16-QAM est très sensible aux interférences, lequelles sont d'autant plus importantes que le nombre de codes utilisés est élevé.

Afin d'éliminer ces interférences, deux principales approches existent dans la littérature. Une approche linéaire basée sur une égalisation linéaire du canal (ZF, LMMSE) [4] avant le désétalement : l'orthogonalité est partiellement rétablie à la sortie de l'égaliseur. La deuxième approche est non linéaire et est basée sur l'annulation d'interférences (Interference Cancellation notée IC) [5]. Selon cette approche, l'interférence est reconstruite à partir de la sortie du récepteur RAKE puis soustraite du signal reçu. Le signal résultant est ensuite désétalé. Cette procédure peut être répétée plusieurs fois de suite (on parle alors d'étages successifs). Nous pouvons distinguer trois méthodes de suppressions. La première est la suppression parallèle des interférences (Parallel IC notée PIC) qui utilise tous les symboles des différents codes à l'entrée de l'étage pour estimer l'interférence de tous les codes puis la supprimer. La deuxième est la suppression en série des interférences (Serial IC notée SIC) qui consiste à supprimer d'abord l'interférence d'un code, puis à ré estimer le symbole étalé par ce code. Ce dernier est utilisé pour l'estimation de l'interférence et la détection du symbole du code suivant. La suppression en série est plus performante en terme de convergence mais elle reste par contre plus complexe que le traitement en parallèle. La dernière méthode est celle qui combine les précédentes et s'appelle "hybride" [6][7].

Afin de mettre en oeuvre l'annulation d'interférence, le récepteur a besoin de savoir quels sont les codes utilisés par la station. Sur la liaison descendante, si toute la ressource est allouée à un seul utilisateur, cette information est disponible au récepteur ; en revanche, si des groupes de codes sont alloués à des utilisateurs différents, chaque récepteur ne connaît que ceux qui lui ont été affectés.

Cependant pour les faibles facteurs d'étalement, le nombre de codes susceptibles d'être utilisés est réduit, ce qui permet d'envisager au niveau du récepteur une détection des codes utilisés par la station, rendant possible la suppression de l'interférence de tous les codes. Dans [8][9][10] plusieurs méthodes ont été proposés pour la détection des codes inconnus. La méthode proposée par [8] est appliquée au HSDPA. Les codes sont détectés symbole par symbole, puis, une suppression en parallèle (PIC) des interférences des tous les codes (connus et estimés) est réalisée en utilisant la sortie souple de l'étage précédent. Dans [9], la détection des codes se base sur la transformée rapide de Walsh. Dans [10] une méthode d'estimation d'interférence des codes par famille suivie d'un SIC a été proposée.

Dans cet article nous considérons un récepteur basé sur une approche hybride de suppression des interférences. Le signal reçu traverse un banc de RAKE associés aux codes connus et inconnus. Les sorties sont injectées dans deux modules PIC connectés en série. Le premier exploite la connaissance de la modulation et les symboles des codes connus pour calculer leur interférence qui est ensuite pondérée en utilisant une approche LLR [11]. Puis il calcule l'interférence des codes détectés et la pondère d'une manière explicite. L'interférence globale est soustraite du signal reçu et seuls les symboles des codes connus sont estimés. Ensuite, la sortie du premier PIC et l'estimation des symboles des codes inconnus sont utilisées par le deuxième PIC afin de calculer puis de supprimer de la même manière l'interference. Tous les symboles des codes inconnus sont estimés et seuls ceux des codes détectés sont considérés à l'étage suivant.

Cette structure a l'avantage d'utiliser cycliquement un groupe de symboles pour améliorer l'estimation d'un autre. La complexité du récepteur n'est pas beaucoup plus élevée que celle d'un PIC global sur tout les codes. Les performances de ces deux récepteurs sont comparées à celles d'un égaliseur linéaire LMMSE.

Cet article est organisé de la manière suivante. Nous présentons d'abord le système de transmission et de réception. Ensuite nous présentons la structure du récepteur proposé. Nous mettons l'accent sur la manière de soustraire les interférences selon que le code est connu ou bien détecté. Après nous présentons l'égaliseur LMMSE au niveau chip. Enfin nous comparons les performances de ce récepteur avec d'autres pour différents scénarios.

2 Modélisation du système

Nous considérons la liaison descendante d'un système HSDPA utilisant N codes d'étalement. Le signal émis associé au n^e code s'écrit :

$$d^{(n)}(t) = \sum_{i=-\infty}^{+\infty} s_i^{(n)} c^{(n)}(t - iT_s) Sr(t - iT_s) \qquad (1)$$

où $s_i^{(n)}$ sont les symboles transmis avec le n^e code d'étalement $c^{(n)}(t)$, T_s est la durée d'un symbole et $Sr(t)$ le code d'embrouillage (scrambling). Les symboles $s_i^{(n)}$ sont ceux d'une modulation QPSK ou 16-QAM, selon le schéma de modulation et de codage choisi par la station de base pour l'utilisateur considéré. Ce signal traverse un canal à L trajets avant d'atteindre le récepteur, de sorte que le signal reçu s'écrit :

$$r(t) = \sum_{n=0}^{N-1} \sum_{l=0}^{L-1} h_l d^{(n)}(t - \tau_l) + w(t) \qquad (2)$$

où h_l et τ_l sont respectivement le coefficient d'atténuation et de déphasage et le retard associés au trajet l, et $w(t)$ est un bruit blanc gaussien complexe circulaire.

Sur la liaison descendante, tous les utilisateurs sont synchronisés. L'émetteur utilise N codes choisis dans un ensemble de N_{max} codes orthogonaux. Parmi ces N codes, N_K sont attribués à l'utilisateur que nous considérons, les $N_U = N - N_K$ codes restant étant alloués à d'autres utilisateurs. Au niveau du récepteur de l'utilisateur considéré, l'entrée du filtre adapté associé au m^e trajet s'écrit :

$$\begin{aligned} y_m = r(t+\tau_m) &= \sum_{n=0}^{N-1} h_m d^{(n)}(t) \\ &+ \sum_{n=0}^{N_K-1} \sum_{l=0, l\neq m}^{L-1} h_l d^{(n)}(t - \tau_l + \tau_m) \\ &+ \sum_{n=N_K}^{N_K+N_U-1} \sum_{l=0, l\neq m}^{L-1} h_l d^{(n)}(t - \tau_l + \tau_m) \\ &+ w(t+\tau_m) \\ &= \sum_{n=0}^{N-1} h_m d^{(n)}(t) + I_{K,m} + I_{U,m} \\ &+ w(t+\tau_m). \end{aligned} \qquad (3)$$

Comme les codes d'étalement sont orthogonaux, le premier terme de cette expression ne contribue pas à l'interférence car tous les termes de la somme disparaissent lors du désétalement, à l'exception du symbole utile. Cependant, les deuxièmes et troisième termes (notés $I_{K,m}$ et $I_{U,m}$) contiennent les interférences causées par les codes connus (les autres codes alloués au même utilisateur) et les codes inconnus (alloués à d'autres utilisateurs). L'objectif de l'annulation d'interférence est de réduire la contribution de ces deux termes.

3 Structure du récepteur

3.1 Description générale

Le récepteur proposé effectue une annulation d'interférence en plusieurs étages. Sa structure est la suivante :

- Puisque certains codes ne sont pas connus du récepteur, le premier étage (numéro 0) consiste à effectuer un RAKE pour chacun des N_{max} codes possibles, de façon à obtenir le vecteur $[\hat{s}_{i,0}^{(0)}, .., \hat{s}_{i,0}^{(N_{max}-1)}]$ où $\hat{s}_{i,0}^{(n)}$ est une estimation du i^e symbole associé au n^e code à l'étage 0. Après avoir procédé à cette estimation sur l'ensemble de la trame, les codes inconnus sont détectés selon la méthode décrite à la section 3.2 et le vecteur $\hat{S}_0 = [\hat{s}_{i,0}^{(0)}, .., \hat{s}_{i,0}^{(N_K+\hat{N}_U-1)}]$ est envoyé vers l'étage suivant du récepteur, $\hat{N}_U$ étant une estimation du nombre N_U de codes inconnus.
- Les étages suivants sont composés de deux blocs PIC connectés en série et d'un détecteur de codes (figure 1). L'entrée $\hat{S}_{j-1}$ est décomposée en deux vecteurs $\hat{S}_{j-1,K} = [\hat{s}_{i,j-1}^{(0)}, .., \hat{s}_{i,j-1}^{(N_K-1)}]$ et $\hat{S}_{j-1,U} = [\hat{s}_{i,j-1}^{(N_K)}, .., \hat{s}_{i,j-1}^{(N_K+\hat{N}_U-1)}]$ contenant respectivement les symboles estimés des codes connus et des codes détectés issus de l'étage précédent.
  Le premier PIC utilise $\hat{S}_{j-1}$ pour annuler les interférences $I_{K,m}, I_{U,m}$ à chaque entrée du filtre adapté (3). Ces interférences sont estimées de la manière suivante :

$$\hat{I}_{K,m} = \sum_{n=0}^{N_K-1} \sum_{l=0, l\neq m}^{L-1} h_l \tilde{d}^{(n)}(t - \tau_l + \tau_m), \qquad (4)$$

et

$$\hat{I}_{U,m} = \sum_{n=N_K}^{N_K+\hat{N}_U-1} \sum_{l=0,l\neq m}^{L-1} h_l \tilde{d}^{(n)}(t-\tau_l+\tau_m). \quad (5)$$

où les signaux $\tilde{d}^{(n)}(t) = \sum_{i=-\infty}^{+\infty} \tilde{s}_i^{(n)} c(t-iT_s) Sr(t-iT_s)$ sont construits à partir des symboles souples $\tilde{s}_i^{(n)}$ estimés à partir des $\hat{s}_i^{(n)}$ selon la méthode décrite aux sections 3.3 et 3.4.
Une fois que ces interférences sont estimées, elles sont soustraites du signal reçu de sorte que la m^e entrée du filtre adapté devient :

$$y_{m,j} = r(t+\tau_m) - \hat{I}_{K,m} - \hat{I}_{U,m}. \quad (6)$$

Seuls les symboles $\hat{S}_{j,K}$ sont récupérés en sortie de ce PIC.
Le second PIC exploite à la fois les sorties du bloc PIC précédent $\hat{S}_{j,K}$ et les symboles d'entrée de l'étage $\hat{S}_{j-1,U}$ pour évaluer l'interférence qui est ensuite retranchée, comme dans le premier PIC. Les symboles $\hat{S}_{j,Umax}$ de tous les codes inconnus sont alors estimés. Ce vecteur de symboles est ensuite envoyé dans le bloc de détection des codes qui ne sélectionne que les symboles $\hat{S}_{j,U}$ des codes détectés. Finalement, les symboles $\hat{S}_{j,K}$ et $\hat{S}_{j,U}$ sont regroupés pour former la sortie $\hat{S}_j$ de l'étage .

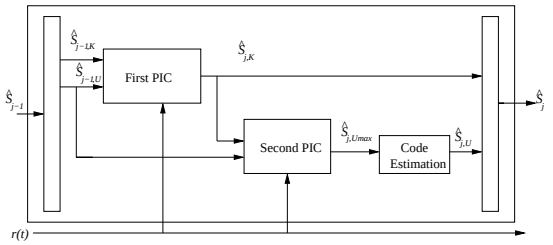


FIG. 1 – Le j^e étage du récepteur hybride

3.2 Détection des codes

La méthode de détection des codes consiste à calculer la puissance $\hat{P}^{(n)}$ des symboles associés à chaque code inconnu. Puisque l'allocation de code n'est mise à jour qu'à chaque trame, l'estimation de la puissance est effectuée en moyennant l'énergie de tous les symboles de la trame (dont le nombre est noté N_f). Un code est supposé utilisé et donc inclus dans (6) lorsque sa puissance dépasse un seuil donné $\mathcal{P}$. L'algorithme est donc le suivant :

$\hat{N}_U = 0$
pour $n = N_c \rightarrow N_{max} - 1$ *faire*
$\hat{P}^{(n)} = 1/N_f \sum_{i=0}^{N_f-1} |\hat{s}_{i,j}^{(n)}|^2$
Si $\hat{P}^{(n)} > \mathcal{P}$ *alors* le code $c^{(n)}(t)$ est utilisé et $\hat{N}_U = \hat{N}_U + 1$
sinon les symboles associés à ce code ne sont pas pris en compte.

Bien entendu, la fiabilité de la détection des codes augmente avec le nombre d'étages. Le fait de ne pas détecter un code pourtant utilisé n'est pas critique : en effet, sa puissance est alors suffisament faible pour ne pas affecter significativement les performances du système.

3.3 Pondération des symboles des codes connus

A chaque étage j, les symboles estimés $\hat{S}_{j,K}$ peuvent être utilisés dans le calcul de l'interférence (4). Cependant, lorsqu'ils ne sont pas fiables, des erreurs supplémentaires peuvent apparaître et se propager aux étages suivants. Afin de réduire l'impact de ce phénomène, un facteur de pondération est appliqué. Puisque la modulation utilisée est connue du récepteur, on peut utiliser une pondération naturelle basée sur une décision souple sur les bits d'information par le calcul du LLR (Log Likelihood Ratio, [4]) :

$$\Psi(b_i^{(n)}(q)) = \ln \frac{p(b_i^{(n)}(q) = +1/\hat{s}_{i,j}^{(n)})}{p(b_i^{(n)}(q) = -1/\hat{s}_{i,j}^{(n)})}, \quad (7)$$

où $b_i^{(n)}(q)$ est le bit en position q dans le symbole $s_i^{(n)}$. Le bit souple $\tilde{b}_i^{(n)}(q)$ est alors obtenu à partir de (7) par :

$$\tilde{b}_i^{(n)}(q) = \tanh(0.5 \times \Psi(b_i^{(n)}(q))). \quad (8)$$

Quand la valeur absolue du LLR d'un bit est proche de zéro, ce qui signifie que ce bit est peu fiable, le bit souple est également proche de zéro, ce qui réduit l'impact de sa contribution.

Le symbole souple pondéré $\tilde{s}_i^{(n)} = \tilde{x}_i^{(n)} + \jmath \tilde{z}_i^{(n)}$ est obtenu par conversion des bits souples

$$\begin{cases} \tilde{x}_i^{(n)} = \frac{1}{\sqrt{2}} \tilde{b}_i^{(n)}(0) \\ \tilde{z}_i^{(n)} = \frac{1}{\sqrt{2}} \tilde{b}_i^{(n)}(1) \end{cases} \quad (9)$$

dans le cas d'une modulation QPSK et

$$\begin{cases} \tilde{x}_i^{(n)} = \frac{1}{\sqrt{10}} \tilde{b}_i^{(n)}(0)(2 - \tilde{b}_i^{(n)}(1)) \\ \tilde{z}_i^{(n)} = \frac{1}{\sqrt{10}} \tilde{b}_i^{(n)}(2)(2 - \tilde{b}_i^{(n)}(3)), \end{cases} \quad (10)$$

dans le cas d'une modulation 16-QAM.

3.4 Pondération des symboles des codes détectés

Même lorsqu'un code inconnu est détecté, sa modulation reste inconnue et le calcul des LLR n'est pas possible. Pour éviter l'effet ping-pong [12] et la propagation d'erreurs, une pondération explicite est utilisée :

$$\tilde{s}_i^{(n)} = \lambda \hat{s}_i^{(n)}, \quad (11)$$

où $0 < \lambda < 1$.

4 Egaliseur LMMSE

Il est intéressant de comparer le récepteur proposé avec l'égaliseur linéaire MMSE qui ne nécessite pas de détecter les codes inconnus. En effet, il égalise le canal avant désétalement.

Le signal reçu, échantillonné aux multiples du temps chip T_c peut s'écrire sous la forme d'un vecteur, de la manière suivante :

$$\mathbf{r} = \mathbf{H}\mathbf{d} + \mathbf{w}, \tag{12}$$

où $\mathbf{r} = [r_{k-F}, r_{k-F+1}, .., r_{k+F}]^T$, $2F+1$ est la taille du filtre LMMSE, $\mathbf{d} = [d_{k-F-P+1}, d_{k-F-P}, .., d_{k+F}]^T$ contient les chips transmis, $\mathbf{w} = [w_{k-F}, w_{k-F+1}, .., w_{k+F}]^T$ est le vecteur de bruit et $\mathbf{H}$ est la matrice canal, de taille $(2F+1) \times (2F+P)$ donnée par :

$$\mathbf{H} = \begin{pmatrix} \tilde{h}_{P-1} & \tilde{h}_{P-2} & \dots & \tilde{h}_0 & & \mathbf{0} \\ & \tilde{h}_{P-1} & \tilde{h}_{P-2} & \dots & \tilde{h}_0 & \\ & & \ddots & & & \ddots \\ \mathbf{0} & & & \tilde{h}_{P-1} & \tilde{h}_{P-2} & \dots & \tilde{h}_0 \end{pmatrix}, \tag{13}$$

où P est la mémoire du canal et $\tilde{h}_p$ sont les coefficients de la réponse impulsionnelle échantillonnée.

L'égaliseur LMMSE est le filtre $\mathbf{v}_{LMMSE}$ qui minimise l'erreur quadratique moyenne entre sa sortie et le signal émis. Ainsi,

$$\mathbf{v}_{LMMSE} = \arg\min_{\mathbf{V}} E\| \mathbf{v}^H\mathbf{r} - d_k \|^2, \tag{14}$$

et la solution optimale est donnée par

$$\mathbf{v}_{LMMSE} = (\mathbf{H}\mathbf{H}^H + \mathbf{R}_w)^{-1}\mathbf{h}_{P+F} \tag{15}$$

où $\mathbf{R}_w$ est la matrice d'autocorrélation du bruit et $\mathbf{h}_{P+F}$ est la $(P+F)^e$ colonne de $\mathbf{H}$.

Le signal reçu est filtré par $\mathbf{v}_{LMMSE}$ afin de rétablir l'orthogonalité entre les codes des différents utilisateurs. Le signal de sortie est désembrouillé et désétalé afin d'obtenir une estimation des symboles ${s_i}^{(n)}$ pour $n = 0, .., N_K - 1$.

5 Simulations

5.1 Scénarios de simulation

Le schéma de modulation et de codage (MCS) de l'utilisateur considéré est décrit dans le tableau 1. Il correspond au MCS 16 de [1]. Les autres utilisateurs HSDPA utilisent $N_U = 5$ codes et leurs modulations ne sont pas connues. Trois niveaux de puissance sont distingués :

- Les N codes sont transmis avec la même puissance.
- La puissance de l'utilisateur considéré est le double de celle des autres.
- La puissance des codes inconnus est le double de celle de l'utilisateur considéré.

TAB. 1 – Paramètres de simulation

Parametres	Valeurs
Chip rate	3.84Mcps
SF	16
N_K	5
N_U	5
Data Modulation	16 QAM
Frame Length	2ms
Channel estimation	Perfect
Channel Coding /Decoding	Turbo Coding (R=0.3714) /Max-Log-Map decoding (8 iterations)
Channel Model	Vehicular-A 3kmph
Number of received antenna	1

Pour chaque scénario, les performances des récepteurs suivants sont évaluées :

- le récepteur RAKE conventionnel,
- le récepteur PIC à deux étages annulant uniquement l'interférence des codes connus,
- le récepteur PIC à deux étages annulant simultanément l'interférence des codes connus et des codes détectés,
- le récepteur LMMSE avec 33 coefficients,
- le récepteur proposé avec deux étages,
- le récepteur PIC à deux étages annulant l'interférence des codes connus en l'absence des codes inconnus.

5.2 Résultats des simulations

Dans un premier temps, (figure 2), l'utilisateur considéré et les autres ont la même puissance. Tandis que le RAKE est inefficace, le récepteur PIC conventionnel (considérant uniquement les codes connus) atteint un taux de trames erronées de 1% à 24 dB. Puisque ce taux peut être atteint à 17 dB en l'absence d'interférences, il existe une importante marge d'amélioration pour les techniques basées sur la détection et l'annulation d'interférence des codes inconnus. Dans cette situation, le récepteur PIC appliqué simultanément à tous les codes a les mêmes performances que le récepteur LMMSE et atteint un taux de trames erronées de 1% à 20 dB, ce qui représente un gain de 4 dB par rapport au récepteur PIC conventionnel. Le récepteur hybride proposé permet de gagner 2 dB par rapport à ces deux récepteurs.

La figure 3 compare les performances des récepteurs lorsque

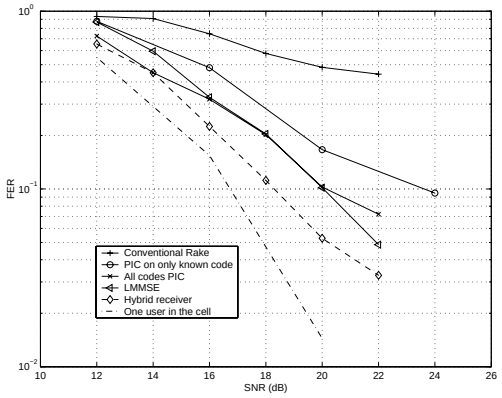


FIG. 2 – Performance des récepteurs : puissances du signal et des interférences équivalentes

la puissance de l'interférence des autres codes est le double de celle de la puissance de l'utilisateur considéré. Dans cette situation, ni le RAKE, ni le PIC conventionnel ne sont capables d'atteindre de bonnes performances. En effet, ils ne peuvent pas supprimer cette forte interférence provenant des autres utilisateurs. Les performances du récepteur PIC appliqué à tous les codes sont sérieusement dégradées par rapport au cas précédent, ce qui montre la limite d'une estimation parallèle des symboles de tous les codes : une erreur dans la détection d'un code ou dans l'estimation d'un symbole se répercute sur tous les étages suivants. Le récepteur LMMSE atteint un taux de trames erronnées de 1% autour de 23 dB et le récepteur hybride

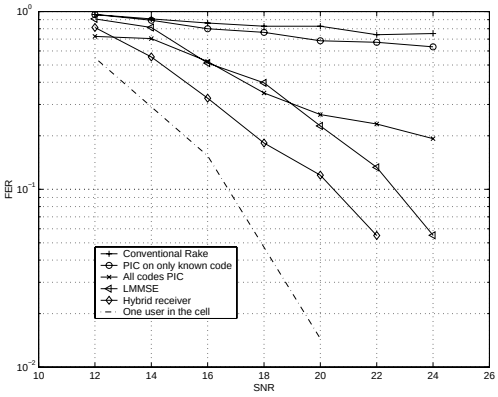


FIG. 3 – Performances des récepteurs : puissance du signal faible

proposé autour de 21 dB, ce qui correspond à un gain de 2 dB.

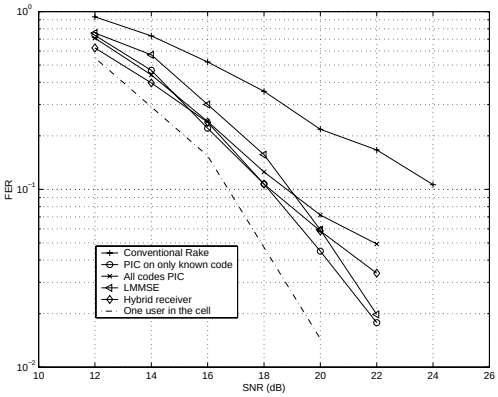


FIG. 4 – Performances des récepteurs : puissance du signal utile élevée

Sur la figure 4, les performances sont comparées lorsque la puissance de l'utilisateur considéré est deux fois plus importante que celle des interférences. Même le récepteur RAKE est capable d'atteindre un taux de trames erronées de 1%. Tous les autres récepteurs sont à peu près équivalents. Le PIC conventionnel est légèrement meilleur que le PIC sur tous les codes. Ceci peut s'expliquer par le fait que dans cette situation, le gain apporté par l'annulation de l'interférence des codes inconnus n'est pas significatif tandis que cette méthode est susceptible d'introduire des erreurs : il est donc parfois préférable d'ignorer les codes inconnus. Le récepteur hybride proposé fait preuve d'une plus grande robustesse vis à vis de ce phénomène puisque ses performances sont équivalentes à celles du PIC conventionnel.

6 Conclusion

Dans ce papier, nous avons présenté un nouveau schéma pour l'annulation d'interférences pour les transmissions HSDPA multi-utilisateurs. L'idée consiste à estimer les symboles des utilisateurs interférents afin de soustraire leur contribution au signal reçu. Comme les codes d'étalement des autres utilisateurs ne sont pas connus, ce récepteur nécessite une étape de détection des codes inconnus. Ce schéma permet d'émettre simultanément vers plusieurs utilisateurs sur la liaison descendante, sans pour autant trop dégrader les performances individuelles de chaque utilisateur, augmentant ainsi la capacité de la cellule.

Références

[1] 3GPP TS 25.214 V5.6.0 : "Physical layer procedures (FDD)", release 5, 09/2003.

[2] 3GPP TS 25.212 V5.2.0 : "Multiplexing and channel coding (FDD)".

[3] 3GPP TS 25.213 V5.2.0 : "Spreading and modulation (FDD)".

[4] T. P. Krauss, M. D. Zoltowski, G. leus, "Simple MMSE Equalizers for CDMA Downlink to Restore Chip Sequence Comparison to Zero-Forcing and Rake", Proc. IEEE ICASSP, Vol. 5 , 5-9 June 2000.

[5] Higuchi, K. Fujiwara, A. M.Sawahashi, "Multipath interference canceller for high-speed packet transmission with adaptive modulation and coding scheme in W-CDMA forward link," IEEE Journal JSAC, Vol. 20 , Issue : 2 , Feb. 2002, Pages :419 - 432.

[6] R. Malik, V. K. Dubey, B.A McGuffin, "hybrid interference canceller for CDMA systems in Rayleigh fading channels" IEEE Proc. VTC 2001 Spring, Vol. 2 , Pages :1523 - 1527.

[7] A. W. Kay, W.A. Krzymieri, "A method of limiting the processing delay of the improved multi-stage variable group hybrid interference cancellation scheme for CDMA systems", IEEE Proc. VTC 1999 - Fall, Vol. 3 , Sept. 1999 Pages :1648 - 1652.

[8] J. Kwun, H. Tu, G. Park, D. Hong, C. Kang, "An MPI Mitigation Technique for High Order Data Modulations in the Presence of Intracell Interference," Proc. of IEEE ICCS, vol. 1, Nov. 2002.

[9] M.F. Madkour, S. C. Gupta, Y.E. Wang, "A Novel Downlink Blind Interference Cancellation in A W-CDMA Mobile Communication System," IEEE, WCNC , Vol. 3 Sept. 1999.

[10] W. Ha, J. Bae, J. Park, "Interference cancellation receiver using groupwise decorrelation for W-CDMA downlink," IEEE Proc. PIMRC, Vol.3, Sept. 2003.

[11] N. Miki, S. Abeta, H. Atrashi, M. Sawahashi, "Multipath Interference Canceller Using Soft-decision Replica Combined with Hybrid ARQ in W-CDMA Forward Link," IEEE Proc. VTC 2001 Fall , Vol. 3 , Oct. 2001.

[12] L.K. Rasmussen, I.J Oppermann, "Ping-pong effects in linear parallel interference cancellation for CDMA," IEEE Trans. on Wireless Comm., Vol. 2 , Issue : 2 , March 2003 Pages :357 - 363.

Construction de filtres LPTV orthogonaux pour un système d'accès multiple par étalement spectral Etude des effets d'une désynchronisation entre utilisateurs

W. CHAUVET[1], B. LACAZE[1], D. ROVIRAS[1] , A. DUVERDIER[2]

[1]ENSEEIHT/IRIT, 2 Rue Camichel, 31071 Toulouse Cedex 7, BP 7122 France

[2]CNES, 18 Avenue E. Belin, 31041 Toulouse Cedex 4 France

wilfried.chauvet@enseeiht.fr

Résumé – Cet article propose une méthode de construction d'une famille de filtres **LPTV** (Linear Periodical Time Varying) orthogonaux pour une application à un système d'accès multiple par étalement. Cette méthode repose sur l'utilisation d'un filtre LPTV générateur. Lorsque les utilisateurs sont synchronisés, le système obtenu est parfaitement orthogonal. L'introduction d'une désynchronisation entre ces utilisateurs a pour effet de dégrader cette orthogonalité en introduisant de l'interférence multiutilisateurs. Nous proposons d'utiliser un entrelaceur matriciel comme filtre LPTV générateur. Pour un tel choix, une étude théorique de l'effet d'une désynchronisation est menée. Des simulations sont ensuite réalisées pour quantifier l'interférence multiutilisateurs introduite par la désynchronisation des différents utilisateurs.

Abstract *– This paper focuses on the design of a set of orthogonal Linear Periodic Time Varying* ***(LPTV)*** *filters and proposes to apply this result to spread spectrum multiuser transmissions. The construction method of such an orthogonal set is presented, emphasizing that only an invertible LPTV filter is required, since the full set is deduced from this kernel element. Synchronisation of all the users results in perfect ortohogonality of the system . But in case of desynchronisation, multiuser interferences occur. As a kernel LPTV filter, we propose to use a matrix interleaver. Effects of desynchronisation are theoretically considered and simulations are performed to evaluate multiuser interferences.*

1. Introduction

Les applications des filtres linéaires périodiques variant dans le temps **(LPTV** pour Linear Periodical Time Varying) en traitement du signal sont nombreuses. Certaines utilisent ces filtres explicitement : transmultiplexage [6], estimation aveugle de canal [2] ou entrelacement fréquentiel [3]. D'autres applications utilisent les filtres LPTV de façon implicite : entrelacement temporel (nature LPTV des entrelaceurs) ou traitement multicadence (nature LPTV de l'interpolation et de la décimation [7]).

Les propriétés d'étalement spectral des filtres LPTV font de ces filtres un outil original pour réaliser un système d'accès multiple par étalement spectral et ont suscité récemment un certain intérêt [4]. Une telle application nécessite d'être en mesure de proposer une famille de filtres LPTV inversibles et orthogonaux. Dans [4], l'inversibilité est assurée par le fait que les filtres LPTV utilisés sont des entrelaceurs blocs (c'est à dire des permutations) qui constituent un sous ensemble particulier de l'ensemble des filtres LPTV. Concernant l'orthogonalité, [4] propose de générer un entrelaceur aléatoirement pour chaque utilisateur.

Dans [1], nous avons proposé une méthode générale pour construire un ensemble de filtres LPTV orthogonaux. Une particularité de cette méthode réside dans le fait qu'elle nécessite seulement la connaissance d'un filtre LPTV générateur inversible.

Dans un contexte d'accès multiple, nous y avons proposé empiriquement [1] l'utilisation de l'entrelaceur matriciel (lignes/colonnes) comme filtre LPTV générateur. Lorsque tous les utilisateurs sont synchronisés, l'orthogonalité du système est parfaite et l'interférence multiutilisateurs est nulle. Dans le cas où une désynchronisation entre les utilisateurs apparaît, l'orthogonalité est perdue ce qui a pour effet d'introduire de l'interférence multiutilisateurs. Dans ce présent papier, nous proposons d'étudier théoriquement l'effet d'une désynchronisation sur le système d'accès multiple construit à partir de l'entrelaceur matriciel puis d'évaluer par simulation l'interférence multiutilisateurs introduite par la désynchronisation des différents utilisateurs.

2. Construction d'un ensemble de filtres LPTV orthogonaux

2.1 Rappel sur les filtres LPTV

Un filtre LPTV est un filtre dont la réponse impulsionnelle est une fonction périodique du temps. Plusieurs représentations, résumées dans [5], sont possibles. Parmi celles ci, une représentation utile pour la suite est la représentation modulateurs (figure 1) pour laquelle un filtre LPTV N périodique est défini par N filtres invariants dans le temps (**LIT**) $T_p(z)$ appelés filtre modulateurs.

2.2 Construction d'une famille de filtres LPTV orthogonaux

Nous considérons $T_p^{(0)}(z)$ $(0 \le p \le N-1)$ les N filtres modulateurs d'un filtre LPTV N périodique que nous appelons filtre LPTV générateur. Nous définissons les N filtres LPTV (pour $0 \le k_0 \le N-1$) par les filtres modulateurs $T_p^{(k0)}(z)$ où $[p]_N$ désigne le reste de la division entière de p par N (1). Nous montrons alors dans [1] que ces N filtres sont inversibles si le filtre LPTV générateur est inversible.

$$T_p^{(k_0)}(z) = T_{[p-k0]_N}^{(0)}(zW_N^{-k_0}) \quad où \quad W_N = \exp(\frac{-2i\pi}{N}) \qquad (1)$$

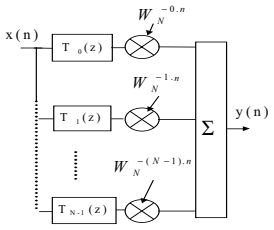


FIG. 1 : Représentation modulateurs d'un filtre LPTV

2.3 Application à un système d'accès multiple

Nous considérons le système multiutilisateurs de la figure 2. A chaque utilisateur i est associée la sortie $u_i(n)$ d'une modulation M-aire de même débit symbole D_S. En émission, chacun de ces signaux $u_i(n)$ est suréchantillonné d'un facteur L par répétition L fois de chaque symbole (non par insertion de L-1 zéros). Ce signal est ensuite filtré par un filtre en Racine de Cosinus Surélevé (**RCS**) de facteur d'arrondi α. Nous considérons alors un filtre LPTV générateur de période N. Nous montrons [1] que parmi les N filtres LPTV inversibles construits précédemment (1), il est possible d'extraire γ de ces N filtres tels que le système de la figure 2 constitue un système multiutilisateurs parfaitement orthogonal en l'absence de désynchronisation entre les utilisateurs. Si nous définissons le paramètre $\beta = \text{int}(N(1+\alpha)/L)+1$, le nombre maximal de filtres LPTV orthogonaux constructibles est donné par $\gamma = \text{int}(N/\beta)$. Pour N grand, nous avons $\gamma \approx L/(1+\alpha)$. Sur la figure 2, $LPTV_i$ est le filtre LPTV de filtres modulateurs $T_p^{(i\beta)}(z)$ définis par (1) à partir du filtre LPTV générateur et $LPTV_i^{-1}$ est le filtre LPTV inverse associé.

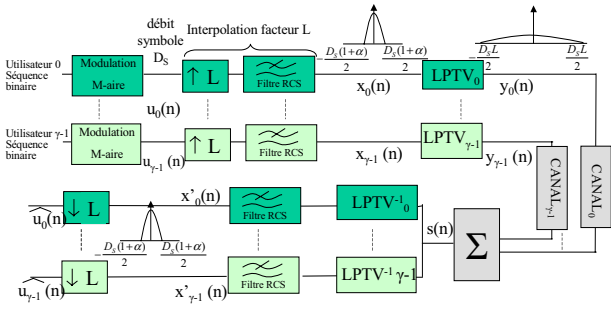


FIG. 2 : Système multiutilisateurs par filtres LPTV

Il est utile de préciser que le problème de la mise en forme analogique des signaux $y_i(n)$ (figure 2) n'est pas discutée ici. Une solution est le filtrage de ces signaux par filtre RCS. Le signal reçu s(n) devra alors être filtré par un filtre adapté RCS. Le point à discuter est alors en réception l'influence d'une imprécision de l'instant d'échantillonnage (lors de la conversion analogique/numérique) en sortie de ce filtre de réception.

3. Etude de la perte d'orthogonalité par désynchronisation des utilisateurs

3.1 Effet d'une désynchronisation sur le système

Le système de la figure 2 définit un système d'accès multiple parfaitement orthogonal lorsque les différents utilisateurs sont tous synchronisés en réception. Lorsque les utilisateurs sont désynchronisés, l'orthogonalité est dégradée. Nous modélisons pour l'utilisateur i une désynchronisation par un canal de type $canal_i = z^{ni}$. Pour cet utilisateur particulier, la cascade du filtre LPTV d'émission, du canal et du filtre LPTV inverse de réception est équivalent à un nouveau filtre LPTV que nous notons $LPTV_i^{ni}$ (figure 3). En l'absence de désynchronisation ($n_i=0$), ce filtre équivalent vaut l'identité. En cas de désynchronisation, ce filtre a une nature variant dans le temps.

Ainsi, si nous considérons un utilisateur i synchronisé en réception, la présence d'une désynchronisation pour les autres utilisateurs j ($j \neq i$) a pour effet de produire de l'interférence multiutilisateurs qui vient bruiter le signal reçu $x_i'(n)$ (figure 2) avant la décision sur l'utilisateur i. Nous montrons [8] que l'interférence multiutilisateurs induite par l'utilisateur j ($j \neq i$) est d'autant plus importante que le filtre $LPTV_i^{nj}$ a une forte nature variant dans le temps. Nous montrons [1] aussi qu'il n'est pas nécessaire d'étudier la sensibilité de chacun des filtres orthogonaux à une désynchronisation, mais qu'il suffit d'étudier la sensibilité du filtre générateur $LPTV_0$ à une désynchronisation.

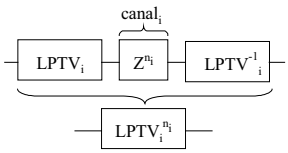


FIG. 3 : Modélisation d'une désynchronisation

3.2 Choix d'un filtre LPTV générateur : l'entrelaceur matriciel

Le choix d'un filtre LPTV générateur quelconque conduit à une sensibilité de ce filtre à la désynchronisation très grande, quelque soit la désynchronisation n_0 non nulle. Le choix du filtre LPTV générateur est donc critique. Nous proposons de choisir notre filtre générateur dans la famille des changements d'horloge (**PCC** pour Periodical Clock Change). Un changement d'horloge N périodique est un filtre LPTV particulier défini par une fonction f(n), de $\mathbb{Z}$ dans $\mathbb{Z}$, de période N, telle que la sortie y(n) est obtenue en fonction de l'entrée x(n) par la relation (2).

$$y(n) = x(n - f(n)) \quad (2)$$

Il est possible de montrer [8] que l'ensemble des entrelaceurs périodiques est confondu avec l'ensemble des PCC numériques. Nous choisissons ici comme filtre LPTV générateur l'entrelaceur matriciel (Q lignes/P colonnes). Cet entrelaceur (de période N=PQ) consiste à écrire les échantillons d'entrée colonne par colonne et à les lire en sortie ligne par ligne (figure 4).

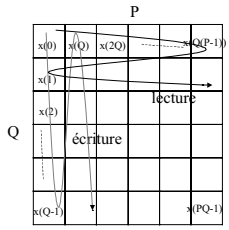


FIG. 4 : Principe de l'entrelaceur matriciel

On montre [8] que la fonction PCC f(n) associée à l'entrelaceur matriciel a l'expression analytique (3). Un exemple de cette fonction est illustré sur la figure 5.

$$f(n) = -(Q-1)c_{P,Q}(n) + (P-1)b_{P,Q}(n) \quad (3)$$

$$avec \quad c_{P,Q}(n) = [n]_P \quad et \quad b_{P,Q}(n) = \left\lfloor (n - c_{P,Q}(n))/P \right\rfloor_Q$$

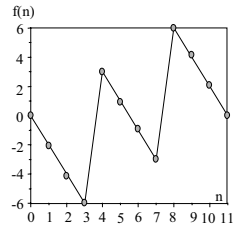


FIG. 5 : Fonction PCC de l'entrelaceur matriciel (Q=3,P=4)

Ce choix particulier d'entrelaceur matriciel comme filtre LPTV générateur résulte finalement en un système assez simple. Il est utile ici de justifier pourquoi nous avons pris le parti d'une introduction très théorique et générale de ce système dans la partie 2. La généralisation de cette partie 2 par un formalisme LPTV a l'intérêt de montrer d'une part qu'un tel système est possible à partir de n'importe quel filtre LPTV générateur et sous entend que la discussion sur un autre choix de filtre LPTV générateur reste ouverte. D'autre part, cette présentation LPTV nous offre un cadre théorique permettant une modélisation intéressante (figure 3) de l'effet d'une désynchronisation. Le point suivant va mettre en évidence l'intérêt de cette modélisation puisque l'étude de l'effet d'une désynchronisation résulte simplement en la détermination théorique d'un PCC équivalent ($LPTV_{n0}$ sur la figure 3).

3.3 Sensibilité de l'entrelaceur matriciel à une désynchronisation

Il s'agit d'évaluer l'expression théorique de $LPTV_0^{n0}$ (figure 3) dans le cas où $LPTV_0$ est un PCC de fonction (3) et où $LPTV_0^{-1}$ est son PCC inverse. Nous montrons alors que $LPTV_0^{n0}$ est un changement d'horloge de fonction PCC $h^{n0}(n)$ (4) où les fonctions $\theta'_{n0}(n)$ et $\theta_{n0}(n)$ sont des fonctions de n.

$$h^{n0}(n) = f(n_0) + \theta_{n0}(n)(PQ-1) - \theta'_{n0}(n)(P-1)Q \quad (4)$$

Un changement d'horloge a une nature variant dans le temps d'autant plus faible que la fonction PCC associée est proche d'une fonction constante (dans quel cas, le PCC est un simple retard (2) et se trouve être invariant dans le temps). Dans notre cas, la fonction $h^{n0}(n)$ (4) ne peut prendre que 4 valeurs puisque l'on montre que les deux fonctions $\theta'_{n0}(n)$ et $\theta_{n0}(n)$ ne prennent que deux valeurs 0 ou 1 en fonction de n conformément à la figure 6.

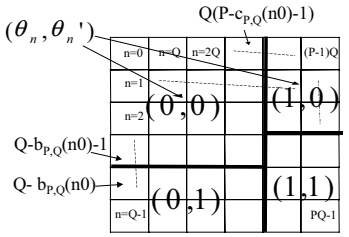


FIG. 6 : Distribution des valeurs de $\theta'_{n0}(n)$ et $\theta_{n0}(n)$

La faible dépendance temporelle du PCC équivalent (figure 3) indique que le système multiutilisateurs (figure 2) est peu sensible à une désynchronisation. Puisque cette sensibilité est à l'origine de l'interférence multiutilisateurs, nous pouvons prévoir une faible interférence multiutilisateurs de notre système dans le cas où les utilisateurs sont désynchronisés. Nous proposons de vérifier à présent ce résultat par des simulations.

4. Quantification de l'interférence multiutilisateurs par simulation

Nous proposons d'illustrer dans un premier temps les capacités d'étalement du système (figure 2) reposant sur l'entrelaceur matriciel. Les paramètres choisis sont α =0.4, L=8, P=16 et Q=12. Cela nous donne un nombre d'utilisateurs maximal γ =5. La figure 7 représente la Densité Spectrale de Puissance (**DSP**) de l'utilisateur i=4 en entrée et en sortie du filtre LPTV d'émission. La figure 8 représente la DSP du signal total s(n) (figure 2). On vérifie bien que le signal total occupe l'intégralité de la bande offerte.

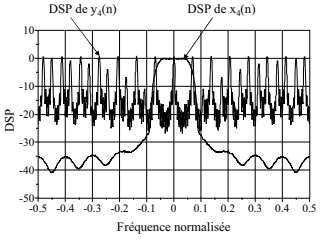


FIG. 7 : Simulation de la DSP après étalement pour un utilisateur

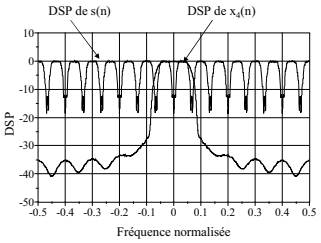


FIG. 8 : Simulation de la DSP après étalement pour l'ensemble des utilisateurs

Afin de quantifier l'interférence multiutilisateurs du système, nous proposons de simuler les performances du système en terme de Taux d'Erreur Binaire (**TEB**) en fonction du nombre d'utilisateurs. Pour cela, nous supposons que l'utilisateur considéré est synchronisé et que les autres utilisateurs sont désynchronisés de façon aléatoire selon une loi uniforme dans l'intervalle [0;N-1]. Les paramètres choisis sont L=31, P=17, Q=188.

Nous proposons de comparer ces performances à celle d'un accès multiple par DS-CDMA (Direct Sequence Code Division Multiple Access) avec séquence de Gold. Le même facteur de suréchantillonnage L=31 est choisi et les deux polynômes générateurs sont $[45,75]_{oct}$. L'utilisateur considéré est synchronisé et les autres utilisateurs sont désynchronisés avec une désynchronisation aléatoire (loi uniforme dans l'intervalle [0;L-1]).

Les performances en TEB obtenues sont présentées sur la figure 9. Rappelons qu'aucun bruit n'est ajouté sur le canal. Les dégradations du TEB ne sont dues qu'à l'interférence multiutilisateurs. Ces résultats nous permettent de vérifier que le système proposé présente une interférence multiutilisateurs faible comme le laissait prévoir l'étude théorique précédente de la désynchronisation.

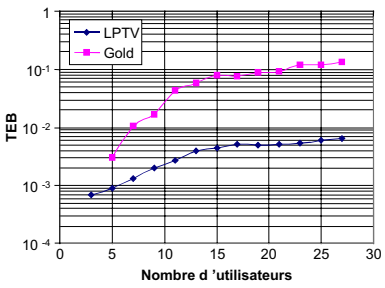


FIG. 9 : Simulation du TEB en fonction du nombre d'utilisateurs

Il est utile ici de justifier le choix d'une comparaison de notre système avec un code des séquences d'étalement qui sont des codes de Gold. En effet, nous aurions aussi pu choisir des codes de Walsh Hadamard. La raison est simple : nous modélisons ici une liaison où les désynchronisations prennent toutes les valeurs équiprobablement entre 0 et L-1 (L est la longueur du code). Cela signifie que nous nous intéressons pas ici seulement à de très faibles désynchronisations (d'une durée de l'ordre de quelques chips du code) mais au cas de désynchronisations équiprobables. Sous une telle hypothèse, l'interférence multiutilisateurs induite pour des codes de Walsh Hadamard serait encore plus handicapantes que pour les codes de Gold car la forme des fonctions d'intercorrélation des codes de Walsh Hadamard implique que ces codes sont bien plus sensibles aux fortes désynchronisations que les codes de Gold.

5. Conclusion

Nous avons d'abord rappelé notre méthode [1] pour construire un système d'accès multiple par étalement à partir d'un filtre LPTV générateur. Ce système présente une orthogonalité parfaite dans le cas où tous les utilisateurs sont parfaitement synchronisés. Le présent papier a discuté l'effet d'une désynchronisation sur les différents utilisateurs et a mis en évidence l'importance du choix du filtre LPTV générateur. Nous avons étudié le cas de l'entrelaceur matriciel (modélisé comme un PCC) et avons mis en évidence la faible sensibilité de ce PCC face à une désynchronisation. L'interférence multiutilisateurs du système est liée à cette sensibilité et des simulations nous ont permis de vérifier que cette interférence était faible. Le système ainsi construit présente une faible interférence mutiutilisateur quelque soit le type de liaison (liaison faiblement ou fortement désynchronisée). Le deuxième point important qui caractérise un système multiutilisateurs est son comportement dans le cas de canaux sélectifs en fréquence (présence d'Interférence Inter Symbole). Ce point est actuellement en cours d'étude.

Références

[1] W.Chauvet, B.Lacaze, D.Roviras, A. Duverdier, « Design of orthogonal LPTV filters: Application to spread spectrum multiple access ». , ICASSP, Montreal, 2004.

[2] K. Tsatsanis et G.B Giannakis, «Transmitter induced cyclostationnarity for blind channel equalization», IEEE Trans. Signal Process., Vol 45, No 7, July 1997.

[3] R. Ishii et M.Kakishita, «A design method for periodically time varying digital filter for spectrum scrambling», IEEE Tans. Acoust. Speech Signal Process., 1990, Vol 38, pp 1219-1222.

[4] D.Roviras, B.Lacaze, N.Thomas, «Effects of discrete LPTV on stationnary signals», ICASSP, Orlando, 2002.

[5] D.Mc Lernon, «One dimensional LPTV structures : derivations, interrelationships and properties», IEE Proc Image Signal Process., Vol 149, No 5, October 1999.

[6] J.S Prater et C.M Loeffler, «Analysis and design of periodically time varying IIR\ filters with aplication to transmultiplexing», IEE Trans. Acoust. Speech Signal Process., 1992, Vol 40, pp. 2715-2725.}

[7] R. Ansari et Bede Liu, «Interpolators and decimators as periodically time varying filters», in Proc. IEEE Int. Symp. Circuits and Systems, 1981, pp. 447-450.

[8] W. Chauvet, «Etude des filtres LPTV numériques : applications aux communications numériques», thèse INP Toulouse, 2004.

HISD : un nouveau décodeur MIMO utilisant une approche géométrique pour des transmissions MAQ

A. NAFKHA, E. BOUTILLON, C. ROLAND

LESTER, Centre de recherche, Université de Bretagne Sud, BP 92116, 56321 Lorient Cedex, France
{emmanuel.boutillon,christian.roland,armor.nafkha}@univ-ubs.fr

Résumé – Dans de nombreux systèmes de communication, l'opération de décodage se résume à résoudre un problème d'optimisation d'un système combinatoire de la forme $y = Hx+b$. Dans le cas général, son décodage au sens du maximum de vraisemblance (ML) est un problème connu de type « NP-Complet ». Dans cet article, afin de résoudre ce problème, un nouvel algorithme sous optimal appelé HISD (Hyperplane Intersection and Selection Detector) est proposé. Cet algorithme repose sur une approche géométrique. Comparé aux algorithmes existants, le HISD possède trois caractéristiques très attrayantes pour une implantation dans des systèmes réels. La première concerne ses performances en terme de TEB qui sont proches de celles de l'optimum (ML). Sa complexité en terme de calcul est faible et sa structure est intrinsèquement parallèle ce qui rend d'autant plus aisée son implantation matérielle. Dans cet article, l'algorithme HISD est utilisé pour résoudre de manière efficace le problème du décodage des systèmes de transmissions MIMO (Multiple Input Multiple Output).

Abstract *– In many efficient wireless communication systems, the demodulation tasks resume to solve a combinatorial optimization problem of the form $y = Hx+b$ where H is a real m×n channel matrix, b is a noise vector, y is the received vector and x is a discrete n-dimensional vector of transmitted information. The Maximum Likelihood (ML) detection of $y = Hx+b$ is known to be a NP-hard problem in the general case. In this paper, a new sub-optimal algorithm for ML detection problem is proposed. This algorithm called HISD (Hyperplane Intersection and Selection Detector) is based on a geometrical approach. Comparing to existing algorithms, HISD has three characteristics that make it very attractive for practical systems. First, it performs very close to the ML limit. Second, it has a low computational complexity. Finally, The inherent parallel structure of the proposed algorithm allows us highly suitable hardware implementation. In this paper the HISD is efficiently employed in the case of MIMO (Multiple Input Multiple Output) systems.*

1. Introduction

L'augmentation de l'efficacité spectrale reste l'un des grands enjeux des communications numériques sans fils. Les transmissions multi-antennes constitue une solution pour améliorer la qualité de la transmission sur des canaux difficiles.

Lors d'une transmission avec un système MIMO (Multiple Input Multiple Output) le maximum de débit transmissible est atteint si on émet à chaque temps symbole un symbole différent sur chaque antenne. Le nombre de symbole émis simultanément est donc égal au nombre N d'antenne émission, c'est ensemble sera représenté dans le reste de l'article par le vecteur du signal bande de base émis $\tilde{x}$.

L'étude portera sur une communication mono utilisateur dans un canal non sélectif en fréquence. Le modèle de canal équivalent en bande de base est représenté par une matrice H, dont chaque coefficient $\tilde{H}_{ij}$ représente la fonction de transfert entre la $j^{ème}$ antenne d'émission et la $i^{ème}$ antenne de réception. Dans ces conditions, le vecteur y du signal bande de base reçu a pour équation :

$$\tilde{y} = \tilde{H}\tilde{x} + \tilde{b} \qquad (1)$$

Dans le cas de transmission numérique avec un codage binaire à symbole de type MAQ les éléments de l'équation (1) sont des complexes. Il est utile pour simplifier le modèle de passer en coordonnées réelles grâce à l'équation (2)

$$\begin{pmatrix}\mathrm{Re}(y)\\ \mathrm{Im}(y)\end{pmatrix} = \begin{pmatrix}\mathrm{Re}(H) & -\mathrm{Im}(H)\\ \mathrm{Im}(H) & \mathrm{Re}(H)\end{pmatrix}\begin{pmatrix}\mathrm{Re}(x)\\ \mathrm{Im}(x)\end{pmatrix} + \begin{pmatrix}\mathrm{Re}(b)\\ \mathrm{Im}(b)\end{pmatrix} \qquad (2)$$

Sous cette condition, l'équation 1 devient $y = Hx + b$.

$x = \{\mathrm{Re}(x_1),\cdots,\mathrm{Re}(x_N),\mathrm{Im}(x_1),\cdots,\mathrm{Im}(x_N)\}^T$ est le vecteur des symboles émis décomposés en partie réel et partie imaginaire, $H = \{h_{ij}\}$ représente la matrice de canal de dimension $m \times n$ ($n=2N$ et $m=2M$ avec N et M respectivement le nombre d'antenne en émission et en réception) et $b = \{\mathrm{Re}(b_1),\cdots,\mathrm{Re}(b_N),\mathrm{Im}(b_1),\cdots,\mathrm{Im}(b_N)\}^T$ le vecteur de bruit gaussien additif.

Concernant le modèle de canal, les hypothèses classiques de stationnarité durant la transmission d'un « bloc » de données contenant plusieurs vecteurs x et de connaissance parfaite par le récepteur de la matrice de canal H sont adoptées.

La solution optimale x_{ML} correspond au minimum d'une fonction de coût définie entre le vecteur du signal reçu y et les éléments x de l'ensemble ξ des solutions possibles tel que :

$$x_{ML} \equiv \mathrm{argmin}\left\|y - Hx\right\|^2 \qquad (3)$$

Dans cette équation, un vecteur $x \in \xi = \{-1,+1\}^n$ pour une MAQ4.

Lorsque n est petit une recherche exhaustive est possible mais pour des valeurs plus importantes les calculs

deviennent rapidement prohibitifs. (par exemple pour n=8 soit 4 antennes en émission, le nombre de calcul est de 64 pour chaque vecteur y du signal reçu). En effet, la complexité en terme de multiplications du décodeur optimal (ML) croit rapidement en fonction du nombre N d'antenne en émission $\mathcal{O}((2N)^{2N})$. Afin de pallier ce problème de nombreux décodeurs ont été proposés.

Le SD (Sphère Décoder) [2] fait partie des méthodes optimales mais sa structure n'est pas régulière et ne permet pas une implantation matérielle optimisée.

Des algorithmes sous optimaux sont aussi proposés dans la littérature comme le décorrélateur ou Zéro-Forcing (ZF) et le Minimum de l'Erreur Quadratique Moyenne (MMSE) [4]. Des solutions à base de détecteurs non linéaires basés sur la soustraction d'interférences multi-utilisateur (SIC ou PIC) ou sur le ZF à retour de décision ont également fait l'objet d'études [7].

Récemment une nouvelle approche [6] appelée SemiDefinite Programming (SDP) a montré quelle pouvait être une proposition intéressante à la résolution du problème posé.

L'algorithme sous optimal HISD (Hyperplane Intersection and Selection Decoder) est proposé, dans cet article, pour résoudre le problème du décodage de transmissions MIMO.

La section n°2 détaille l'algorithme de décodage HISD pour des transmissions de signaux modulées en MAQ4. Celle ci est divisée en deux parties qui sont les phases d'intensification et de diversification. Dans la section suivante, la complexité de calcul du HISD est étudiée. Enfin, la dernière section quand à elle présentera les résultats du HISD pour des transmissions MAQ avec un code MIMO de rendement plein. L'algorithme est comparé en terme de complexité et de résultat au décodeur optimal SD d'une part et à des algorithmes sous optimaux de type MMSE et PIC.

2. Description de l'algorithme HISD

L'algorithme HISD fait partie des méthodes sous optimales du ML (Maximum Likelihood) qui consistent à diminuer la taille de l'ensemble des solutions ξ en sélectionnant un sous ensemble de points ξ_{start}.

L'algorithme HISD est une nouvelle méthode de sélection de ξ_{start}. Il s'agit de trouver des solutions proches de la solution optimale à l'aide d'une approche géométrique tirée des travaux particulièrement originaux de H. Artes, D. Seethaler et F. Hawatsch [5]. Le HISD utilise une amélioration de cette approche associée à une méthodologie d'optimisation de processus qui contient deux stratégies complémentaires. L'approche méthodologique du HISD consiste à voir la résolution du problème de décodage posé comme un problème de Recherche opérationnelle. En conséquence nous l'algorithme est composé de deux phases, à savoir « l'intensification » et « la diversification ». L'amélioration de l'approche géométrique concerne la possibilité de calcul avec des éléments réels plutôt que complexes, ce qui facilité grandement l'implantation de l'algorithme. La phase de diversification consiste à rechercher un ensemble ξ_{start} de points susceptibles de parvenir dans la phase d'intensification au point correspondant à la solution optimale.

L'algorithme HIS utilisent deux paramètres qui sont gérés dans la phase de diversification. Le premier D est le nombre de vecteurs singuliers obtenus à partir de la décomposition de la matrice de canal H et le second C représente le nombre d'éléments de ξ_{start} candidats à la phase d'intensification.

2.1 Phase d'intensification

La fonction d'intensification $gr()$ est basée sur l'itération de la minimisation de la fonction de coût. Cette minimisation est obtenue en calculant les gradients entre un élément de ξ_{start} et ses plus proches voisins. Soit x_t un point de ξ_{start} et de $\mathcal{N}(x_t)$ sous ensemble de ξ qui contient les points situés à une distance de Hamming de 1 de x_t. Par convention on notera $x_t^p \in \mathcal{N}(x_t)$ le point résultant de x_t par l'inversion de signe de la $p^{\text{ème}}$ coordonnée. Soit $x_t^p(p) = -x_t(p)$ pour une MAQ4. À partir d'un point obtenu dans la phase de diversification, nous calculons le meilleur des voisins en utilisant la formule suivante :

$$x_{t+1} = \arg\left\{ \min_{x \in \mathcal{N}(x_t)} f_y(x) \right\} \quad (4)$$

Pour passer d'une itération à la suivante la fonction de coût f_y doit décroître ($f_y(x_{t+1}) < f_y(x_t)$) sinon l'algorithme s'arrête et retourne x_t. L'implantation exhaustive de l'équation (4) comporte n calculs de $f_y(x_t^p)$ qui peuvent être remplacer par le calcul suivant :

$$\delta(x_t^p, x_t) = \tfrac{1}{4}\left(f_y(x_t^p) - f_y(x_t) \right) \quad (5)$$

$$\delta(x_t^p, x_t) = G(p,p) + \eta\left[x_t{}^T G(:,p) - y^T H(:,p) \right] \quad (6)$$

avec la matrice de Gram $G = H^T H$, $\eta = -\text{sign}(x_t(p))$, $G(:,p)$ et $H(:,p)$ représentant respectivement la $p^{\text{ème}}$ colonne des matrices de Gram et de canal. La matrice G est calculée seulement lors de la remise à jour des coefficients de la matrice canal suite à une phase d'estimation. Ce calcul est réalisé à chaque nouveau « bloc » de symbole.

La phase de minimisation de la fonction de coût ainsi effectuée est itérée jusqu'à ce que la solution trouvée soit un minimum. Pour une implantation matérielle d'une MAQ4, le nombre d'itération θ est limité à 2 car le gain obtenu ensuite est pratiquement nul. Le problème est que ce minimum n'est pas pour autant l'optimum, ceci est du à la présence de minima locaux dans l'hyperespace des solutions ξ.

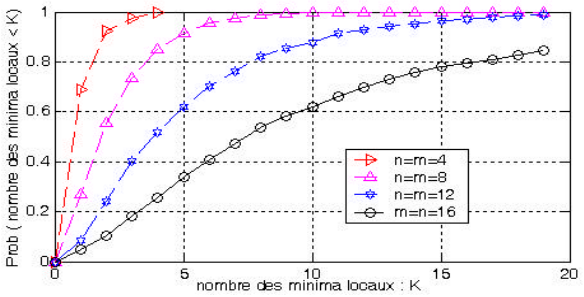


FIG. 1 : Probabilités d'apparition des minima locaux

Dans l'exemple de la figure 1, le rapport Signal à Bruit est de 10 dB et $n = m = 16$. Les probabilités d'avoir une

valeur supérieure à 6 ou à 12 minima est de respectivement de 60% et 30% (cf. fig.1).

Plus le nombre d'antenne augmente plus ces probabilités deviennent importantes. De même, le nombre de minima augmente en fonction du nombre de points de la constellation.

2.2 Phase de diversification

La phase de diversification doit donc permettre de trouver autour du point optimal un ensemble ξ_{start} de points candidats. Ceux ci doivent être le moins redondant possible de façon à couvrir un espace suffisant renfermant le point optimal et de manière à ce que la phase d'intensification évite l'écueil des minima locaux.

Pour être efficace la méthode doit fournir un sous ensemble ξ_{start} de ξ qui contient les points candidats vérifiant l'intersection :

$$\xi_{start} \cap \prod \neq \varnothing \text{ où } \forall (x,y) \in \prod \quad gr(x) = gr(y) \qquad (7)$$

La fonction $gr()$ est la fonction itérative d'intensification et Π est l'ensemble des points qui convergent vers la solution optimale dans la phase d'intensification.

L'idée de base repose sur la recherche de solutions géométriques au problème combinatoire posé. Cela consiste dans un premier temps à obtenir une solution sous optimale ρ_{zf} de type zéro forcing (ZF) ou (MMSE). À chaque fois que la matrice de canal est estimée une décomposition en valeurs singulières (SVD) est réalisée : $H = U\Sigma V^T$. Où les valeurs singulières $\{\lambda_k\}_{k=1}^n$ sont les composantes de la matrice diagonale Σ. Ces composantes sont classées suivant un ordre croissant. Les vecteurs singuliers $\{u_k\}_{k=1}^n$ et $\{v_k\}_{k=1}^n$ qui proviennent respectivement des matrices unitaires U et V^T. Les vecteurs v_k sont axes principaux des ellipsoïdes d'équipotentiel de bruit définies par la fonction $f_y(a)$= C^{ste}.

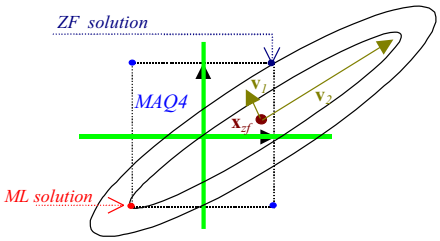


FIG. 2 : Projection à partir du ZF suivant une direction dominante de bruit avec une modulation MAQ4

Pour tout point a de ξ, $a-\rho_{zf}$ est un vecteur qui est solution de l'équation $a - \rho_{zf} = \sum_{k=1}^{n} \alpha_k v_k$ où les α_k sont des coefficients réels. Il est ainsi possible de réécrire la fonction d'intensification par rapport au point candidat a :

$$f_y(a) = \| y - Ha \|_2^2 = \| \sum_{k=1}^{n} \alpha_k \lambda_k v_k \|_2^2 = \sum_{k=1}^{n} \alpha_k^2 \lambda_k^2 \qquad (8)$$

Cette égalité est obtenue grâce au fait que les vecteurs v_k forment une base orthonormée de l'espace.

Considérons les droites Δ_k définies par :

$$\Delta_k = \left\{ z \in \Re^n / z = \rho_{zf} + \alpha v_k, \alpha \in \Re \right\} \qquad (9)$$

Ces droites Δ_k passent par le point ZF $\rho_{zf} = G^{-1}H^T y$ et ont pour direction les vecteurs singuliers v_k (cf. fig. 2). Il faut noter que ces axes sont d'autant plus importants que les valeurs respectives de λ_k sont faibles, donc que k est faible.

Au maximum les D plus petites valeurs des v_k sont retenues (en général de 1 à 4). Le fait de prendre D droites Δ_k va générer un premier degré de diversification. Un second degré de diversification est obtenu grâce aux intersections (cf. fig. 3) entre les droites Δ_k et les plans de l'hypercube des solutions de l'ensemble ξ.

En effet, à partir d'une droite Δ_k on obtient $2n$ intersections avec les hyperplans $P(i,s)$ définis de la manière suivante :

$$P(i,s) = \left\{ z \in \Re^n / z(i) = s \right\} \qquad (10)$$

i varie de 1 à n puisqu'il existe autant d'hyperplan que de coordonnées d'un vecteur x et s vaut 1 ou -1 .

L'ensemble $\mathcal{I}_k$ des solutions doit vérifier pour chaque coordonnées d'un de ses éléments l'équation suivante.

$$\rho_{zf}(i) + \alpha_k^{s,i} v_k(i) = s \qquad (11)$$

Sur l'exemple à 2 dimensions (n=2) de la figure 3 $\mathcal{I}_k$ correspond à l'ensemble : $\{s_1, s_2\}$.

Ensuite, une stratégie de choix des meilleurs candidats est mise en place de façon à limiter le nombre d'élément de ξ_{start} et par la même le nombre de calcul.

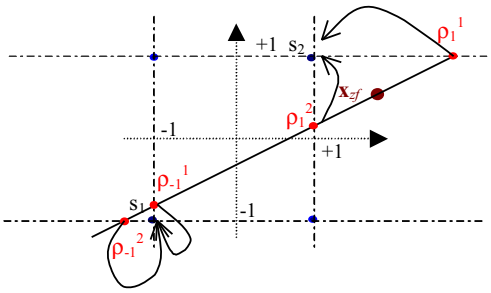


FIG. 3 : Première étape du choix des candidats potentiels

2.3 Description de l'algorithme HISD :

Un composant matériel est défini par ses entrées, ses sorties, ses paramètres et le traitement proprement dit. Nous découpons en 2 parties l'opération de traitement : le prétraitement (qui est lié à l'évolution du canal) et les différentes phases opératoires qui seront intégrées sur un composant hardware de type FPGA.

Les entrées:

Le vecteur y du signal reçu et la matrice de canal.

La sortie

Une solution proche de l'optimal x_{hisd}

Le traitement

Pré-traitement : chercher les D vecteurs singuliers associés aux valeurs singulières de la matrice de canal H (décomposition en valeur singulières SVD).

Phase 1 : Calcul de $\rho_{zf} = Fy$ et $y^T H$.

Phase 2 : pour chaque direction $k \in \{1 \cdots D\}$, générer $\mathcal{I}_k$.

Notons qu'il est possible à ce niveau de supprimer les points redondants.

Phase 3 : pour chaque direction $k \in \{1 \cdots D\}$, extraction des C points candidats de $\mathcal{I}_k$ ayant une distance minimale.

Phase 4 : Création de l'ensemble ξ_{start}. puis à partir de chaque point de ξ_{start} utiliser la méthode du gradient itératif.

Phase 5 : Enregistrement de la meilleure solution trouvée x_{hisd}.

3. Complexité

Concernant la complexité du HIS pour les multiplications, celle-ci est en $(D+C)\times \mathcal{O}(n^3)$ où D représente le nombre de directions évaluées et C le nombre de points candidats retenus de ξ_{start}. De manière plus fine intervient le nombre d'itérations θ. En examinant le tableau 1 et la figure 4, les valeurs C =4, D =2 et θ =2 sont suffisants.

La complexité dépend des deux paramètres du HISD (D et C). Dans cette étude, nous considérerons uniquement le cas où n=m. Le HISD est constitué de deux parties. La première concerne le calcul des matrice F et G ainsi que l'extraction des D plus petits vecteurs singuliers de H. Par « bloc » de vecteur x le nombre d'opérations est en $\mathcal{O}(n^3)$. Cette première partie est évaluée à chaque « bloc » le canal étant supposé invariant pendant cet espace de temps.

Le SD fait partie des méthodes optimale mais sa structure n'est pas régulière et ne permet pas une implantation optimale. L'algorithme HISD étant lui essentiellement parallèle, le nombre de cycle moyen est égal à deux. Le HISD est donc particulièrement bien adapté à une architecture matérielle haut débit.

TAB. 1 : complexité phase par phase du HISD

Pour un « bloc » de vecteur x		
$H^T H$ and $\{v_k\}_{k=1}^{D}$	$\otimes$	$\oplus$
	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$
Pour chaque vecteur y reçu		
HISD	$\otimes$	$\oplus$
Phase (1) diversification	$2n^2$	$2n^2$
Phase (2) diversification	Dn^2	$4Dn^2 + Dn(\frac{n+1}{2})$
Phase (3) diversification	$2Dn^2$	$2Dn^3 + DCn$
Phase (4) intensification	$DC\theta\, n^2$	$DC\theta\, n(n+2)$

4. Performance

On considérera un nombre d'antenne N = 5 identique en émission et en réception. La matrice de canal possède des entrées gaussiennes indépendantes et identiquement distribuées de variance unitaire. Le codage MIMO est de rendement plein c'est à dire qu'à chaque temps symbole un symbole S_i différent est émis sur chaque antenne.

Des résultats en terme de TEB en fonction du rapport S/B pour une constellation MAQ4 sont présentés pour les différents algorithmes ZF, MMSE, PIC, SD et HISD (cf. fig. 4). Pour le HISD les résultats sont présentés en fonction du nombre de directions D et du nombre de points candidats sélectionnés C. Les résultats du HISD à 10^{-4} sont dégradés de seulement 0,25 dB par rapport à l'optimum (SD) mais présentent un gain de 2,25 dB par rapport au SDP [6] et au PIC [4].

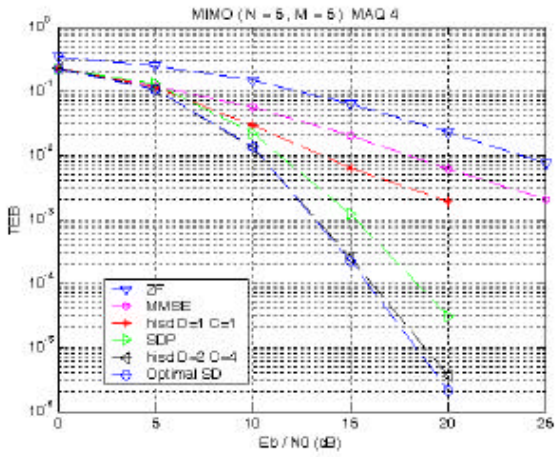


FIG. 4 : Taux d'Erreur Binaire pour un système MIMO 5x5

5. Conclusion

Dans cet article une nouvelle méthode le HISD pour résoudre le problème « NP-complet » $y = Hx+b$ est appliquée aux transmissions de type MAQ en MIMO. En utilisant une approche géométrique qui utilise les directions des axes de bruits dominants obtenus par une décomposition SVD de la matrice du canal, cette méthode offre des performances significatives par rapport aux algorithmes sous optimaux existants comme les décodeurs SDP and PIC. De plus, l'algorithme HISD est implantable aisément sur un composant hardware de type FPGA grâce à un parallélisme inhérent à sa structure. Des essais très prometteurs sur des modulations à un nombre d'état plus important (MAQ16, MAQ64) sont en cours. De plus, une version à sortie souples est actuellement testée. Enfin le HISD peut-être adapté à d'autres problèmes de même type comme le décodage MC-CDMA [3].

Références

[1] G. J. Foschini. *Layered Space-Time Architecture for Wireless Communication in a Fading Environment when Using Multiple Antennas.* Bell Labs Technical Journal, 1(2):41–59, Autumn 1996.

[2] E. Viterbo and J. Boutros. *A universal lattice code decoder for fading channels.* IEEE Trans. on Inf. Theory, pp. 1639-1642, July 1999.

[3] A. Nafkha, C. Roland, E. Boutillon. *A Near-optimal Multiuser Detector for MC-CDMA Systems using Geometrical Approach.* ICASSP'05, Philadelphia, March 2005.

[4] S. Moshavi. *Multi-user Detection for DS-CDMA Communications.* IEEE Communications Magazine, pp. 124-136, Oct. 1996.

[5] H. Artes, D. Seethaler, F. Hawatsch. *Efficient Detection Algorithms for MIMO Channels: A Geometrical Approach to Approximate ML Detection.* IEEE Transactions on Signal Processing, Vol. 51 (2003), 11; 2808 - 2820.

[6] P. Huitian and L. K. Rasmussen. *The application of semidefinite programming for detection CDMA.* IEEE Select. Areas Commun., vol. 19, pp. 1442-1449, Aug. 2001.

[7] S. Verdú, "Optimum Multiuser Signal Detection in Gaussian Noise is NP-hard," Abs. 1985 IEEE Int. Symp. Information Theory, p. 75, Brighton, UK, June 1985.

Comparaison du précodeur Max-$\mathrm{d}_{\min}$ avec les OSTBC et le précodeur Max-SNR pour les systèmes MIMO

Baptiste VRIGNEAU[1], Philippe ROSTAING[1], Ludovic COLLIN[2], Koffi YAO[1]

[1]LEST UMR-CNRS 6165
6 Av. Le Gorgeu, CS 93837, 29238 Brest Cedex 3, France

[2]E3I2 ENSIETA
2 rue Francois Verny 29806 Brest Cedex 09

[1]prenom.nom@univ-brest.fr, [2]ludovic.collin@ensieta.fr

Résumé – Nous avons comparé trois techniques de transmission permettant d'obtenir l'ordre de diversité maximal d'un système MIMO : le précodeur Max-$d_{\min}$, le précodeur Max-SNR et le code d'Alamouti. Les deux précodeurs se distinguent par la nécessité de connaître l'état du canal à l'émission contrairement au code spatio-temporel en bloc. Pour ce faire, nous avons simulé les trois méthodes en utilisant un modèle de canal utilisant des antennes à polarisations orthogonales. Nous avons quantifié l'influence de cette connaissance sur le taux d'erreur binaire ainsi que l'augmentation du nombre d'antennes à la réception.

Abstract – We compared three methods which achieve the maximum diversity order in a MIMO system: the Max-$d_{\min}$ precoder, the Max-SNR precoder and the Alamouti code. The two precoders require the channel state information contrarely to the space-time block code.With this intention, we simulated the three methods with a channel model employing dual-polarized antennas. In order to quantify the influence of this knowledge and the transmit antennas number, simulations of bit error rate are carried out.

1 Introduction

Les communications sans fil évoluent sans cesse à l'image de la norme IEEE 802.11 qui présente plusieurs versions dont les plus connues sont les normes 802.11b et 802.11g. Tout en restant compatibles, elles intégrent les nouveautés technologiques des communications numériques et notamment les systèmes multi-antennes ou MIMO (Multiple-Input Multiple Output). Ces derniers améliorent une liaison dans un environnement à trajets multiples comme un réseau sans fil domestique [1] en augmentant l'efficacité spectrale ou la robustesse de la transmission. Par exemple, le multiplexage spatial (MS) permet une augmentation significative de l'efficacité spectrale [2]. D'autres systèmes comme les codes spatio-temporel en bloc orthogonaux ou OSTBC (Orthogonal Space-Time Bloc Code) et le précodeur maximisant le rapport signal sur bruit (RSB) post-traitement noté Max-SNR améliorent la fiabilité de la transmission. Cette robustesse se traduit par un ordre de diversité maximal. Pour un canal de Rayleigh, sa valeur est $n_T \times n_R$, produit du nombre d'antennes à l'émission avec celui à la réception et correspond à la pente de l'asymptote de la probabilité d'erreur à fort RSB. Ces schémas sont basés sur deux stratégies différentes. Le système OSTBC introduit de la redondance en transmettant plusieurs combinaisons de symboles utilisant plusieurs périodes symbole [3][4]. Cette solution n'exige pas de connaissance du canal au niveau de l'émetteur notée Tx-CSI (Channel State Information) mais diminue l'efficacité spectrale par rapport au MS. Le précodeur Max-SNR émet les symboles dans la direction la plus favorable du canal maximisant le RSB à la réception [5]. Cela implique un précodeur avec Tx-CSI. De plus, un symbole est émis pendant une période diminuant ainsi l'efficacité spectrale par rapport au MS.

Le précodeur Max-$d_{\min}$ introduit récemment [6, 7] dont le principe est basé sur la maximisation de la distance euclidienne minimale entre les symboles reçus, a démontré sa capacité à améliorer à la fois l'efficacité spectrale et la robustesse de la transmission [8]. Comme le Max-SNR, l'information du canal est nécessaire à l'émission.

De plus, nous considérons un modèle de système utilisant des antennes polarisées prenant en compte la corrélation de transmission et de réception ainsi que la capacité de l'antenne à discriminer les polarisations [9].

Comme ces solutions assurent l'ordre de diversité maximal, nous nous sommes intéressés aux bénéfices de la connaissance du canal à l'émission (Tx-CSI) sur les performances du système en terme de probabilité d'erreur bit. Dans la partie 2 de cet article, le modèle du canal utilisé sera détaillé. Les deux parties suivantes présenteront respectivement les STBC puis les précodeurs max-SNR et Max-$d_{\min}$. La partie 5 étudiera l'influence du nombre d'antennes de réception et l'importance de la précision de l'estimation du canal. Enfin, nous présenterons nos conclusions.

2 Modèle du canal de propagation

2.1 Système de base

Comme présenté dans [9], le système utilise une antenne à l'émission et une antenne à la réception avec une double polarisation orthogonale. Ainsi, deux signaux peuvent être émis et reçus indépendamment, ce qui permet de voir ce système

mono-antenne comme un système MIMO avec deux antennes à l'émission et deux antennes à la réception. Cette solution permet de diminuer l'encombrement et le coût. Le modèle utilisé est un canal de Rayleigh et prend en compte les couplages introduits par les antennes et leur capacité à discriminer les deux polarisations au travers de la matrice de corrélation du canal [9]. Les éléments de la matrice du canal de base, $\mathbf{H}_{2\times 2} = \begin{pmatrix} h_{0,0} & h_{0,1} \\ h_{1,0} & h_{1,1} \end{pmatrix}$, sont des variables aléatoires gaussiennes complexes corrélées avec les propriétés suivantes :

$$\begin{array}{rcl} E\{|h_{0,0}|^2\} & = & E\{|h_{1,1}|^2\} \ = \ 1 \\ E\{|h_{0,1}|^2\} & = & E\{|h_{1,0}|^2\} \ = \ \alpha \end{array} \quad (1)$$

où $0 \leq \alpha \leq 1$ dépend de la discrimination de polarisation des antennes et du couplage entre les polarisations induit par les réflexions. Quand α tend vers 0, la discrimination entre les polarisations devient parfaite et le système peut être vu comme deux systèmes SISO (Single Input Single Ouput). La corrélation des éléments de $\mathbf{H}_{2\times 2}$ est définie comme suit :

$$\begin{array}{l} t = \frac{E\{h_{0,0}h_{0,1}^*\}}{\sqrt{\alpha}} = \frac{E\{h_{1,0}h_{1,1}^*\}}{\sqrt{\alpha}} \\ r = \frac{E\{h_{0,0}h_{1,0}^*\}}{\sqrt{\alpha}} = \frac{E\{h_{0,1}h_{1,1}^*\}}{\sqrt{\alpha}} \end{array} \quad (2)$$

où t est le cœfficient de correlation induit lors de la transmission et r celui de la réception. Nous supposons que $E\{h_{0,0}h_{1,1}^*\} = E\{h_{1,0}h_{0,1}^*\} = 0$. En considérant toutes ces hypothèses, la matrice de corrélation du système de base est définie par $\mathbf{R}_{\mathbf{H}_{2\times 2}} = E\{\mathrm{vect}\{\mathbf{H}_{2\times 2}\}\,\mathrm{vect}\{\mathbf{H}_{2\times 2}\}^*\}$ et peut être exprimée comme :

$$\mathbf{R}_{\mathbf{H}_{2\times 2}} = \begin{pmatrix} 1 & r\sqrt{\alpha} & t\sqrt{\alpha} & 0 \\ r\sqrt{\alpha} & \alpha & 0 & t\sqrt{\alpha} \\ t\sqrt{\alpha} & 0 & \alpha & r\sqrt{\alpha} \\ 0 & t\sqrt{\alpha} & r\sqrt{\alpha} & 1 \end{pmatrix} \quad (3)$$

où $\mathrm{vect}(\mathbf{H}_{2\times 2})$ est un vecteur de taille 4×1 obtenu par concaténation de chaque colonne de la matrice $\mathbf{H}_{2\times 2}$.

2.2 Extension à plusieurs antennes

Par la suite, nous augmentons le nombre d'antennes de réception tout en gardant une seule antenne d'émission. Ainsi, nous avons un système physique SIMO (Single Input Single Output) avec une antenne d'émission et une ou plusieurs antennes de réception et, grâce à la double polarisation, un système MIMO équivalent avec deux antennes d'émission et un nombre n_R d'antennes de réception égal au double de celui du système physique.

Par souci de simplification, nous considérons que le système SIMO peut être décomposé en $n_R/2$ systèmes de base indépendants ayant les caractéristiques statistiques décrites précédement. La matrice de corrélation est alors :

$$\mathbf{R}_{\mathbf{H}} = \mathbf{I}_{n_R/2} \otimes \mathbf{R}_{\mathbf{H_{2\times 2}}} \quad (4)$$

avec $\otimes$ le produit de Kronecker et $\mathbf{I}_{n_R/2}$ la matrice identité de taille $n_R/2$.

3 Les codes spatio-temporels orthogonaux (OSTBC)

Pour la suite de l'article, nous fixons la puissance totale moyenne émise P_0 et supposons que l'énergie moyenne des symboles est unitaire, $E[|s_i|^2] = 1$. Le RSB est alors défini par P_0/σ^2 où σ^2 représente la variance du bruit.

Supposant un canal stationnaire sur N_P périodes symbole, un flux de n_S symboles est transmis sur les n_T antennes d'émission grâce à une matrice de code $\mathbf{C}$ $[n_T \times N_P]$. Typiquement, les éléments de la matrice $\mathbf{C}$ sont de la forme $\pm s_i$ ou $\pm s_i^*$. Par exemple, le célèbre code d'Alamouti [3] correspondant à $n_T = n_S = 2$ est $\mathbf{C} = \begin{pmatrix} s_0 & -s_1^* \\ s_1 & s_0^* \end{pmatrix}$. La figure 1 illustre

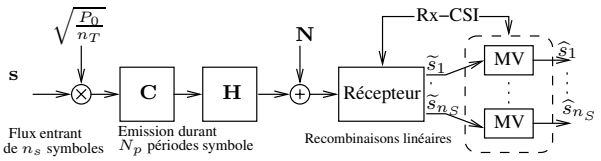


FIG. 1 – Synoptique d'une transmission OSTBC

une telle transmission. L'utilisation d'un OSTBC signifie que l'émetteur utilise N_P périodes symboles. Par conséquent, le taux de transmission est affecté par le rendement $R = n_S/N_P$ et retarde l'estimation de N_P périodes. Le signal reçu est donné par :

$$\mathbf{Y} = \sqrt{P_0/n_T}\mathbf{H}\mathbf{C} + \mathbf{N} \quad (5)$$

où $\mathbf{H} = [h_{i,j}]_{i,j=1}^{n_R,n_T}$ $[n_R \times n_T]$ est la matrice de canal, $\mathbf{Y}$ $[n_R \times N_P]$ est le vecteur d'échantillons reçus et $\mathbf{N}$ $[n_R \times N_P]$ représente un bruit blanc additif gaussien temporellement et spatialement décorrélé de variance σ^2. Un avantage important des OSTBC est la simplicité du récepteur. La première étape consiste à estimer la matrice de canal $\mathbf{H}$. Ensuite, les estimations des symboles peuvent être découplées grâce à l'orthogonalité du code [4]. Le critère du maximum de vraisemblance (MV) peut être appliqué à chaque symbole de la constellation $\mathcal{C}$:

$$\widehat{s}_i = \arg\min_{s_l \in \mathcal{C}} \left| \widetilde{s}_i - \sqrt{P_0/n_T}\|\mathbf{H}\|_F^2 s_l \right| \quad (6)$$

pour $i = 1, \ldots, n_S$ et $\|\mathbf{H}\|_F^2 = trace(\mathbf{H}\mathbf{H}^H) = trace(\mathbf{H}^H\mathbf{H})$ désigne la norme de Frobenius au carré. Le système MIMO est équivalent à n_S voies indépendantes illustrées par la figure 2 où n_i est un bruit blanc gaussien de variance σ^2. La force des OSTBC est d'obtenir le gain $||\mathbf{H}||_F^2/n_T$ sur le RSB et un ordre de diversité $n_T \times n_R$ sans connaissance du canal à l'émission.

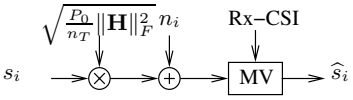


FIG. 2 – Schéma équivalent d'une transmission OSTBC

Le code d'Alamouti correspondant à notre configuration matérielle pour $n_T = 2$ offre le meilleur rendement $R = 1$.

4 Précodeur linéaire possédant l'information du canal à l'émission

4.1 Diagonalisation de la matrice canal

Les précodeurs utilisent la connaissance du canal via une matrice de précodage $\mathbf{F}$ à l'émetteur et une matrice de décodage $\mathbf{G}$ au récepteur. Une première opération permet de simplifier le canal en décomposant $\mathbf{F} = \mathbf{F}_v\mathbf{F}_d$ et $\mathbf{G} = \mathbf{G}_d\mathbf{G}_v$ puis

en définissant le canal virtuel $\mathbf{H}_v = \mathbf{G}_v\mathbf{H}\mathbf{F}_v$. Les matrices $\mathbf{F}_v$ et $\mathbf{G}_v$ permettent de blanchir le bruit et de diagonaliser le canal à l'aide de la décomposition en valeurs singulières de $\mathbf{H}$ (Figure 3). Note : dans le cas d'un bruit blanc, $E[\mathbf{n}\mathbf{n}^*] = \sigma^2\mathbf{I}_{n_R}$, $\mathbf{H}_v = \mathtt{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_b})$ où $b = \min(n_T, n_R)$ et les λ_i sont directement les valeurs propres de $\mathbf{H}\mathbf{H}^*$ rangées par ordre décroissant.Le canal virtuel est représenté par b voies indépendantes avec chacune un gain donné par la valeur singulière $\sqrt{\lambda_i}$ et un bruit blanc additif de variance σ^2.

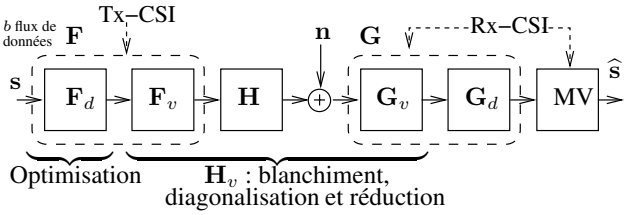


FIG. 3 – Synoptique de la diagonalisation du canal

La nouvelle relation d'entrée-sortie est :

$$\mathbf{y} = \mathbf{G}_d\mathbf{H}_v\mathbf{F}_d\mathbf{s} + \mathbf{G}_d\mathbf{n}_v \tag{7}$$

La matrice $\mathbf{G}_d$ n'ayant aucune influence sur la décision selon le maximum de vraisemblance est choisie égale à la matrice identité $\mathbf{I}_2$. Il reste à trouver la matrice $\mathbf{F}_d$ qui résulte de l'optimisation d'un critère à définir. Nous allons en voir deux exemples.

4.2 Optimisation au sens du RSB

Le premier précodeur Max-SNR, bien connu dans la littérature [5], maximise le RSB post-traitement. La matrice de précodage $\mathbf{F}_d$ permet d'optimiser ce critère sous la contrainte de puissance $\|\mathbf{F}_d\|_F^2 = P_0$:

$$\mathbf{F}_d = \arg\max_{\mathbf{F}'_d} ||\mathbf{H}_v\mathbf{F}'_d||_F^2 \tag{8}$$

Le résultat de cette optimisation donne la matrice de précodage :

$$\mathbf{F}_d = \sqrt{P_0}\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \tag{9}$$

Maximiser le RSB consiste donc à émettre toute la puissance sur la voie virtuelle la plus favorable correspondant à la valeur singulière la plus forte. La relation d'entrée-sortie devient :

$$y = \sqrt{P_0\lambda_1}s + n \quad \text{avec } E[|n|^2] = \sigma^2 \tag{10}$$

4.3 Optimisation de la distance minimale

La distance euclidienne minimale d_{min} entre deux points de la constellation de réception influence les performances du système, spécialement quand un estimateur utilise le MV. Le précodeur $\mathbf{F}_d$ maximise ce critère sous la contrainte de puissance $\|\mathbf{F}_d\|^2 = P_0$ selon le critère suivant :

$$\mathbf{F}_d = \arg\max_{\mathbf{F}'_d}\left\{\min_{\mathbf{s}_k,\mathbf{s}_l\in\mathcal{C}^b,\mathbf{s}_k\neq\mathbf{s}_l} \|\mathbf{H}_v\mathbf{F}'_d(\mathbf{s}_k - \mathbf{s}_l)\|\right\} \tag{11}$$

Dans notre cas, $b = 2$ donnant deux valeurs propres λ_1 et λ_2 et la solution de (11) pour une MAQ-4 s'écrit [7] :

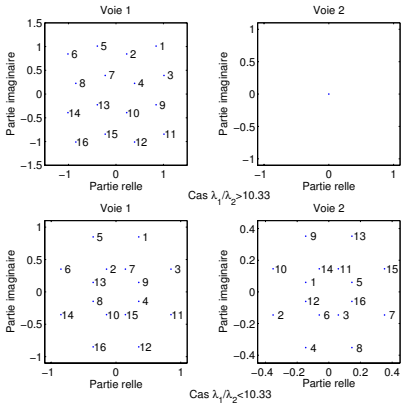


FIG. 4 – Deux types de contellation de réception optimisant d_{min} pour une modulation MAQ-4 et deux voies. Les points de la constellation sont numérotés afin d'identifier sur chaque voie les 16 vecteurs possibles.

- si $\lambda1/\lambda2 \geq 10.33$

$$\mathbf{F}_d = \sqrt{P_0}\begin{pmatrix} \sqrt{\frac{3+\sqrt{3}}{6}} & \sqrt{\frac{3+\sqrt{3}}{6}}e^{j\frac{\pi}{12}} \\ 0 & 0 \end{pmatrix} \tag{12a}$$

- si $\lambda1/\lambda2 < 10.33$

en posant $\psi = \arctan\left((\sqrt{2}-1)\sqrt{\frac{\lambda_1}{\lambda_2}}\right)$

$$\mathbf{F}_d = \frac{\sqrt{P_0}}{2}\begin{pmatrix} \cos\psi & 0 \\ 0 & \sin\psi \end{pmatrix}\begin{pmatrix} \sqrt{2} & 1+i \\ -\sqrt{2} & 1+i \end{pmatrix} \tag{12b}$$

Le rapport des valeurs propres λ_1/λ_2 permet de choisir le précodeur. Lorsque le rapport est supérieur à 10.33, le précodeur prend la première forme (12a) et ne dépend pas des λ_i. Toute la puissance est émise sur la première voie offrant une solution proche du Max-SNR à une rotation près. Lorsque la seconde voie a un gain suffisant, $\lambda_1/\lambda_2 < 10.33$, le précodeur (12b) émet du signal sur les deux voies afin de profiter des deux gains pour optimiser d_{min}. L'expression de $\mathbf{F}_d$ dépend alors que du rapport λ_1/λ_2. Les constellations à la réception sont illustrées par la figure (4), en haut la solution (12a) et en bas celle de (12b).

5 Simulations du taux d'erreur binaire

Les différentes simulations de type Monte Carlo sont obtenues en utilisant les caractéristiques statistiques d'un canal réaliste donné dans [9]. Les valeurs des paramètres définis précédemment sont $t = 0.5$, $r = 0.3$ et $\alpha = 0.4$. Le précodeur Max-SNR et le code d'Alamouti utilisent une MAQ-16 afin de maintenir la même efficacité spectrale que le précodeur Max-$d_{\min}$.

La figure 5 donne les résultats du TEB des trois systèmes pour $n_R = 2$ correspondant à une configuration physique ne comportant qu'une seule antenne à l'émission et à la réception. Le précodeur Max-$d_{\min}$ et Max-SNR sont équivalents avec cependant un avantage au précodeur Max-SNR à faible RSB. Le code d'Alamouti accuse une perte de gain de 2 dB pour un fort RSB. Nous avons estimé l'ordre de diversité proche de 4 ($n_T \times n_R$) pour les trois cas. Rappelons que lorsque $\alpha = 0$, les termes croisés sont nuls diminuant la redondance de l'information et l'ordre de diversité. Malgré la faible valeur de α

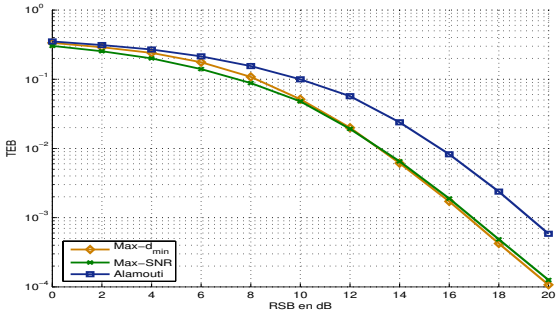


FIG. 5 – Simulation dans un canal réaliste avec $n_R = 2$.

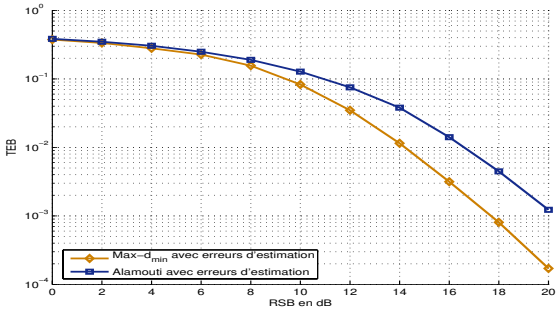


FIG. 6 – Influence des erreurs d'estimation du canal sur le précodeur Max-$d_{\min}$ et le code d'Alamouti dans un canal réaliste avec $n_R = 2$.

($\alpha = 0.4$), les systèmes peuvent exploiter la diversité de polarisation offerte.

Lors de l'émission, la CSI permet un gain important pour les deux précodeurs et c'est pourquoi nous avons introduit des erreurs d'estimation pour évaluer leur robustesse. L'émetteur et le récepteur utilisent une matrice d'estimation du canal, $\mathbf{H}_{est}$, comprenant des erreurs. Le modèle utilisé est défini par $\mathbf{H}_{est} = \mathbf{H} + \varepsilon$ où les éléments de ε sont des variables aléatoires gaussiennes centrées, indépendantes et de variance 1/(4RSB) [7]. La figure 6 présente les TEB pour $n_R = 2$ lorsque l'estimation du canal est imparfaite. Les performances du précodeur Max-$d_{\min}$ et Max-SNR sont proches et seul le TEB du précodeur Max-$d_{\min}$ est représenté. Une perte de 1 dB est constatée pour les trois systèmes (figures 5 et 6). Le code d'Alamouti est pénalisé autant que les autres alors qu'il n'utilise l'estimation du canal seulement qu'à la réception. La perte est due à une apparition d'interférences entre symboles.

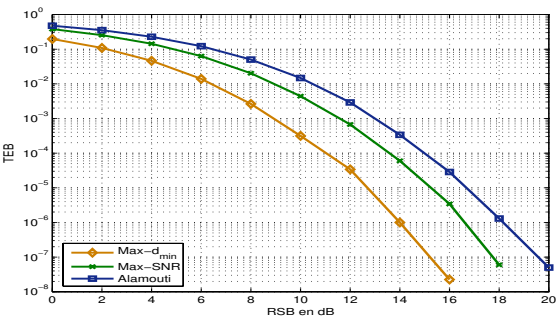


FIG. 7 – Simulation dans un canal réaliste avec $n_R = 6$ (3 antennes physiques à double polarisation).

La valeur de n_R est augmentée à 6 correspondant à 3 antennes physiques de réception. Les résultats du TEB sont donnés sur la figure 7. Le précodeur Max-$d_{\min}$ se distingue nettement du précodeur Max-SNR avec un gain de 3 dB pour un fort RSB. Ce gain s'explique par l'utilisation de la voie 2 avec le gain λ_2. En effet, la statistique du rapport λ_1/λ_2 dépend du nombre d'antennes et les probabilités d'utilisation du précodeur (12b) sont $P(\lambda_1/\lambda_2 < 10.33|_{n_R=2}) = 0.01$ et $P(\lambda_1/\lambda_2 > 10.33|_{n_R=6}) = 0.44$. En supprimant λ_2, le précodeur Max-SNR montre alors ses limites. De plus, son gain par rapport au code d'alamouti diminue de 2 dB pour $n_R = 2$ à 1.6 dB pour $n_R = 6$. Le code d'Alamouti utilise aussi la valeur propre λ_2 au travers du gain sur le RSB $||\mathbf{H}||_F^2/2 = (\lambda_1 + \lambda_2)/2$. Finalement, seul le précodeur Max-$d_{\min}$ utilise à la fois λ_2 et la CSI et obtient logiquement les meilleurs performances.

6 Conclusion

Nous avons simulé les taux d'erreurs binaires de trois systèmes assurant un ordre de diversité maximum dont la différence majeure est la connaissance du canal à l'émission. Pour ce faire, nous avons utilisé un modèle de canal prenant en compte la diversité de polarisation et la corrélation. Pour une seule antenne physique ($n_R = 2$), le précodeur Max-$d_{\min}$ et Max-SNR sont équivalents. Le code d'Alamouti offre une solution simple à mettre en œuvre mais accuse une perte de 2 dB sur le précodeur Max-SNR. Ce code reste un compromis intéressant pour $n_R = 2$ mais perd de son attrait lorsque ce nombre augmente. En effet, le précodeur Max-$d_{\min}$ tire parti de toute la configuration matérielle et obtient un gain significatif de 3 dB sur le Max-SNR et près de 5 dB sur le code d'Alamouti. De plus lorsque $n_R = 2$, les trois méthodes ont la même robustesse aux erreurs d'estimation du canal que l'émetteur utilise ou non la CSI.

Références

[1] G.J. Foschini and M.J. Gans. On limits of wireless communications in fading environment when using multiple antennas. *Wireless Personal Communications*, 6 :331–335, 1998.

[2] I.E. Telatar. Capacity of multi-antenna gaussian channels. *Eur. Trans. Telecommun.*, 10(6) :585–595, Nov./Dec. 1999.

[3] S. M. Alamouti. A simple diversity technique for wireless communications. *IEEE J. Select. Areas Commun.*, 16 :1451–1458, Oct. 1998.

[4] V. Tarokh, N. Seshadri, and A. R. Calderbank. Space-time codes for high data rate wireless communication : Performance criterion and code construction. *IEEE Trans. Inform. Theory*, 44 :744–765, Mar. 1998.

[5] P. Stoica and G. Ganesan. Maximum-SNR spatial-temporal formatting designs for MIMO channels. *IEEE Trans. Signal Processing*, 50(12) :3036–3042, Dec. 2002.

[6] L. Collin. *Optimisation de systèmes multi-antennes basée sur la distance minimale.* PhD thesis, Université de Bretagne Occidentale, `http://www.univ-brest.fr/lest/tst`, décembre 2002.

[7] L. Collin, O. Berder, P. Rostaing, and G. Burel. Optimal minimum distance-based precoder for mimo spatial multiplexing systems. *IEEE Trans. Signal Processing*, 52(3) :617–627, March 2004.

[8] O. Berder, L. Collin, P. Rostaing, and G. Burel. Précodeur linéaire basé sur la distance euclidienne minimale pour les systèmes de transmission multi-antennes. *GRETSI*, 2003.

[9] R.U. Nabar, H. Bolcskei, V. Erceg, D. Gesbert, and A.J. Paulraj. Performance of multiantenna signaling techniques in the presence of polarization diversity. *IEEE Trans. Signal Processing*, 50(10) :2553 – 2562, Oct 2002.

Construction d'une Modulation Codée Non Cohérente MAQ-32

S.HADDA, S.JARBOUI,

Laboratory of communication systems, National School of engineers of Tunis,
Campus Universitaire, Le Belvedere 1002 Tunis, TUNISIA

e-mail: slaheddine.jarboui@fss.rnu.tn

Résumé – Dans cet article nous présentons un nouveau schéma de la modulation codée MAQ - 32 non-cohérente (MAQ-32 NC). Ce schéma est obtenu après modifications de la structure conventionnelle de la modulation codée en treillis MAQ-32 (MCT-MAQ-32). Les résultats de simulation montrent que les performances de la MAQ-32 NC avec détection non-cohérente sont très proches de celles de la MCT-MAQ-32 optimale, proposée par Ungerboeck, avec détection cohérente. La MAQ-32 NC proposée dans cet article est une extension de la MAQ-16 NC déjà présentée par P. Fan et X.-G. Xia.

Abstract *– In this letter, we present a new non-coherent coded 32-QAM (NC 32-QAM) schema derived from the trellis coded 32-QAM (TC 32-QAM) with several modifications. The simulation results show that the performance of the NC 32-QAM with non-coherent detection is close to the one of the optimal TC 32-QAM with coherent detection. The NC 32-QAM in this work is an extension of the NC 16-QAM proposed by P. Fan and X.-G. Xia.*

1. Introduction

La détection cohérente des modulations codées en treillis (MCT), telles qu'elles ont été définies par Ungerboeck dans [1] et [2], nécessite au niveau du récepteur la disposition locale d'une onde porteuse verrouillée en fréquence et en phase sur celle utilisée à l'émission. De plus le récepteur doit disposer d'une horloge, permettant d'échantillonner les signaux reçus à la même cadence que celle de l'émission. Dans le cas de la détection non-cohérente, les signaux de synchronisation sont récupérés à partir des signaux de données uniquement. Le spectre du signal reçu ne contient pas de raie, ni à la fréquence de l'onde porteuse, ni aux fréquences multiples de la fréquence de l'horloge. Ceci permet de mieux gérer la bande passante du signal et de contrôler plus efficacement la puissance émise. Toutefois la conception des MNCs devient d'autant plus compliquée que la constellation des signaux est plus dense. Les cas de la MDP-8 NC et la MAQ-16 NC ont été déjà traités respectivement dans [3] et [4]. Dans cet article nous discutons le cas de la MAQ-32 NC. Nous avons défini une méthode une nouvelle méthode d'affectation des symboles binaires aux points de la constellation afin d'éliminer toute sorte d'ambiguïté lors de la détection. De plus nous avons optimisé les paramètres du décodeur afin d'obtenir une performance maximale. Les résultats de simulations montrent que les performances de la MAQ-32 NC sont très proches de celle de la MAQ-32 cohérente

2. Construction de la MAQ-32 NC

Dans [3], il a été spécifié qu'une MCT avec des branches parallèles ne peut être convertie en modulation non cohérente. Pour cette raison, tous les bits présents à l'entrée de la MCT doivent être codés. A chaque instant iT, où T désigne l'intervalle de modulation, le codeur convolutif à 32 états, présenté par la figure 1, reçoit un bloc de 4 bits et génère un symbole codé sur 5 bits. Les coefficients de parités du codeur convolutif exprimés en forme octale sont $\underline{h}^0 = 41$, $\underline{h}^1 = 2$, $\underline{h}^2 = 4$, $\underline{h}^3 = 10$ et $\underline{h}^4 = 20$.

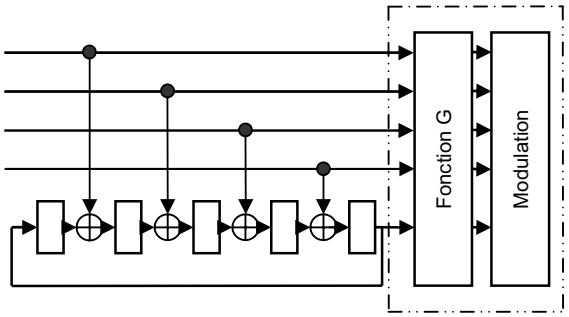


FIG. 1 : structure de la MAQ-32 NC

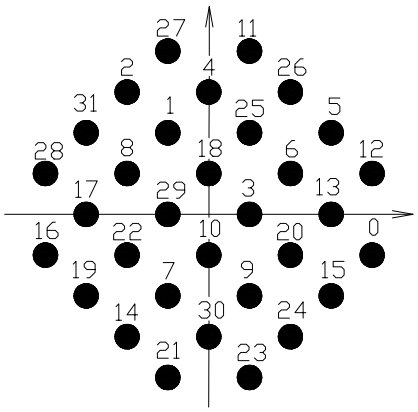


FIG. 2 : la constellation MAQ-32 sélectionnée

Si on suppose que le canal de transmission est à bruit blanc Gaussien additif (BBGA) et une absence totale de synchronisation, l'échantillon prélevé à l'instant *iT* à la sortie du canal de transmission peut être exprimé en bande de base de la manière suivante :

$$r_i = a_i e^{j\theta_0} + \eta_i \ , \tag{1}$$

où a_i est un nombre complexe discret qui représente le signal MAQ-32 délivré par la MCT à l'instant *iT*, θ_0 est une erreur de phase arbitraire due à l'absence de synchronisation et η_i représente la contribution du bruit.

Pour reconstruire la phase θ_0 on procède d'une manière analogue à [4]. La métrique suivante est alors utilisée pour détecter les *n* premiers symboles reçus,

$$M(\mathbf{R},\mathbf{S}) = \sum_{i=0}^{i=n-2} \left| r_i r_{i+1}^* - a_i a_{i+1}^* \right|^2 , \tag{2}$$

où, $\mathbf{R} = \{r_0, r_1, ..., r_{n-1}\}$ représente la séquence des *n* premiers échantillons reçus, r_i^* désigne le conjugué de r_i et $\mathbf{S} = \{a_0, a_1, ..., a_{n-1}\}$ est une séquence candidate de *n* signaux MAQ-32 qui peut être générée par la MCT. La séquence détectée est la séquence particulière $\mathbf{S_0} = \{\widetilde{a}_0, \widetilde{a}_1, ..., \widetilde{a}_{n-1}\}$ qui vérifie la condition suivante :

$$M(\mathbf{R},\mathbf{S_0}) = \min_S \{M(\mathbf{R},\mathbf{S})\}. \tag{3}$$

La séquence $\mathbf{S_0}$ est utilisée pour déduire une valeur approchée de θ_0 en utilisant la formule suivante :

$$\widetilde{\theta}_0 = Arg\left(\frac{\sum_{k=0}^{n-1} r_i \widetilde{a}_i}{\sum_{k=0}^{n-1} |\widetilde{a}_i|^2} \right) \tag{4}$$

Soit $\mathbf{S}_1 = \{a_0, a_1, ..., a_{n-1}\}$ et $\mathbf{S}_2 = \{b_0, b_1, ..., b_{n-1}\}$ deux séquences différentes et de même longueur *n*. Evidemment, si $\{a_0 a_1^*, a_1 a_2^*, ..., a_{n-2} a_{n-1}^*\} = \{b_0 b_1^*, b_1 b_2^*, ..., b_{n-2} b_{n-1}^*\}$ alors $\mathbf{S}_1$ et $\mathbf{S}_2$ sont équivalentes d'après ce processus de détection. Afin d'éviter toute sorte d'ambiguïté au niveau du décodeur, il faut s'assurer que de telles séquences ne peuvent être générées par la MCT. Généralement, les deux séquences $\mathbf{S}_1$ et $\mathbf{S}_2$ sont dites différentiellement équivalentes par cohérence (DEC) si pour chaque entier *i*, ($0 \le i \le n$-1) l'égalité suivante est vérifiée [4] :

$$a_i a_{i+1}^* = b_i b_{i+1}^* \ . \tag{5}$$

Soit z_0 le nombre complexe tel que :

$$\frac{b_0}{a_0} = z_0 . \tag{6}$$

On peut vérifier par un simple calcul que $\mathbf{S}_1$ et $\mathbf{S}_2$ sont deux séquences DEC si et seulement si on a :

$$\begin{cases} \dfrac{b_i}{a_i} = \dfrac{1}{z_0^*} & si \quad i = 2p+1 \\ \dfrac{b_i}{a_i} = z_0 & si \quad i = 2p. \end{cases} \tag{7}$$

Par conséquence, on peut exprimer $\mathbf{S}_2$ comme suit :

$$\mathbf{S}_2 = \left\{ z_0 a_0, \frac{a_1}{z_0^*}, z_0 a_2, \frac{a_3}{z_0^*}, ..., z_0 a_{2p}, \frac{a_{2p+1}}{z_0^*}, ... \right\}. \tag{8}$$

La MCT présentée par la figure.1 peut générer des séquences DEC. Pour éliminer ces séquences on procède d'une manière analogue à [4]. Ainsi, à chaque mot code $c_i \in \{0, ..., 31\}$ issu du codeur convolutif de la MCT on associe l'image du signal a_i par la fonction G définie comme suit :

Si a_i désigne le signal canal de rang c_i dans la constellation MAQ-32 alors $\mathrm{G}(a_i)$ est le signal canal de rang :

$$\mathrm{rang}(\mathrm{G}(a_i)) = \begin{cases} c_i & si \quad i = 2p \\ (2 + c_i) \bmod (32) & si \quad i = 2p+1, \end{cases} \tag{9}$$

On remarque que la fonction G laisse a_i invariant lorsque *i* est pair. Si on applique G aux deux séquences $\mathbf{S}_1$ et $\mathbf{S}_2$ on obtient :

$$\begin{aligned} \mathrm{G}(\mathbf{S}_1) &= \{\mathrm{G}(a_0), \mathrm{G}(a_1), \cdots, \mathrm{G}(a_{2p}), \mathrm{G}(a_{2p+1}), \cdots\} \\ &= \{a_0, \mathrm{G}(a_1), \cdots, a_{2p}, \mathrm{G}(a_{2p+1}), \cdots\}, \end{aligned} \tag{10}$$

et

$$\mathrm{G}(\mathbf{S}_2) = \left\{ z_0 a_0, \mathrm{G}(\frac{a_1}{z_0^*}), \cdots, z_0 a_{2p}, \mathrm{G}(\frac{a_{2p+1}}{z_0^*}), \cdots \right\}. \tag{11}$$

Les séquences $\mathrm{G}(\mathbf{S}_1)$ et $\mathrm{G}(\mathbf{S}_2)$ ne sont plus DEC si et seulement si la condition suivante est vérifiée :

$$\mathrm{G}(\frac{a_{2p+1}}{z_0^*}) \neq \frac{\mathrm{G}(a_{2p+1})}{z_0^*} \tag{12}$$

En remarquant que

$$\frac{a_{2p+1}}{z_0^*} \text{ et } a_{2p+1}$$

sont deux signaux MAQ-32, on peut montrer que tous les séquences DEC sont entièrement éliminées si pour chaque couple de signaux MAQ-32 différents, la condition suivante est vérifiée :

$$\frac{\mathrm{G}(a)}{\mathrm{G}(a')} \neq \left. \frac{a}{a'} \right|_{\forall (a \neq a')} \tag{13}$$

La figure 2 montre le résultat d'une recherche effectuée à l'aide de l'ordinateur concernant une constellation MAQ-32 qui vérifie la condition (13).

3. Détection Non-cohérente

La figure 3 présente la structure en blocs de l'algorithme de décodage. La première étape consiste à détecter les n premiers échantillons reçus en utilisant la métrique en (2). Le résultat de détection sera utilisé par la suite pour déterminer une valeur approximative de θ_0. En deuxième étape le décodeur effectue une correction de phase des signaux reçus et débute l'opération de décodage en utilisant l'algorithme conventionnel de Viterbi. Les signaux décodés sont utilisés pour effectuer la mise à jour de la phase tout le long du décodage.

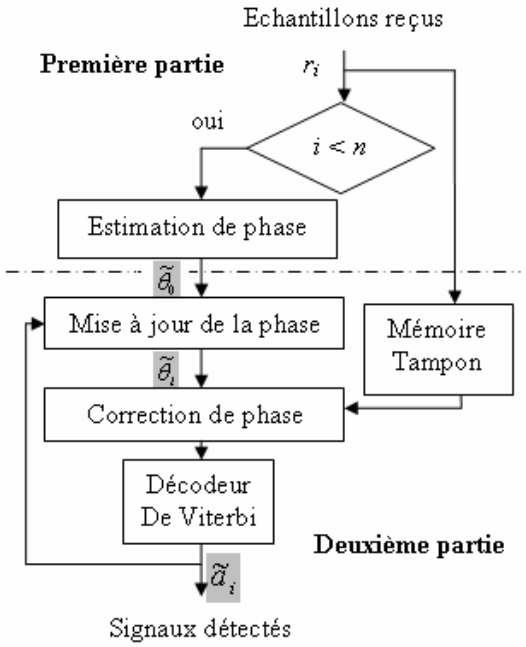


FIG. 3 : la structure en blocs de l'algorithme de décodage

Evidement on aurait pu utiliser la métrique en (2) pour décoder l'ensemble des échantillons reçus. Cependant ceci nécessite énormément d'espace mémoire et de temps de calcul. De plus les performances de décodage relatives à cette métrique ne sont pas satisfaisantes. La figure 4 donne à titre d'exemple le taux d'erreur binaire (TEB) en fonction du rapport signal bruit pour n = 300, et $\theta_0 = \pi/9$. Pour ces raisons l'utilisation de la métrique en (2) a été limitée à la détermination d'une valeur approchée de θ_0.

L'étape suivante de décodage utilise l'algorithme conventionnel de Viterbi. Cependant afin d'effectuer une correction de phase aux échantillons reçus, l'expression de la métrique des branches sera modifiée comme suit :

$$M_b(r_i, a^{(q)}) = \left| r_i e^{-j\tilde{\theta}_i} - a_b \right|^2 . \qquad (14)$$

Où, a_b désigne le signal MAQ-32 associé à la branche et $\tilde{\theta}_i$ représente la valeur estimée de θ_0 à l'instant iT. Le calcul de $\tilde{\theta}_i$ est fait à partir des équation suivantes [4]:

$$\begin{aligned} \phi_i &= \frac{\sum_{k=D}^{K+D} r_{i-k}\tilde{a}_{i-k}}{\sum_{k=D}^{K+D} |\tilde{a}_{i-k}|^2} \\ \theta_i &= \alpha e^{j\tilde{\theta}_{i-1}} + (1-\alpha)\phi_i \qquad (15) \\ \tilde{\theta}_i &= Arg\left(\frac{\theta_i}{|\theta_i|}\right), \end{aligned}$$

Où, α est un paramètre récursif appartenant à l'intervalle]0, 1[, $\tilde{a}_i$ représente un signal canal détecté, ϕ_i est une mesure du déphasage à partir des K derniers signaux détectés et D est un entier naturel supérieur à 5 fois la longueur de contrainte du codeur convolutif. L'optimisation des performances du décodeur consiste à chercher des valeurs convenables des paramètres α et K. La figure 5 est obtenue par simulation et montre la variation du TEB en fonction de α pour K = 10, 15 et 20, et en présence d'un rapport signal bruit typique (E_b/N_0=10 dB). D'après ces courbes on voit clairement que seules les valeurs de α entre 0.85 et 0.95 sont intéressantes puisqu' elles assurent le plus petit TEB.

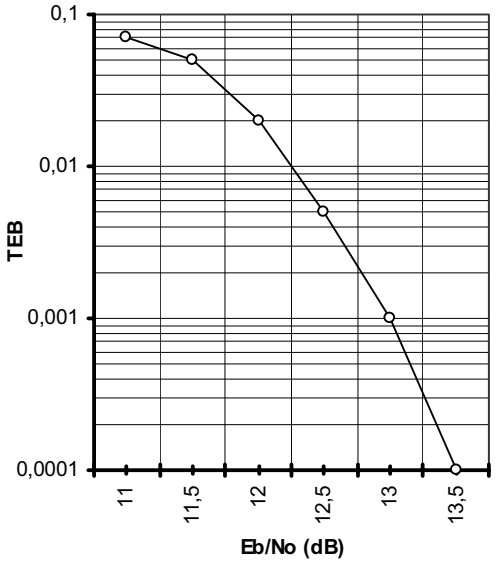


FIG. 4 : performance de la métrique M

Les résultats de simulation de la MAQ-32 NC pour K = 10 et 20, et $\alpha = 0.95$ sont représentées par la figure 6. Les troisième et quatrième courbes indiquées par la légende correspondent respectivement au cas de la MAQ-32 NC avec détection cohérente et la modulation codée en treillis MAQ-32 (MCT-MAQ32) optimale telle qu'elle a été décrite par Ungerboeck [2]. On voit clairement que pour un TEB de l'ordre de 10^{-6} la MAQ-32 NC avec détection non-cohérente perd respectivement 0.15 dB et 0.3 dB seulement par rapport à la MAQ-32 NC avec détection cohérente et la MCT-MAQ-32.

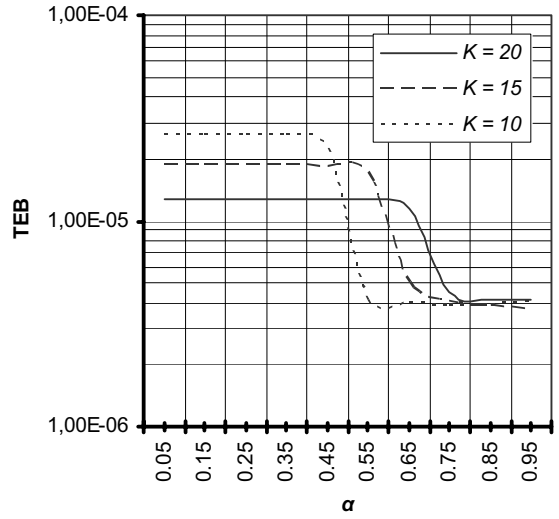


FIG. 5 : le taux d'erreur binaire en fonction de α

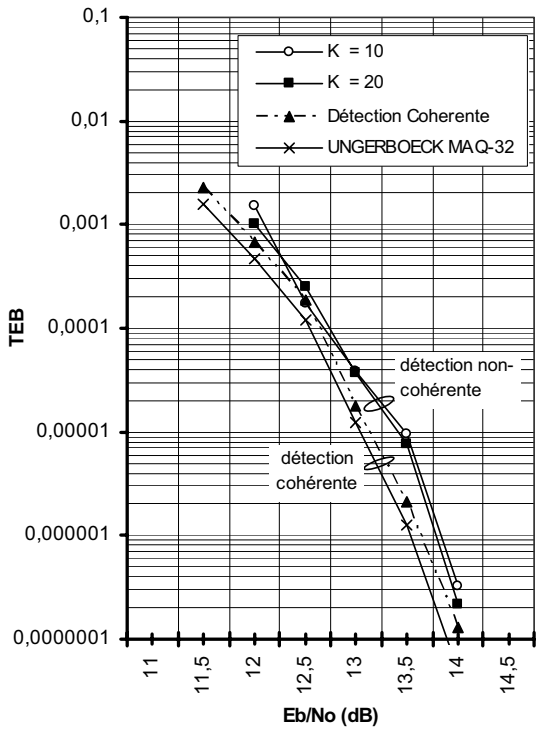


FIG. 6 : performance de la détection MAQ-32 NC

4. Conclusion

Dans cet article nous avons présenté un nouveau schéma de modulation MAQ-32 non cohérente et développé une nouvelle méthode d'affectation des symboles binaires aux points de la constellation afin d'éliminer les séquences DCE. La structure à deux étages de l'algorithme de décodage a garanti au décodeur une quasi autonomie et une indépendant vis a vis des données reçues. Seulement 0.3 dB sont perdus par rapport à la modulation codée en treillis MAQ-32 optimale. La méthode proposée peut être généralisée à fin de simplifier la conception d'autres schémas de modulateurs non cohérents.

Références

[1] G. UNGERBOECK. *Channel coding with multilevel/phase signals*. IEEE Transactions Inform. Theory, vol. IT-28, pp. 55-66, Jan.1982

[2] G. UNGERBOECK. *Trellis-Coded Modulation with Redundant Signal Sets Part II: State of the Art*. IEEE Communications Magazine, Vol. 25, No. 2, February 1987.

[3] RUEY-YI WEI AND MAO-CHAO LIN. *Noncoherent coded modulation constructed from conventional trellis coded modulation*. IEEE Communications. Letters, vol. 2, pp. 260-262, Sept.1998.

[4] PINGYI FAN AND XIANG-GEN XIA. *A noncoherent coded modulation for 16QAM*. IEEE Transactions on Communications, vol. 49, pp. 585-588, April 2001.

Synthèse de la notion de facteur de crête et application aux modulations monoporteuse

Yves LOUËT, Jacques PALICOT

IETR - SUPELEC, Campus de RENNES
Équipe SCEE (Signal, Communications, Électronique Embarquée)
Avenue de la Boulaie, BP 81127 35511 CESSON SEVIGNE - France
Yves.Louet@rennes.supelec.fr, Jacques.Palicot@rennes.supelec.fr

Résumé – Les modulations à enveloppes non constantes présentent l'inconvénient d'être sensibles aux non linéarités (principalement de l'amplificateur de puissance) du fait d'importantes fluctuations du signal modulé. Pour quantifier ces fluctuations, on définit le " facteur de crête ", rapport de la puissance instantanée maximale et moyenne du signal en question. Ce facteur de crête est cependant décliné dans la littérature sous différentes appellations comme CF pour Crest Factor, PAPR pour Peak to Average Power Ratio ou encore PMEPR pour Peak to Mean Envelope Power. De plus, les signaux peuvent prendre des formes différentes (continue ou échantillonnée), être observés sur des durées différentes ou être traités à des niveaux différents (en bande de base ou en radio fréquence).

Devant cette multiplicité, nous établissons dans un premier temps une définition générique du facteur de crête, le PR pour Power Ratio synthétisant les définitions existantes. Ensuite, nous proposons une étude théorique du PR en modulations monoporteuse en fonction du facteur de suréchantillonnage, puis du facteur de retombée d'un filtre de Nyquist. Des résultats de simulations sont présentés en modulations QPSK et MAQ 16.

Abstract – In the past few years, many research activities have concerned the Peak to Average Power Ratio (PAPR) especially in an Orthogonal Frequency Division Multiplexing (OFDM) context. Nevertheless, we have noticed that the PAPR definition is not always the same from one author to the other, depending on the paper context which leads to confusions and bad interpretations. That's why we propose in this article to generalize the PAPR definition by introducting the Power Ratio (PR) parameter from which we can derive all the possible PAPR versions. Then, we propose a theoretical analysis of the PR vs oversampling and the roll off factor in single carrier modulation when a Nyquist filter is considered. We give an upper bound of the PR and simulations for QPSK and 16QAM modulations.

1 Introduction

L'amplificateur de puissance est un des éléments clés d'un système de télécommunications, tout spécialement dans des contextes radiomobiles. Le rendement des amplificateurs est cependant faible (de l'ordre de 50%), leurs caractéristiques ne sont pas linéaires et une partie non négligeable (environ 60% dans le cas d'un terminal 2.5G) de la consommation est attribuée à l'amplificateur de puissance. Ainsi, l'utilisation de modulations à enveloppes constantes telles que la GMSK pour le GSM se justifie, optimisant ainsi l'efficacité de la fonction d'amplification en se plaçant au point de saturation, évitant ainsi tout traitement de non linéarités. Pour autant, la plupart des modulations standardisées aujourd'hui ne sont pas à enveloppes constantes, qu'elles soient mono ou multi porteuses et les conséquences d'une amplification de puissance non linéaire du signal radio fréquence doivent être examinées avec la plus grande attention.

Le facteur de crête est un paramètre défini comme le rapport entre la puissance instantanée et moyenne d'un signal donné et permettant de dimensionner ce dernier vis à vis de l'amplificateur de puissance (et vice versa). Ce paramètre apparait dans la littérature sous plusieurs acronymes tel que PAPR (Peak To Average Power Ratio), PMEPR (Peak to Mean Enveloppe Power Ratio) ou encore CF (Crest Factor). Devant cette multiplicité, il nous a paru nécessaire de synthétiser la notion de facteur de crête en introduisant celle du Power Ration (PR), plus générale et déclinable en différentes versions permettant de retrouver celles pré-citées. Cette approche fait l'objet de la première partie de cet article.

Dans la deuxième partie de cet article, nous étudions séparément l'influence du suréchantillonnage et du filtrage sur le PR. En effet, tous les étages d'une chaîne de communications modifient le signal instantané et donc la valeur du PR. Cette étude est menée dans un cas monoporteuse.

2 Définition du PR (Power Ratio)

2.1 Définitions théoriques

Plusieurs cas de figure doivent être considérés suivant l'état du signal : continu (désigné par **c**) ou échantillonné (désigné par **e**) et suivant le temps d'intégration utilisé dans le calcul du PR : fini (désigné par **f**) ou infini (désigné par **i**). Nous obtenons ainsi quatre états possibles pour le PR : $PR_{c,f}$, $PR_{c,i}$, $PR_{e,f}$ et $PR_{e,i}$ illustrés sur la figure 1 :

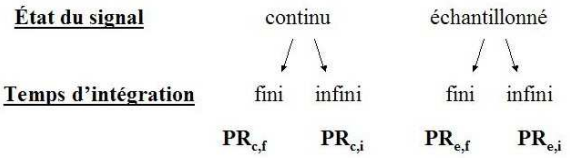


FIG. 1: Les différentes configurations du PR

2.1.1 Définition du PR infini

Nous présentons le PR infini dans le cas continu ($PR_{c,i}$) sachant que cette définition est valable aussi dans un cas échantillonné. Dans ces conditions, soit $s(t)$ un signal réel ou complexe, en bande de base ou radio fréquence. Ainsi, le $PR_{c,i}$ d'un signal $s(t)$ est :

$$\mathbf{PR}_{c,i}\{s(t)\} = \frac{Max|s(t)|^2}{E\{|s(t)|^2\}} = \frac{P_{max}}{P_{mean}}. \quad (1)$$

La variable t est réelle et $E\{.\}$ représente l'espérance mathématique. P_{max} est le maximum de la puissance instantanée, pris sur un intervalle de temps infini. Ce $PR_{c,i}$ est associé au signal présenté à l'entrée de l'amplificateur de puissance. Les autres définitions du PR ne sont que des approximations de celui-ci.

2.1.2 Définition du PR fini

Nous nous plaçons dans le cas d'un signal continu, sachant encore une fois que cette défintion s'applique aussi dans un cas échantillonné. $\boldsymbol{PR}_{c,f}$, pour un temps d'intégration T, est alors défini par :

$$\mathbf{PR}_{c,f}(T)\{s(t)\} = \frac{Max_{[0,T]}|s(t)|^2}{\frac{1}{T}\int_T |s(t)|^2 dt}. \quad (2)$$

Dans ces conditions, nous avons la relation :

$$\lim_{T\to+\infty} \mathbf{PR}_{c,f}(T) = \mathbf{PR}_{c,i}. \quad (3)$$

2.1.3 Définition du PR dans le cas échantillonné

Soit N le nombre d'échantillons du signal $s(k)$. Le PR d'un signal numérique $s(k)$ (k variable entière) est :

$$\mathbf{PR}_{e,f}(N)\{s(k)\} = \frac{Max_{k\in[0,N-1]}|s(k)|^2}{\frac{1}{N}\sum_{k=0}^{N-1}|s(k)|^2}. \quad (4)$$

Comme précédemment, nous avons la relation suivante pour le cas échantillonné :

$$\lim_{N\to\infty} \mathbf{PR}_{e,f}(N) = \mathbf{PR}_{e,i}, \quad (5)$$

avec

$$\mathbf{PR}_{e,i}\{s(k)\} = \frac{Max|s(k)|^2}{E\{|s(k)|^2\}}. \quad (6)$$

3 Liens entre le PR et les définitions usuelles du facteur de crête

Dans la suite, le signal considéré est supposé continu et le temps d'intégration infini.

3.1 Cas du PMEPR

Le $PMEPR$ (Peak to Mean Envelop Power Ratio) est lié à un signal en bande de base (en référence à la notion d'enveloppe complexe) [3]. En utilisant la notation $\widetilde{PR}$ pour spécifier que le signal est complexe, $PMEPR = \widetilde{PR}_{c,i}$.

3.2 Cas du PAPR

Le $PAPR$ (Peak to Average Power Ratio) est généralement utilisé pour définir le facteur de crête de signaux radio fréquence [2]. Soit $\widetilde{s}(t)$ le signal en bande de base et $f_0 = \frac{\omega_0}{2\pi}$ la fréquence porteuse. Le PAPR est alors défini par :

$$PAPR\{\widetilde{s}(t)\} = \frac{Max|Re(\widetilde{s}(t)e^{j\omega_0 t})|^2}{E\{|Re(\widetilde{s}(t)e^{j\omega_0 t})|^2\}}. \quad (7)$$

En utilisant la notation $\mathcal{PR}$ pour spécifier que le signal est radio fréquence, $PAPR = \mathcal{PR}_{c,i}$.

3.3 Relation entre le PMEPR et le PAPR

De nombreux articles ont proposé une relation entre le PAPR et le PMEPR, tout en donnant des résultats contradictoires. En effet, dans [2] ou [4] par exemple, les auteurs montrent que $PAPR \leq PMEPR$, et dans [5], il est montré que $PAPR \approx 2PMEPR$. Cette dernière approximation est correcte mais à notre sens imprécise. En effet, comme :

$$|Re(\widetilde{s}(t)e^{j\omega_0 t})|^2 \leq |\widetilde{s}(t)|^2 \quad (8)$$

et que la puissance moyenne en fréquence porteuse est la moitié de celle en bande de base, alors :

$$PAPR \leq 2PMEPR \quad (9)$$

L'égalité est obtenue lorque les puissances instantanées en radio fréquence et bande de base ont même maximum et au même instant. Ceci est en particulier réalisé lorsque $f_0 \ll \frac{1}{T_s}$ où T_s est le temps symbole. Cette condition est toujours vérifiée en télécommunications. C'est pourquoi l'on peut affirmer qu'en pratique, $PAPR \approx 2PMEPR$, soit $PAPR_{dB} = PMEPR_{dB} + 3dB$.

3.4 Cas du CF

Le facteur de crête (CF pour Crest Factor) peut être utilisé et est défini comme la racine carrée du PAPR [7] ou du PMEPR [6] suivant le cas d'étude :

- $CF = \sqrt{PAPR} = \sqrt{\mathcal{PR}_{c,i}}$ en fréquence porteuse
- $CF = \sqrt{PMEPR} = \sqrt{\widetilde{PR}_{c,i}}$ en bande de base

4 Analyse du PR en modulations monoporteuse

L'analyse du PR est effectuée selon deux axes : le premier consiste à étudier l'influence du facteur de suréchantillonnage

sur le PR ; le second consiste à prendre en compte l'effet du filtrage (ici de Nyquist) et de propose une borne supérieure du PR.

4.1 Influence du suréchantillonnage

En prenant comme exemple la modulation QPSK, le $PR_{e,i}$, calculé à la fréquence symbole vaut $0dB$ car les puissances instantanées et moyennes sont identiques. Cette valeur n'a cependant pas de sens dans un cas continu car elle ne prend pas en compte les valeurs des transitions entre symboles. En prenant ainsi un suréchantillonnage de 2, les constatations sont les suivantes :
- pendant une demie période, on se situe à P_{max}
- pendant l'autre demie période, on se situe à mi chemin entre 2 symboles avec :

- une chance sur 4 que les symboles soient les mêmes ($P = P_{max}$)
- une chance sur 2 que les symboles soient voisins ($P = \frac{P_{max}}{2}$)
- une chance sur 4 que les symboles soient diamétralement opposés ($P = 0$)

Dans ces conditions, le calcul de la puissance moyenne, pour une QPSK avec un facteur de suréchantillonnage de 2 donne :

$$P_{mean} = \frac{P_{max}}{2} + \frac{1}{2} * (\frac{P_{max}}{4} + \frac{P_{max}}{2*2} + \frac{0}{4}), \qquad (10)$$

soit $1.25dB$.
En effectuant le même raisonnement sur une QPSK pour un facteur N, on montre que l'on a :

- une probabilité de $\frac{1}{4}$ de rester sur le même symbole ($P_0 = \frac{P_{max}}{2}$)
- une probabilité de $\frac{1}{2}$ pour une transition d'un symbole à un symbole voisin avec une probabilité globale égale à :

$$P_1 = \frac{P_{max}}{2N} \sum_{i=0}^{N-1} \frac{(N/2-i)^2}{2*(N/2)^2} = \frac{1+2N^2}{6N^2} P_{max}. \quad (11)$$

- une probabilité de $\frac{1}{4}$ pour une transition d'un symbole à un autre diamétralement opposé avec une probabilité globale égale à :

$$P_2 = \frac{P_{max}}{4N} \sum_{i=0}^{N-1} \frac{(N/2-i)^2}{(N/2)^2} = \frac{2+N^2}{12N^2} P_{max}. \quad (12)$$

Le PR global est alors $\frac{3N^2}{1+2N^2}$ soit $1.76dB$ pour valeur limite quand N tend vers l'infini. Cette valeur est cependant sous estimée car elle ne prend pas en compte l'effet du filtrage. Ce point est développé dans la partie suivante.

4.2 Influence du filtre de Nyquist

Dans cette partie, on considère un filtre de Nyquist de facteur de retombée β :

$$p(t) = \frac{sin(\frac{\pi}{T_s}t)}{\frac{\pi}{T_s}t} \frac{cos(\frac{\pi\beta}{T_s}t)}{1-\frac{4\beta^2t^2}{T_s^2}}. \qquad (13)$$

T_s est le temps symbole et le $\widetilde{PR}_{c,i}$ sera analysé pour un temps continu (t réel) ; le signal considéré sera alors $\widetilde{s}(t)$ défini par :

$$\widetilde{s}(t) = \sum_{k=0}^{N-1} (a_k + jb_k) p(t - kT_s). \qquad (14)$$

L'objectif de cette partie est de majorer le $\widetilde{PR}_{c,i}$ de $\widetilde{s}(t)$ par le biais du calcul de la moyenne et du maximum de la puissance instantanée de $\widetilde{s}(t)$. Ces développements seront effectués en modulations BPSK et étendus ensuite à toutes sortes de modulations.

Analyse de la puissance moyenne en BPSK: en considérant $e(t) = \sum_{k=0}^{N-1} a_k \delta(t - kT_s)$ et *P(f)* la transformée de Fourier de $p(t)$, la puissance moyenne du signal $\widetilde{s}(t)$ est :

$$P_m = E\{|\widetilde{s}(t)|^2\} = \int_{-\infty}^{\infty} \gamma_e(f) P^2(f) df, \qquad (15)$$

En supposant que les symboles sont non corrélés et de moyennes nulles, $\gamma_e(f) = \frac{E\{a_k^2\}}{T_s} = \frac{\sigma_a^2}{T_s}$. On montre de plus que $\int_{-\infty}^{\infty} P^2(f) df = T_s(1 - \frac{\beta}{4})$

En conclusion, $P_m = \sigma_a^2(1 - \frac{\beta}{4})$.

Analyse du maximum de la puissance instantanée : pour que l'équation (24) soit valable, il faut étudier séparément les cas $\beta = 0$ et $\beta \neq 0$.
Pour $\beta = 0$, le signal pour un temps donné prend la forme d'une série dont il faut étudier la convergence. Si on considère $a_k = (-1)^k$ et $t = \frac{-T_s}{2}$, on obtient :

$$\widetilde{s}(t = -\frac{T_s}{2}) = \frac{2}{\pi} \sum_{k=0}^{N-1} \frac{1}{1+2k}, \qquad (16)$$

Cette série est divergente. Cela nous assure qu'il existe bien une configuration de symboles c_k et un instant donné pour lesquels le signal et donc sa puissance instantanée tendent vers l'infini, ce qui se traduit par :

$$\widetilde{PR}_{c,i}\{\widetilde{s}(t), \beta = 0\} = +\infty. \qquad (17)$$

Pour une meilleure lisibité, considérons $T_s = 1$. Pour $\beta \neq 0$, on montre que $|s(t)| \leq \frac{1}{|1-4\beta^2(t-k)^2|}$. La puissance instantanée est ainsi majorée par une série de Rieman qui converge.
On peut donc conclure que :

$$\widetilde{PR}_{c,i}\{\widetilde{s}(t), \beta \neq 0\} \neq +\infty. \qquad (18)$$

Analysons maintenant le maximum de $|\widetilde{s}(t)|^2$ pour $\beta \neq 0$. Dans ce cas,

$$\widetilde{s}(t) = \sum_{k=0}^{N-1} a_k \frac{sin(\pi(t-k))}{\pi(t-k)} \frac{cos(\pi\beta(t-k))}{1-4\beta^2(t-k)^2}. \qquad (19)$$

$|\widetilde{s}(t)|^2$ peut alors être majoré par :

$$\sum_{k=0}^{N-1}(\frac{sin(\pi(t-k))}{\pi(t-k)})^2 \sum_{k=0}^{N-1}(\frac{cos(\pi\beta(t-k))}{1-4\beta^2(t-k)^2})^2. \qquad (20)$$

Le premier terme est majoré par 1. Analysons le second terme par le biais de la fonction $\varphi(t) = \sum_{k=0}^{N-1}(\frac{cos(\pi\beta(t-k))}{1-4\beta^2(t-k)^2})^2$. On montre aisément que φ est maximum en $t_n = \frac{N}{2}$. Ceci est illustré sur la figure 2 où la courbe supérieure est la somme des N fonctions $(\frac{cos(\pi\beta(t-k))}{1-4\beta^2(t-k)^2})^2$ décalées.

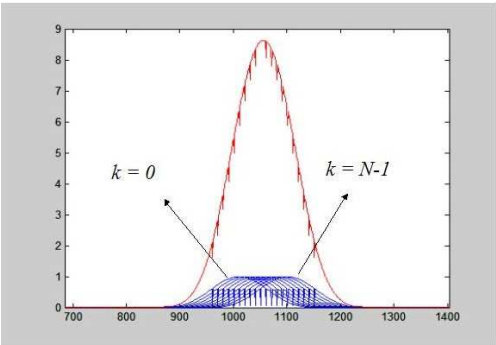


FIG. 2: Fonction $\varphi(t)$

Pour $t = \frac{N}{2}$, $\varphi(t = \frac{N}{2}) = 1 + 2\sum_{k=1}^{N/2}(\frac{cos(\pi\beta k)}{1-4\beta^2k^2})^2$. En montrant que

$$\lim_{n\to\infty}\int_0^n (\frac{cos(\pi\beta t)}{1-4\beta^2t^2})^2 dt = \frac{\pi^2}{16\beta} \qquad \beta \in]0,1], \qquad (21)$$

on en déduit que $|\widetilde{s}(t)|^2 \leq \frac{\pi^2}{8\beta}$ et donc

$$\widetilde{PR}_{c,i}\{\widetilde{s}(t)\}_{BPSK} \leq \frac{\pi^2/8\beta}{(1-\beta/4)}. \qquad (22)$$

Cas des autres modulations : en considérant que les symboles a_k et b_k ont la même statistique, on montre que dans le cas général d'une modulation MAQ à M états,

$$\widetilde{PR}_{c,i}\{\widetilde{s}(t)\}_{QAM} \leq (\frac{Max(a_k)}{\sigma_a})^2 \frac{\pi^2/8\beta}{(1-\beta/4)}. \qquad (23)$$

Résultats de simulations : la figure 3 compare les majorations obtenues en modulations BPSK, QPSK et MAQ 16 avec des simulations pour quatre valeurs différentes de β (0.9, 0.6, 0.3 et 0.1). Pour des valeurs de β proches de 1, les simulations ont été menées en générant 10^4 signaux de $N = 10^3$ symboles. Pour ces valeurs de β, les densités de probabilités de $\widetilde{PR}_{c,i}$ sont très peu étalées, donnant ainsi des valeurs de facteur de crête très proches les unes des autres. Par contre pour des valeurs de β proches de 0, du fait de la divergence de la série décrite au paragraphe 4.2, les valeurs de $\widetilde{PR}_{c,i}$ sont beaucoup plus dispersées et cela nécessite de générer 10^6 signaux de $N = 10^5$ symboles.

Ces résultats montrent ainsi les très fortes adéquations entre les majorations théoriques proposées et les simulations.

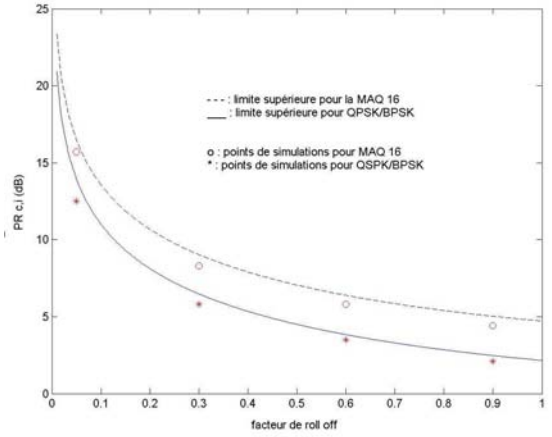


FIG. 3: Le $PR_{c,i}$ en fonction de β

5 Conclusion

L'étude théorique menée montre que le PR est fortement influencé par le suréchantillonnage et le filtrage. Dans un contexte monoporteuse, une borne supérieure très fine du PR avec un filtre de Nyquist est proposée. Des travaux sont en cours pour étendre ces résultats à un filtre en racine de Nyquist pour formaliser de façon plus générale l'influence du filtrage sur le PR, aussi bien en mono qu'en multiporteuses.

References

[1] Nati Dinur, Dov Wulich, "Peak-to-Average Power Ratioin High-Order OFDM", IEEE Transactions on Communications, Vol 49, No.6, June 2001, pp 1063-1072.

[2] M. Sharif, M. Gharavi-Alkhansari and B.H. Khalaj, "On the Peak-to-Average Power of OFDM Signal Based on Oversampling", IEEE Transactions on Communications, Vol 51, No 1, January 2003, pp 72-78.

[3] M. Sharif and B.H. Khalaj, "Peak to Mean Envelope Power Ratio of oversampled OFDM Signals: an analytical approach", Proc. of IEEE Int. Conf. Communications, Vol 5, Helsinki, Finland, June 2001, pp 1476-1480.

[4] "A theoretical framework for the peak-to-average power control problem in OFDM transmission" Gerhard Wunder, PhD from Berlin University, 2004.

[5] "Peak to average power reduction for multicarrier modulation", Jose Tellado, Stanford University, Septembre 1999.

[6] M.Friese, "Multitone Signals with Low Crest Factor", IEEE Trans. On Communications, Vol 45, No 10, October 1997, pp 1338-1344.

[7] B.M. Popovic, "Synthesis of Power Efficient Multitone Signals with Flat Amplitude Spectrum", IEEE Transactions on Communications, Vol 39, No 7, July 1991, pp 1031-1033. 2002.

TEB de système OFDM codé multi-utilisateurs

Adrien RENOULT[1,2], Inbar FILAKOW[1], Marc CHENU-TOURNIER[2]

[1]ETIS/ENSEA - Univ de Cergy-Pontoise - CNRS
6 av du Ponceau, 95014 Cergy-Pontoise France

[2]Thales Communication
160 bld de Valmy BP 92, 92704 Colombes Cedex, France
adrien.renoult@fr.thalesgroup.com, renoult@ensea.fr
fijalkow@ensea.fr
marc.chenu@fr.thalesgroup.com

Résumé – Le but de cet article est de présenter les performances analytiques d'un système OFDM codé multi-utilisateurs sans méthode d'accès multiple spécifique. Nous déterminerons les performances du système à la sortie de décodeur canal et à la sortie du détecteur de symboles. Le calcul est effectué de manière exacte dans le cas mono-utilisateur ou lorsque nous prenons les performances en sortie de détecteur de symboles, mais dans le cas multi-utilisateurs en sortie de décodeur canal une approximation est nécessaire. La valadité de l'analyse est montrée par simulations.

Abstract – The BER of a multi-users multi-sensors array coded Orthogonal Frequency Division Multiplexing (OFDM) system is investigated. We consider an almost optimal receiver with a soft output symbol detector and a soft input Viterbi channel decoder. We use the Viterbi bound and extend it to a multi-users case. We then obtain the performance conditionally to the channel and we integrate this performance over the different channel realisations. Finally the validity of our analysis is shown by simulations.

1 Introduction

Les systèmes à modulation multi-porteuses de type Orthogonal Frequency Division Multiplexing (OFDM) ont été retenus pour les systèmes de transmission haut-débits. En effet, cette modulation permet une égalisation du canal très simple. Différents systèmes d'accès multiple peuvent être associés à cette modulation. Nous allons dans cet article étudier un système d'accès multiple sans séparation en code ou en fréquence. Nous ferons la détection des symboles envoyés par les différents utilisateurs en utilisant un critère au sens du maximum à posteriori afin de passer le plus d'information possible au décodeur de Viterbi.

A cause de la complexité d'un tel système il devient temporellement difficile de simuler ses performances dès que le nombre d'antennes et d'émitteurs dépasse trois ou quatre. Nous allons donc tenter en isolant un petit nombre de paramètres significatifs de dériver les performances analytiques du système. Ceci nous permettra ainsi d'obtenir une borne pour comparer les performances des systèmes utilisant des récepteurs sous-optimaux.

A notre connaissance aucune publication ne présente les performances d'un système OFDM codé multi-utilisateurs complet. Dans un article précédant [3], nous avions modélisé les Log Rapport de Vraisemblance (LLRs) en sortie d'estimateur de symboles comme des processus Gaussien consistant et simulé les performances de l'algorithme de Viterbi lorsque nous lui rentrions ces LLRs. Cependant deux limitations étaient présentes dans cet article, premièrement nous n'obtenions pas une formule analytique des Taux d'Erreur Binaire (TEB) et de plus nous ne prenions pas en compte les interférences entre les différents utilisateurs ce qui amenait un décalage entre les performances analytiques et les simulations.

Dans ce papier nous étudierons le TEB en sortie de décodeur canal. Ce TEB sera exprimé en fonction des caractéristiques du système (nombre d'utilisateurs, nombre de capteurs à l'antenne de réception , type de code convolutif) et du Rapport Signal sur Bruit (RSB).

Dans une première partie, nous présenterons le modèle utilisé ensuite nous présenterons l'évènement d'erreur. Les parties trois et quatre comporterons la dérivation des performances. Ces performances étant validées par simulation. Enfin la section 5 conclura ce papier.

2 Modèle

Nous allons utiliser dans cet article un modèle OFDM multi-utilisateurs dans le domaine fréquentiel de la forme :

$$\mathbf{y}(k,n) = \sum_{t=1}^{N_T} \mathbf{h}_t(k,n)\, x_t(k,n) + \mathbf{w}(k,n)$$

où $\mathbf{y}(k,n)$ est le vecteur contenant les N_R échantillons reçus par les N_R capteurs de réception, $x_t(k,n)$ est le symbole envoyé par l'utilisateur t. $\mathbf{h}_t(k,n)$ est le canal entre l'utilisateur t et les différents capteurs du récepteur. Enfin $\mathbf{w}(k,n)$ est le vecteur de bruit additif blanc Gaussien associé aux différents capteurs de réception. Tous ces éléments étant définis pour la sous-porteuse k et le symbole OFDM n.

Nous supposerons dans la suite de l'article une connaissance parfaite des coefficients du canal. Pour la détection des symboles nous utiliserons l'algorithme du Maximum A Posteriori (MAP). Pour le décodage canal nous utiliserons l'algorithme de Viterbi. Enfin afin de rester cohérent lorsque nous ferrons varier le nombre d'utilisateur de notre système nous appliquerons le coefficient de pondération $\frac{1}{\sqrt{N_T}}$ aux coefficients du canal afin de garder une puissance d'émission constante.

3 Performance à la sortie du décodeur de Viterbi

Dans cette section nous définissons le TEB d'un utilisateur et nous dérivons ces performances en moyennant sur les différentes réalisations du canal.

3.1 Formulation du problème

Les différents utilisateurs de notre système ayant les mêmes performances moyennes, nous allons considérer les performances de l'utilisateur 1. De même nous considérons que les symboles envoyés sont une séquence de 1 car les performances ne dépendent pas des symboles émis. Nous utiliserons la borne de Viterbi [4] :

$$P_b \leq \sum_{d=d_{free}}^{\infty} c_d P_d \tag{1}$$

Dans cette équation P_d représente la probabilité de choisir un chemin faux de distance de Hamming d au chemin correct. c_d représente la moyenne du nombre de bits faux pour un chemin de distance de Hamming d.

Le calcul de P_d ce fait par intégration sur les différents éléments du canal :

$$P_d = \int P_{d|\mathbf{H}} p\left(\mathbf{H}\right) d\mathbf{H}$$

où $P_{d|\mathbf{H}}$ est la probabilité de mauvais choix sachant la réalisation du canal $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_{N_T}]$, et $p\left(\mathbf{H}\right)$ est la probabilité d'avoir le canal $\mathbf{H}$.

3.2 Vraisemblance d'un chemin faux sachant le canal

Un chemin faux (indexé par i) de distance de Hamming d du chemin correct (indexé par j), sur le treillis du décodeur de Viterbi et pour le premier utilisateur est défini par le fait que sa métrique est plus petite que la métrique du bon chemin :

$$\sum_{(k,n)\in\mathcal{C}_d} \|\mathbf{y}(k,n) - \sum_{t=1}^{N_T} \mathbf{h}_t(k,n) x_t^i(k,n)\|^2 < \sum_{(k,n)\in\mathcal{C}_d} \|\mathbf{y}(k,n) - \sum_{t=1}^{N_T} \mathbf{h}_t(k,n) x_t^j(k,n)\|^2 \tag{2}$$

où $\mathcal{C}_d$ est définis un l'ensemble des d états erronés a travers le treillis pour le premier utilisateur et ayant n'importe quel nombre d'erreurs pour les autres utilisateurs. Nous ne pouvons pas restreindre $\mathcal{C}_d$ aux seuls trajets avec exactement d erreurs pour l'utilisateur d'intérêt et un décodage correct pour les autres utilisateurs, car comme dans [3] ceci reviendrai à négliger l'interférence entre symboles.

La probabilité de choisir un mauvais chemin en faisant une erreur de d peut être bornée par la somme des probabilités de couple (i,j) avec i un chemin à une distance de Hamming de d pour l'utilisateur d'intérêt. Ceci revient à utiliser la borne de l'union :

$$P_{d/\mathbf{H}} \leq \sum_{i\in\mathcal{I}_d} Prob\left(\sum_{(k,n)\in\mathcal{C}_d} \|\mathbf{y}(k,n) - \sum_{t=1}^{N_T} \mathbf{h}_t(k,n) x_t^i(k,n)\|^2 < \sum_{(k,n)\in\mathcal{C}_d} \|\mathbf{y}(k,n) - \sum_{t=1}^{N_T} \mathbf{h}_t(k,n) x_t^j(k,n)\|^2 \right) \tag{3}$$

où $\mathcal{I}_d$ est l'ensemble des chemins faux de distance de Hamming d pour le premier utilisateur.

Dans le cas d'un système mono-utilisateur (3) deviens une égalité et la sommation sur tous les trajets i ce réduit à l'influence d'un seul trajet de distance de Hamming d. Ceci expliquera dans la suite l'analyse plus précise dans le cas mono-utilisateur.

3.3 Vraisemblance d'un chemin faux

Nous noterons

$$D_{i,j} = \sum_{(k,n)\in\mathcal{C}_d} \|\mathbf{y}(k,n) - \sum_{t=1}^{N_T} \mathbf{h}_t(k,n) x_t^i(k,n)\|^2 - \|\mathbf{y}(k,n) - \sum_{t=1}^{N_T} \mathbf{h}_t(k,n) x_t^j(k,n)\|^2$$

la différence de métrique entre un mauvais chemin i et le bon chemin j. La probabilité de l'évènement d'erreur de distance de Hamming d sera donc égale à la probabilité d'avoir une métrique $D_{i,j}$ négative. Comme dans [1], dans le cas d'une détection conjointe MIMO, nous pouvons exprimer $D_{i,j}$ comme la somme de dN_R termes :

$$D_{i,j} = \sum_{(k,n)\in\mathcal{C}_d} \sum_{r=1}^{N_R} \mathbf{z}_r^{\dagger}(k,n) \mathbf{F}_{i,j}(n,k) \mathbf{z}_r(k,n) \tag{4}$$

où $\mathbf{z}_r(k,n) = (y_r(k,n), h_{1,r}(n,k), \cdots, h_{N_T,r}(n,k))^t$ et $\mathbf{F}_{i,j}(n,k)$ la matrice de rang deux définis par $\mathbf{u}_i(n,k)\mathbf{u}_i(n,k)^{\dagger} - \mathbf{u}_j(n,k)\mathbf{u}_j(n,k)^{\dagger}$ où $\mathbf{u}_i(n,k) = (1, -x_1^i(n,k), \cdots, -x_{N_T}^i(n,k))^t$. Nous supposerons sans perte de généralité que le chemin correct est le chemin correspond à la séquence de bits tout à zéro ou de façon équivalente la séquence de symboles tout à un. Nous voulons maintenant moyenner sur toutes les réalisations du canal la probabilité de choisir le chemin faux i.

$$P_i = \int Prob(D_{i,j} < 0) p(\mathbf{H}) d\mathbf{H} \tag{5}$$

cette expression peut ce développer comme dans [2] en utilisant les valeurs propres de $\mathbf{R}\mathbf{F}_{i,j}(n,k)$ ($\mathbf{R}$ étant définis

comme $\frac{1}{2}E[\mathbf{z}_r(k,n)\mathbf{z}_r^\dagger(k,n)|\mathbf{x}_j])$. Le résultat de [2] peut être appliqué si $\mathbf{F}_{i,j}(n,k)$ n'est une fonction de i seulement. Dans notre cas, ceci sera vrai sous la condition suffisante suivante : pour un chemin faux i les utilisateurs de 2 à N_T seront soit entièrement correcte soit entièrement juste pour les indices $(k,n) \in \mathcal{C}_d$. Comme certains trajets de $\mathcal{I}_d$ ne suivent pas cette condition nous devrons restreindre la sommation de l'équation (3). Même si cette condition nécessaire n'est pas suffisante elle reste plus fine que de négliger intégralement l'interférence entre utilisateurs comme dans [3]. Nous obtenons alors le résultat suivant :

Lemme 1 *Le i^{th} terme de la partie droite de l'équation (3) peut être approximé par :*

$$P_i = \frac{1}{(1-\mu_i)^{2dN_R-1}} \sum_{r=0}^{dN_R-1} \left(\begin{array}{c} 2dN_R-1 \\ r \end{array} \right) (-\mu_i)^r \quad (6)$$

avec $\mu_i = \frac{a_i\gamma+\sqrt{a_i^2\gamma^2+2a_i\gamma}}{a_i\gamma-\sqrt{a_i^2\gamma^2+2a_i\gamma}}, a_i = \sum_{t=1}^{N_T}(1-x_t^i)$, $\gamma = \frac{E_b}{N_0}$ *et* $i \in \mathcal{J}_d$. x_t^i *est le symbole de l'utilisateur* t *pour le chemin* i.

3.4 Probabilité d'erreur à la sortie du décodeur canal

En utilisant l'équation (6), qui nous donne la probabilité de choisir un chemin faux i de distance de Hamming d, nous pouvons trouver la probabilité de choisir un faux chemin et que ce chemin soit à la distance de Hamming d :

$$P_d \leq \sum_{i\in\mathcal{J}_d} \frac{1}{(1-\mu_i)^{2dN_R-1}} \sum_{r=0}^{dN_R-1} \left(\begin{array}{c} 2dN_R-1 \\ r \end{array} \right) (-\mu_i)^r \quad (7)$$

En utilisant alors l'équation (1), nous obtenons alors l'expression de la probabilité d'erreur à la sortie du décodeur.

$$P_b \leq \sum_{d=d_{free}}^{\infty} c_d \sum_{i\in\mathcal{J}_d} \frac{1}{(1-\mu_i)^{2dN_R-1}} \sum_{r=0}^{dN_R-1} \left(\begin{array}{c} 2dN_R-1 \\ r \end{array} \right) (-\mu_i)^r \quad (8)$$

4 Résultats de simulations

Nous présenterons dans cette section les résultats de simulation et leur comparaison avec les performances analytiques. Dans tous les cas le signal émis par les différents utilisateurs sera codé en utilisant un code convolutif non systématique non récursif de polynôme générateur $(5,7)$. Ceci nous donne un code de distance libre 3 et nous utiliserons les 20 premiers coefficients de c_d. Les courbes sont moyennées sur le canal et le bruit en utilisant une méthode de Monté-Carlo. Les coefficients du canal sont supposés parfaitement connus au récepteur.

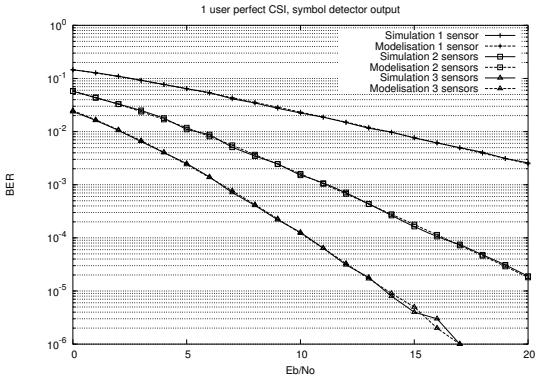


FIG. 1 – TEB à la sortie du détecteur de symboles, $N_T = 1$, présentant l'égalité entre les simulations et les performances analytiques du modèle (7) avec $d = 1$ dans un cas mono-utilisateur multi-capteurs.

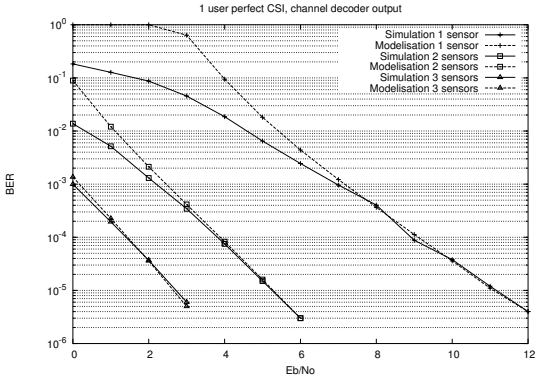


FIG. 2 – TEB à la sortie du décodeur canal, $N_T = 1$, présentant l'égalité entre les simulations et les performances analytiques du modèle (8) dans une cas mono-utilisateur multi-récepteurs.

4.1 Cas mono-utilisateur

Durant ces simulations le nombre d'utilisateurs est fixé à 1 et le nombre de capteurs varis de 1 à 3. La figure 1 montre la grande adéquation entre les performances analytiques et les performances simulés. A la sortie de détecteur de symboles. La figure 2 présente le cas en sortie de décodeur canal.

4.2 Cas multi-utilisateurs

Dans le cas multi-utilisateurs nous utilisons des systèmes avec le même nombre de récepteurs que d'émetteurs. La figure 3 présente les résultats en sortie de détecteur de symboles. Les résultats de la simulation et des performances analytiques sont identiques. Cependant dans le cas de la figure 4 présentant les résultats en sortie de décodeur canal les performances analytiques présentent un écart. Ce écart provient probablement de l'hypothèse que nous avons du faire pour dériver les performances du système.

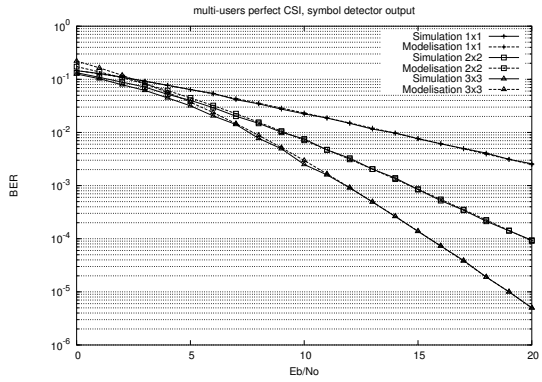


FIG. 3 – Mise en avant de la forte adéquation entre les performances analytiques en sortie de détecteur de symboles (équation (7) with $d = 1$) et les simulations dans le cas multi-utilisateurs multi-capteurs.

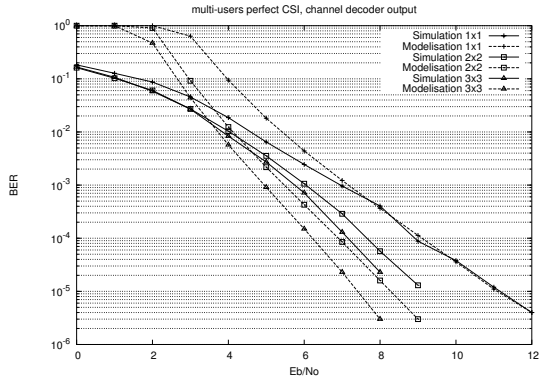


FIG. 4 – Mise en avant de l'écart entre les performances analytiques en sortie de decodeur canal (équation (8)) et les simulations dans un cas multi-utilisateurs multi-capteurs.

5 Conclusion

Dans cet article nous avons dérivé les performances d'un système OFDM multi-utilisateurs codé. Le critère utilisé pour caractériser les performances a été le TEB. Les performances ont été calculées en sortie de détecteur de symboles et en sortie de décodeur canal. Comme les performances ne dépendent ni de l'utilisateur ni de la séquence envoyée nous avons choisi de les calculer pour le premier utilisateur et une séquence de symboles émise constitué de un.

Nous avons utilisé la borne de Viterbi. Cette dernière exprime le résultat comme une sommation des probabilités de choisir un mauvais chemin de distance de Hamming d.

Nous avons dérivé la probabilité P_d de choisir un mauvais chemin à la distance de Hamming d du chemin correct et pour ce faire nous avons du faire l'approximation que le chemin choisis par les utilisateurs de 2 à N_T était soit entièrement juste soit entièrement faux. Nous aboutissons ainsi à une formule analytique complète.

Dans la dernière partie de notre article nous avons comparé les performances analytiques trouvées avec les simulations du système. Nous obtenons une très bonne adéquation entre les performances analytiques et les simulations pour le cas mono-utilisateur ou en sortie de détecteur de symboles. Cependant dans le cas multi-utilisateurs en sortie de décodeur canal les performances analytiques présentent un écart avec les simulations. Ce écart est probablement dût à l'approximation dans la borne de Viterbi.

Références

[1] Stephen J. Grant and James K. Cavers. Performance enhancement through joint detection of cochannel signals using diversity arrays. *IEEE Transactions on Communications*, Volume 46(8) :Pages 1038–1049, August 1998.

[2] Stephen J. Grant and James K. Cavers. Further analytical on the joint detection of cochannel signals using diversity arrays. *IEEE Transactions on Communications*, Volume 48(11) :Pages 1788–1792, November 2000.

[3] Adrien Renoult, Charly Poulliat, Marc Chenu-Tournier, and Inbar Fijalkow. On the performance of a multi-user multi-array cofdm system. In *proceeding of IEEE SPAWC*, July 2004.

[4] Andrew J. Viterbi. Convolutional codes and their performance in communication system. *IEEE Transactions on communications technology*, Volume 19(5) :Pages 751–772, October 1971.

Algorithmes à norme constante pour les systèmes de communication MIMO

A. IKHLEF [1], D. LE GUENNEC [1], J. PALICOT [1]

[1] IETR / Supelec-Campus de Rennes, Avenue de la Boulaie, BP 81127, 35511 Cesson-Sévigné, France

aissa.ikhlef@supelec.fr, daniel.leguennec@supelec.fr, jacques.palicot@supelec.fr

Résumé – Un nouvel algorithme de séparation aveugle de sources pour les systèmes MIMO (Multiple-Input Multiple-Output) de type BLAST (Bell Labs Layered Space-Time) est proposé, basé sur un critère de norme constante CN (Constant Norm), associé à la procédure d'orthogonalisation de Gram-Schmidt afin d'assurer l'indépendance des sorties de l'égaliseur. De cette approche deux nouveaux algorithmes sont déduits. Le premier appelé CQA (Constant sQuare Algorithm), est mieux adapté pour les modulations QAM que le classique CMA (Constant Modulus Algorithm), il fournit un niveau de bruit plus faible avec une complexité comparable. Le second est une pondération entre le CMA et le CQA pour tirer avantage des deux. Le coefficient de pondération est évalué dynamiquement, d'où son nom de CDNA (Constant Dynamic Norm Algorithm). Les algorithmes proposés reposent sur la minimisation d'une fonction de coût, construite à partir de normes, par un gradient stochastique sous la contrainte d'orthogonalité. En simulation, ces algorithmes montrent de meilleures performances comparés aux algorithmes CMA et MUK avec une complexité comparable. Le CQA atteint un meilleur état permanent que le CMA (gain de 3 dB) et le CDNA tend vers le meilleur algorithme dynamiquement entre le CMA et le CQA.

Abstract *– In this paper we present a new algorithm for blind source separation (BSS) based on the Constant Norm (CN) criterion for Multiple-Input Multiple-Output (MIMO) communication systems. The treated problem consists in blindly recovering (i.e. without the use of training sequences) the signals transmitted over a linear MIMO memoryless system, which introduces only Inter-User Interference (IUI). From the proposed algorithm, we deduce two other new algorithms designed especially for QAM modulation. The first one is named Constant sQuare Algorithm (CQA) and the second one, which is a weighting between the Constant Modulus Algorithm (CMA) and the CQA to get the advantages of both, is named Constant Dynamic Norm Algorithm (CDNA). At each iteration, the algorithms combine a stochastic gradient update and a Gram-Schmidt orthogonalization procedure. The simulation results show that the proposed algorithms have better performances compared to CMA and Multiuser Kurtosis Algorithm (MUK) with comparable complexity.*

1. Introduction

Le problème de la séparation aveugle de sources, apparu il y a vingt ans, est rencontré dans une large variété d'applications en traitement du signal comme par exemple le traitement de la parole, l'analyse sismique, les applications médicales (EEG) [1] et d'autres encore. Depuis dix ans l'application de cette technique en communications numériques a reçu un intérêt croissant, comme dans [2] pour les communications multi-utilisateurs (CDMA) modélisées comme un système MIMO. Dans la littérature nous trouvons aussi les algorithmes de type MUK [3], basés sur les Kurtosis, qui sont une extension de l'algorithme proposé initialement pour l'égalisation aveugle de systèmes SISO présenté dans [4]. Plus récemment une technique de séparation aveugle de sources a été appliquée au système de communications BLAST (Bell Labs Space-Time) construite sur l'algorithme multi-modules MMA (multimodulus algorithm) [5].

Nous présentons ici un nouvel algorithme de séparation aveugle de sources pour les systèmes MIMO (Multiple-Input Multiple-Output) de type BLAST. Il est basé sur le critère de la norme constante CN (Constant Norm) introduit récemment par [6], associé à la procédure d'orthogonalisation de Gram-Schmidt. Nous déduisons ensuite de cette approche deux nouveaux algorithmes. Le premier appelé CQA (Constant sQuare Algorithm), est mieux adapté pour les modulations QAM que le classique CMA (Constant Modulus Algorithm), il fournit un niveau de bruit plus faible avec une complexité comparable. Le second est une pondération entre le CMA et le CQA pour tirer avantage des deux. Le coefficient de pondération est évalué dynamiquement, d'où son nom de CDNA (Constant Dynamic Norm Algorithm). Les algorithmes proposés reposent sur la minimisation d'une fonction de coût, construite à partir de normes, par un gradient stochastique sous une contrainte d'orthogonalité afin d'assurer l'indépendance des sorties de l'égaliseur à chaque itération.

Le reste du papier est organisé de la façon suivante : formulation du problème et hypothèses sont introduites en section 2. En section 3, nous présentons notre nouvel algorithme CNA. Les deux algorithmes dérivés CQA et CDNA sont développés en section 4 et 5 respectivement. Les performances des algorithmes proposés et des algorithmes CMA et MUK sont comparées en section 6. Enfin une conclusion est donnée en section 7.

2. Formulation du problème

On considère un système de transmission MIMO comprenant respectivement n_T et n_R antennes en émission et en réception (voir figure 1). Le canal est supposé linéaire, sans mémoire, introduisant uniquement de l'interférence

inter-antennes, plus communément appelée IUI pour « Inter User Interference ».

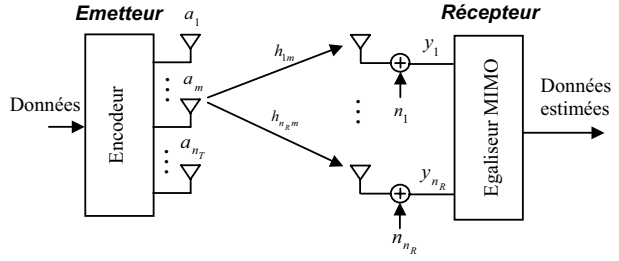


Figure 1. Modèle d'un système MIMO

Le signal reçu prend alors la forme générale :

$$\mathrm{Y}(k) = \mathbf{H}\,\mathrm{A}(k) + \mathrm{n}(k) \tag{1}$$

où $\mathrm{A}(k) = [a_1(k),...,a_{n_T}(k)]^T$ est le vecteur $(n_T \times 1)$ des signaux émis, $\mathbf{H}$ représente la matrice $(n_R \times n_T)$ de canal instantané, $\mathrm{Y}(k)$ est le vecteur $(n_R \times 1)$ des signaux reçus et $\mathrm{n}(k)$ est le vecteur $(n_R \times 1)$ de bruit.

Nous considérons les hypothèses suivantes :

- $n_R \geq n_T$,
- $\mathbf{H}$ est de rang plein n_T, avec des éléments complexes indépendants et identiquement distribuées (i.i.d.), de moyenne nulle et de variance unité.
- Le bruit est additif, blanc gaussien, de moyenne nulle, indépendant des signaux source, et de matrice de covariance : $\mathrm{R_n} = E[\mathrm{n\,n}^H] = \sigma_{\mathrm{n}}^2 \mathbf{I}_{n_R}$ où $\mathbf{I}_{n_R}$ représente la matrice identité $(n_R \times n_R)$.
- Les signaux source sont des séquences à temps discret, centrées, indépendantes et identiquement distribuées (i.i.d), de matrice de covariance $\mathrm{R_A} = E[\mathrm{A\,A}^H] = \mathbf{I}_{n_T}$.

Pour récupérer les signaux transmis, le signal reçu est filtré par une matrice $\mathbf{W}(k)$, $(n_R \times n_T)$, appelée matrice d'égalisation. La sortie de l'égaliseur devient alors :

$$\begin{aligned} \mathrm{Z}(k) &= \mathbf{W}^T(k)\,\mathrm{Y}(k) = \mathbf{W}^T(k)\,\mathbf{H}\,\mathrm{A}(k) + \overline{\mathrm{n}}(k) \\ &= \mathbf{G}(k)\,\mathrm{A}(k) + \overline{\mathrm{n}}(k) \end{aligned} \tag{2}$$

où $\mathrm{Z}(k)$ est le vecteur $(n_T \times 1)$ de sortie,

$\mathbf{G}(k) = \mathbf{W}^T(k)\,\mathbf{H}$ est la matrice globale $(n_T \times n_T)$ canal-égaliseur et $\overline{\mathrm{n}}(k)$ le bruit filtré en sortie du récepteur.

La séparation des signaux source est obtenue à une permutation près et à un déphasage près [3].

3. Algorithme MIMO-CNA

Pour réaliser la séparation, nous proposons un algorithme appelé (MIMO-CNA), basé sur le critère CN [6]. Le critère (MIMO-CNA) à minimiser s'écrit :

$$J_{MIMO-CNA}(\mathbf{W}) = \sum_{i=1}^{n_T} E[\mathcal{N}(z_i)^P - R]^q \tag{3}$$

où $\mathcal{N}(.)$ est une norme sur $\Re^2$ et R est une constante donnée par : $R = \dfrac{E[\mathcal{N}^{2p}(a)]}{E[\mathcal{N}^p(a)]}$, pour $q = 2$.

Dans la suite nous considérons uniquement le cas où : $p = q = 2$.

Afin d'éviter une convergence vers un ensemble de signaux en sortie contenant plusieurs fois la même source, une contrainte est rajoutée au critère :

$$\begin{aligned} &\min_{\mathbf{W}} \; J_{MIMO-CNA}(\mathbf{W}) = \sum_{i=1}^{n_T} E[\mathcal{N}(z_i)^2 - R]^2 \\ &\text{Sous} \quad \mathbf{G}^H\mathbf{G} = \mathbf{I}_{n_T} \end{aligned} \tag{4}$$

La contrainte est issue de la condition $E[\mathrm{Z\,Z}^H] = \mathbf{I}_{n_T}$ pénalisant l'extraction du même signal sur plusieurs sorties. Pour satisfaire cette contrainte, la matrice de canal $\mathbf{H}$ est supposée être unitaire, la contrainte se ramène alors à :

$$\mathbf{W}^H(k+1)\,\mathbf{W}(k+1) = \mathbf{I}_{n_T}$$

Quand la matrice de canal n'est pas unitaire, une étape de pré-blanchiment est réalisée simplement par décomposition en valeurs propres.

Pour minimiser (4) nous utilisons un algorithme du gradient stochastique associée à une orthogonalisation de Gram-Schmidt [3] pour satisfaire la contrainte.

Lorsque la norme représente le module, c.à.d. $\mathcal{N}(.) = |.|$, nous retrouvons l'algorithme classique CMA, conçu initialement pour des signaux PSK.

L'utilisation du CMA pour des signaux QAM est possible [7], mais l'algorithme de descente génère alors un niveau de bruit significatif. C'est pour cette raison que nous proposons ci-dessous un nouvel algorithme spécialement conçu pour les signaux QAM.

4. Algorithme MIMO-CQA

De l'algorithme précédent, nous déduisons, pour le contexte MIMO, un algorithme CQA (Constant sQuare Algorithm), mieux adapté aux signaux QAM.

Les points de la constellation QAM sont en moyenne plus proches d'un carré que du cercle donné par le CMA (voir figure 2). L'idée est donc, plutôt que de contraindre les sorties de l'égaliseur à être sur un cercle comme dans le CMA, de les contraindre à être sur un carré. Le bruit résiduel est alors plus faible [6].

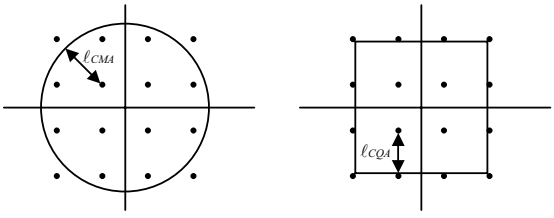


Figure 2. Principe du CMA et du CQA

Pour ceci, nous utilisons la norme infinie donnée par : $\|z_i\|_\infty = Max\left(|\Re z_i|,|\Im z_i|\right)$ où $\Re z_i$ et $\Im z_i$ sont respectivement les parties réelle et imaginaire de z_i. La fonction de coût (4) s'écrit alors:

$$\min_{\mathbf{W}} J_{MIMO-CQA}(\mathbf{W}) = \sum_{i=1}^{n_T} E\left[\|z_i\|_\infty^2 - R\right]^2$$
$$\text{sous} \ : \mathbf{G}^H\mathbf{G} = \mathbf{I}_{n_T}$$

où R est donné par: $R = \dfrac{E\left[\|z\|_\infty^4\right]}{E\left[\|z\|_\infty^2\right]}$.

L'algorithme CQA force l'égaliseur à donner en sortie une constellation qui corresponde à celle des signaux sources à une rotation de phase près, modulo π/2.

Après calculs, nous obtenons pour l'algorithme du gradient stochastique :

$$\overline{\mathbf{W}}(k+1) = \mathbf{W}(k) - \mu\,[\Delta_1(k)\ \ldots\ \Delta_{n_T}(k)]\,\mathrm{Y}^*(k)$$

où $\Delta_i(k) = \left(\|z_i(k)\|_\infty^2 - R\right)\|z_i(k)\|_\infty F(z_i(k))$

et $F(z_i) = \begin{cases} \mathrm{sgn}(\Re z_i), & \text{si } |\Re z_i| > |\Im z_i| \\ j.\mathrm{sgn}(\Im z_i), & \text{sinon} \end{cases}$

où *sgn* est la fonction signe et $j = \sqrt{-1}$.

Pour satisfaire la contrainte, nous appliquons une orthogonalisation de Gram-Schmidt à $\mathbf{W}$:

$$\begin{cases} W_1(k+1) = \dfrac{\overline{W}_1(k+1)}{\left\|\overline{W}_1(k+1)\right\|}, \\ W_p(k+1) = \dfrac{\overline{W}_p(k+1) - Q}{\left\|\overline{W}_p(k+1) - Q\right\|} \quad , \quad p = 2,\ldots,n_T \\ Q = \sum_{l=1}^{p-1}\left[W_l^H(k+1)\,\overline{W}_p(k+1)\right]W_l(k+1) \end{cases}$$

où : W_i et $\overline{W}_i$ sont la *i-ème* colonne des matrices $\mathbf{W}$ et $\overline{\mathbf{W}}$ respectivement.

5. Algorithme MIMO-CDNA

Avec la norme infinie, il est évident que, contrairement au CMA, l'algorithme CQA est sensible au résidu de porteuse puisqu'il récupère la phase. En effet, dans le transitoire, l'algorithme CQA a besoin d'un certain nombre d'itérations pour compenser le résidu, les performances en terme de rapidité de convergence sont alors moins bonnes comparées à celles du CMA. Une fois le résidu compensé, le CQA atteint un meilleur régime permanent.

Pour régler ce problème, nous proposons d'utiliser la norme définie par :

$$\|z_i\|_\lambda = \alpha\ \lambda\ \|z_i\|_\infty + (1-\lambda)|z_i|$$

pour les systèmes MIMO. L'algorithme obtenu baptisé CDNA est un cas particulier de CNA, il combine à la fois les avantages du CMA et du CQA, il utilise le CMA en phase de démarrage et ensuite le CQA pour son meilleur régime permanent. La norme dépend d'un paramètre de pondération λ ; pour $\lambda = 0$ et $\lambda = 1$ nous obtenons respectivement le module et la norme infinie. La constante $\alpha > 0$ offre un degré de liberté supplémentaire. Afin d'obtenir le meilleur de cette norme, le paramètre de pondération λ est adapté lui-même par un gradient stochastique.

La fonction de coût modifiée est définie par :

$$\min_{\mathbf{W}} J_{MIMO-CDNA}(\mathbf{W}) = \sum_{i=1}^{n_T} E\left[\|z_i\|_\lambda^2 - R(\lambda)\right]^2$$
$$\text{sous la contrainte :} \mathbf{G}^H\mathbf{G} = \mathbf{I}_{n_T}$$

où $$R(\lambda) = \dfrac{E\left[\|z\|_\lambda^4\right]}{E\left[\|z\|_\lambda^2\right]}$$

et l'algorithme du gradient stochastique est donné par :

$$\begin{cases} \overline{\mathbf{W}}(k+1) = \mathbf{W}(k) - \mu_{\mathbf{W}}\,[\Delta_1(k)\ \ldots\ \Delta_{n_T}(k)]\,\mathrm{Y}^*(k) \\ \lambda_{k+1} = \lambda_k - \mu_\lambda \sum_{i=1}^{n_T}\left[\|z_i(k)\|_{\lambda_k}^2 - R(\lambda_k)\right]\left[2\|z_i(k)\|_{\lambda_k}\right. \\ \qquad\qquad \left.\left(\alpha\|z_i(k)\|_\infty - |z_i(k)|\right) - R'(\lambda_k)\right] \end{cases}$$

où:

$$\Delta_i(k) = \left[\|z_i(k)\|_{\lambda_k}^2 - R(\lambda_k)\right]\|z_i(k)\|_{\lambda_k}\left[\alpha\lambda_k F(z_i(k)) + (1-\lambda_k)\,z_i(k)/|z_i(k)|\right]$$

et $R'(\lambda)$ est la dérivée de $R(\lambda)$ par rapport à λ.

On applique ici aussi une orthogonalisation de Gram-Schmidt à chaque itération.

6. Résultats de simulation

Les performances sont évaluées en considérant l'interférence inter-antennes de la *i*-ème source à la *k*-ème sortie, définie par :

$$IUI = \frac{\sum_{i=1}^{n_T}|g_{ki}|^2 - Max_i|g_{ki}|^2}{Max_i|g_{ki}|^2}$$

où $g_{ki} = [\mathbf{G}(k,i)] = W_k^T H_i$, W_k et H_i sont la *k*-ème et la *i*-ème colonne des matrices $\mathbf{W}$ et $\mathbf{H}$ respectivement.

Les signaux source sont i.i.d., mutuellement indépendants et issus d'une constellation 64-QAM pour 4 antennes émettrices et 4 antennes réceptrices supposées non corrélées. La matrice de canal (4×4) est choisie aléatoirement. Le bruit est complexe blanc gaussien de moyenne nulle et de variance déterminée à partir du rapport signal sur bruit défini (SNR). Pour chacune des simulations, le signal reçu est pré-blanchi avant d'appliquer les algorithmes. Les performances sont données ici pour une seule sortie, les autres sorties ayant le même comportement.

Dans un premier temps, nous comparons l'algorithme CQA avec les algorithmes CMA [2] et MUK [3]. Les pas d'adaptation ont été choisis pour obtenir sensiblement la même vitesse de convergence pour tous les algorithmes.

Nous considérons une moyenne sur 1000 exécutions indépendantes.

La figure 1 présente l'IUI d'une des sorties, en dB pour un SNR = 30 dB. On voit que l'algorithme CQA atteint un meilleur état permanent que les algorithmes CMA et MUK, avec un gain de 3 dB entre le CQA et le CMA.

En figure 2, les algorithmes CMA et CQA sont comparés en terme de rapidité de convergence, les pas d'adaptation ayant été choisis pour atteindre approximativement la même valeur d'IUI finale. On voit clairement que le CQA converge plus vite que le CMA.

Dans la figure 3, nous comparons le CDNA avec le CMA et le CQA. Le paramètre λ est initialisé à zéro (ainsi le CDNA démarre comme un CMA), le coefficient α est fixé à 1 et le pas μ_λ est fixé à 1.10^{-3}. Le pas d'adaptation pour tous les algorithmes est fixé à 8.10^{-3}. Le CDNA se conduit comme le CMA au début (exactement pour les 300 premières itérations et approximativement jusqu'à 1500 itérations). Après 2500 itérations le CQA devient meilleur que le CMA, et alors le CDNA tend vers le CQA. Ce résultat confirme que le CDNA tend à choisir le meilleur algorithme entre le CMA et le CQA.

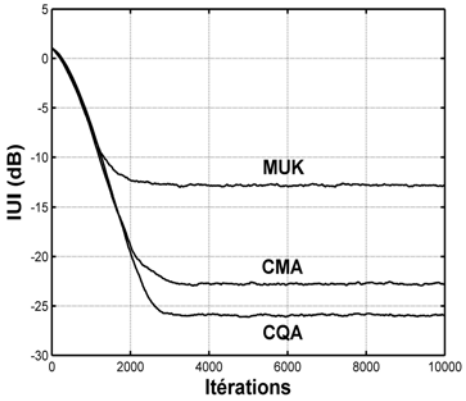


Figure 1. Performances moyennes comparées

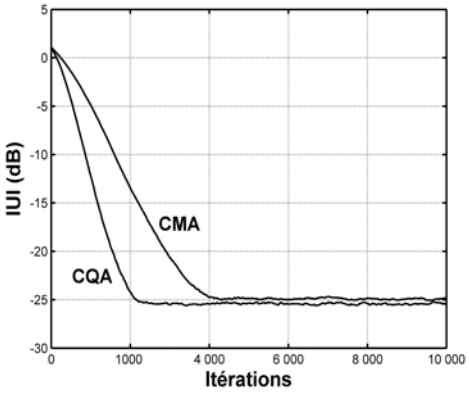


Figure 2. Comparaison du CMA et du CQA

7. Conclusion

Nous avons proposé un nouvel algorithme (CNA) pour les systèmes de communication MIMO à partir duquel nous avons déduit deux algorithmes baptisés CQA et CDNA pour la séparation aveugle de sources identiquement distribuées et mutuellement indépendantes reçues en présence d'interférences linéaires inter-antennes.

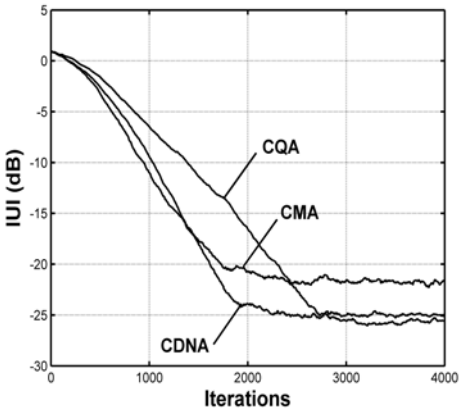


Figure 3. Comparaison CMA, CQA et CDNA

La fonction de coût est basée sur un critère à Norme Constante (CN) associé à une contrainte permettant d'éviter l'extraction du même signal simultanément sur plusieurs sorties. Les algorithmes proposés montrent de meilleures performances comparés aux algorithmes CMA et MUK avec une complexité comparable. Le CQA atteint un meilleur état permanent que le CMA (gain de 3 dB) et le CDNA tend vers le meilleur algorithme dynamiquement entre le CMA et le CQA. Ces techniques seront étendues lors de travaux futurs aux canaux avec mélanges convolutifs et à d'autres situations (CDMA, OFDM) pouvant se modéliser comme des systèmes MIMO. L'étude de la convergence sera aussi abordée.

Références

[1] A. Rouxel, D. Le Guennec et O. Macchi, *Unsupervised adaptive separation of impulse signals applied to EEG analysis*. ICASSP, Istanbul, Turkey, June 2000.

[2] L. Castedo, C. J. Escudero et A. Dapena, *A blind signal separation method for multiuser communications*, IEEE Trans. Signal Processing, vol. 45, pp. 1343-1348, May 1997.

[3] C. B. Papadias, *Globally convergent blind source separation based on a multiuser kurtosis maximization criterion*. IEEE Trans. Signal Processing, vol.48, pp. 3508-3519, Dec. 2000.

[4] O. Shalvi et E. Weinstein, *New criteria for blind deconvolution of non minimum phase systems (channels)*, IEEE Trans. Information Theory, vol. 36, pp. 312-321, Mar. 1990.

[5] P. Sansrimahachai, B. B. Ward et A. G. Constantinides, *Blind source separation for BLAST*. In *Proc. 14 th Int. Conf.* Digital Signal Processing 2002, Santarino, Greece, July 2002.

[6] A. Goupil et J. Palicot, *Constant norm algorithms class*. In Proc. EUSIPCO 2002, Toulouse, France, September 2002.

[7] J. Lundell et B. Widrow, *Applications of the constant modulus adaptive beamformer to constant and non-constant modulus signals*. In Proc. 21st Asilomar Conf. Signals, Systems, Computers, Pacific Grove, CA, USA, pp. 432-436, Nov. 1987.

Etude Statistique du seuil dans la détection d'entrelaceur

Guillaume SICOT[1], Sébastien HOUCKE[1]

[1]ENST Bretagne - Département Signal & Communication
Technopôle Brest-Iroise - CS 83818 - 29238 Brest Cedex 3
{guillaume.sicot} {sebastien.houcke} @enst-bretagne.fr

Résumé – Les entrelaceurs sont de plus en plus utilisés dans les systèmes de communications numériques utilisant des codes correcteurs d'erreurs. Dans un contexte non coopératif, tel que l'écoute passive, il se peut que le récepteur doit être capable d'estimer les différents paramètres de l'entrelaceur. Nous avons proposé un algorithme permettant d'estimer la taille de l'entrelaceur, le rendement de codage ainsi que le début de la trame entrelacée. Cet algorithme utilise un seuil de détection que nous proposions de fixer de manière heuristique. Dans cet article, nous montrons qu'il existe une valeur optimale de ce seuil et nous proposons une procédure itérative permettant de l'atteindre.

Abstract – Interleaving is a key component of many digital communication systems involving error correction schemes. In a non-cooperative context, such as passive listening, it is a challenging problem to estimate the interleaver parameters. Recently, we proposed an algorithm to estimate the size of the interleaver block, the code rate and to perform the blind frame synchronization of the interleaver blocks. This algorithm used an empirical threshold. In this paper we show that there exists an optimal threshold and we propose an iterative procedure that allows to converge to this optimal value.

1 Introduction

Les codes correcteurs d'erreurs possèdent généralement de meilleures propriétés de correction lorsque les erreurs sont équiréparties sur la trame plutôt que sous la forme de paquets d'erreurs. C'est pourquoi on entrelace les données avant émission. Ainsi, au niveau du récepteur, le désentrelacement permet de répartir uniformément les erreurs sur toute la trame.

Dans un contexte non-coopératif, les paramètres de l'entrelaceur et du schéma de codage doivent être estimés en aveugle afin de récupérer le message émis. Dans ce document nous allons nous intéresser à la détection des paramètres de l'entrelaceur. La technique proposée dans [1] résout ce problème dans le cas d'une transmission sans erreur. Afin d'étendre cette solution en présence d'erreurs de transmission une solution a été proposée dans [2]. Ces deux algorithmes permettent de retrouver des caractéristiques de l'entrelaceur telles que sa taille, le début du bloc de l'entrelaceur, le motif d'entrelacement utilisé mais aussi des informations sur le code utilisé telles que son rendement.

Dans un premier temps, nous présentons l'algorithme publié dans [2]. Ensuite une étude théorique des performances de cet algorithme est réalisée en particulier en ce qui concerne les limites de détection pouvant être attendues suivant le taux d'erreur présent dans le canal de transmission. Cette étude permet de mieux comprendre le fonctionnement de notre méthode et surtout donne des critères permettant de choisir objectivement la valeur d'un paramètre primordial de l'algorithme.

2 Détection en aveugle d'un entrelaceur

Nous nous intéressons au cas où le code correcteur utilisé est un code en bloc et l'entrelaceur est également un entrelaceur par bloc. L'opération de codage à l'aide d'un code en bloc est complètement définie par une matrice génératrice de rang plein G, qui transforme chaque bloc de k_c bits d'information en n_c bits codés (avec $k_c < n_c$).

Un entrelaceur peut être modélisé par une matrice de permutation P de taille $S \times S$ où S représente la taille de l'entrelaceur. En général, la taille de l'entrelaceur est un multiple de taille d'un mot de code, hypothèse que nous ferons ici. Ainsi, nous avons la relation : $S = N.n_c$, où N représente le nombre de mots de code dans le bloc de l'entrelaceur.

Par la suite, $\mathbf{X}$ désigne la séquence émise composée de M blocs entrelacés, et $\mathbf{Z}$ la séquence interceptée. $\mathbf{Z}$ est une version de $\mathbf{X}$, décalée de t_0 bits (retard de transmission) et entachée d'erreurs dues au canal de propagation. Le canal est modélisé par un canal binaire symétrique de probabilité d'erreur P_e.

2.1 Sans erreur de transmission

Dans cette partie, nous supposons que le canal de transmission n'introduit aucune erreur de transmission (*i. e.* $P_e = 0$). Burel *et al.* proposent [1] de construire une matrice $H(n_a, d)$ à partir de la séquence $\mathbf{Z}$. Pour ce faire, on découpe la séquence $\mathbf{Z}$ à laquelle on a préalablement enlevé les d premiers bits, en blocs K contenant n_a bits. Ensuite ces K blocs de taille n_a forment les lignes de la matrice $H(n_a, d)$. Ainsi la matrice $H(n_a, d)$ est une matrice de dimension $K \times n_a$. Burel *et al.* proposent d'étudier le coefficient $\rho(n_a, d)$ défini par (1) pour différentes valeurs de n_a et de d.

$$\rho(n_a, d) = \frac{\text{rank}(H(n_a, d))}{n_a} \tag{1}$$

Il apparaît que $\forall\, d$, $\rho(n_a, d)$ est égal à 1, sauf lorsque n_a est un multiple de la taille de l'entrelaceur, c'est-à-dire lorsque $n_a = kS$, avec k un entier non nul. En effet, dans ce cas particulier, des corrélations entre colonnes apparaissent dues à la redondance introduite par le code correcteur d'erreur.

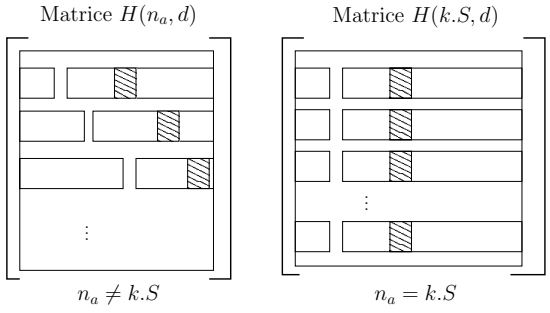


FIG. 1 – La matrice $H(n_a, d)$

La figure 1 présente les deux cas de figure qui peuvent se présenter au niveau de la matrice $H(n_a, d)$. L'élément grisé dans la figure 1 représente un bit de redondance : ce bit est donc une combinaison linaire d'autres bits appartenant au même bloc. Lorsque $n_a = kS$, cette même relation linéaire est aussi vérifiée par la ligne suivante de la matrice $H(kS, d)$ et donc par toute la colonne. La colonne hachurée est une combinaison linéaire d'autres colonnes et la matrice $H(kS, d)$ n'est pas de rang plein. Dès que $n_a \neq kS$ cette propriété n'est plus vérifiée. La déficience du rang de $H(n_a, d)$ permet ainsi d'estimer la taille de l'entrelaceur. Une fois S estimée, l'argument minimum de $\rho(kS, d)$ en fonction de d permet de trouver le début de la trame de l'entrelaceur et de déduire le rendement du code utilisé.

2.2 Avec erreurs de transmission

Dans le contexte d'une écoute passive, la séquence $\mathbf{Z}$ est vraisemblablement entachée d'erreurs, ainsi la matrice $H(n_a, d)$ devient généralement de rang plein même pour $n_a = kS$. Par conséquent la méthode précédemment citée devient peu performante. Pour pallier à ce problème, nous avons proposé dans [2] de construire la matrice $H(n_a, d)$ de la même manière que dans la méthode précédente. La matrice $H(n_a, d)$ est ensuite trianguralisée à l'aide de l'algorithme du pivot de Gauss modifié pour fonctionner dans l'espace binaire $\mathbb{F}_2$. Il est possible de définir l'opération réalisée par cet algorithme par la relation linéaire suivante :

$$A_1 H(n_a, d) A_2 = L(n_a, d) \tag{2}$$

où A_1 représente les permutations sur les lignes de $H(n_a, d)$, et A_2 représentant les opérations sur les colonnes. Enfin $L(n_a, d)$ représente la matrice triangulaire inférieure résultant de cet algorithme.Soit B_i la variable qui représente le nombre de 1 dans la partie inférieure de la colonne i de la matrice $L(n_a, d)$ (cf. figure 2).

Définissons par $\phi_{n_a}(i)$ la variable aléatoire suivante : $\phi_{n_a}(i) = \frac{B_i}{m_B}$ où m_B est un facteur de normalisation égal à

$$m_B = \frac{1}{2}\left(\lfloor \frac{MS}{n_a} \rfloor - n_a\right).$$

Il correspond à la moyenne de B_i si les bits de la colonne i sont indépendants et équiprobables (ce qui est le cas si $n_a \neq kS$).

Si $n_a \neq kS$, $\forall\, i \in \{1, .., n_a\}$, B_i suit une loi binomiale de paramètres $(K - n_a, \frac{1}{2})$ et il est facile de montrer que :

$$\forall i \in \{1, .., n_a\},\ \lim_{M \to \infty} \phi_{n_a}(i) \xrightarrow{\mathcal{P}} 1 \tag{3}$$

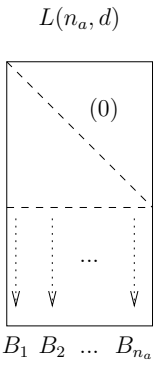


FIG. 2 – La matrice $L(n_a, d)$

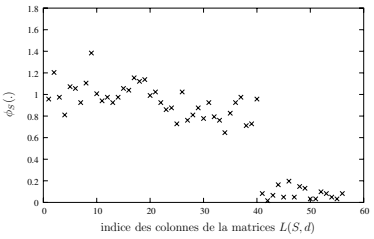


FIG. 3 – Exemples des valeurs prises par ϕ pour $n_a = kS$

où $\xrightarrow{\mathcal{P}}$ signifie la convergence en probabilité.

Si $n_a = kS$, alors nous devrions trouver des colonnes dépendantes dans $H(n_a, d)$, c'est-à-dire nous devrions avoir des colonnes de zéros dans la matrice $L(n_a, d)$. Soit $\mathcal{I}$ l'ensemble de ces colonnes. En supposant que la partie triangulaire supérieure de $H(n_a, d)$ ne contienne aucune erreur alors pour $i \in \mathcal{I}$, B_i suit une loi Binomiale de paramètres $(K - n_a, P)$ et nous avons :

$$\forall i \in \mathcal{I} \lim_{M \to \infty} \phi_{kS}(i) \xrightarrow{\mathcal{P}} = 2P$$

avec :

$$P = 1 - \sum_{l=0}^{\lfloor \frac{p_i}{2} \rfloor} \binom{p_i}{2l} P_e^{2l} (1 - P_e)^{p_i - 2l} \tag{4}$$

où p_i représente le nombre minimal de combinaisons linéaires nécessaires pour obtenir la colonne i de $L(kS, d)$ et P est la probabilité qu'un élément de la colonne i est égale à 1. De plus, si $n_a \neq kS$, $\phi_{n_a}(i)$ possède le même comportement que $\phi_{kS}(i)$ pour $i \notin \mathcal{I}$.

La figure 3 nous montre que l'écart entre les deux comportements de $\phi_S(.)$ est significatif même pour des valeurs de M finies. Cette figure a été réalisée avec une code de Hamming $(7, 4)$, un entrelaceur de taille 56, un taux d'erreur binaire du canal de propagation $p_e = 0.08$ et une durée d'observation de 10000 bits. Pour ce code particulier, p_i est égal à $4\ \forall i$.

Précisons que les erreurs se produisant dans la partie triangulaire supérieure de $H(kS, d)$ ont l'effet suivant : nous risquons d'ajouter une colonne que nous ne devrions pas ajouter et ainsi nous risquons de "perdre" une corrélation. Dans [2], nous avons proposé une procédure itérative permettant de contrer ce phénomène. En effet le Pivot de Gauss utilise exclusivement la partie triangulaire supérieure de $H(n_a, d)$ pour obtenir $L(n_a, d)$,donc si nous relançons l'algorithme sur une autre matrice $H(n_a, d)$ obtenue par une permutation de ses lignes, nous

détecterons éventuellement d'autres corrélations et les performances de l'algorithme sont ainsi améliorées.

Pour estimer la taille de l'entrelaceur, nous procédons par une recherche exhaustive sur le paramètre n_a. Pour chaque valeur de n_a, nous calculons $\phi_{n_a}(i)$. L'estimateur $\hat{S}$ de la taille de l'entrelaceur est alors :

$$\hat{S} = \text{Argmax}_{n_a} \left(\text{Card}(\{i = 1, \ldots, n_a | \phi_{n_a}(i) < \beta\})\right)$$

où β est un seuil fixé. Le choix de ce seuil conditionne les performances de notre algorithme. Dans cette article nous proposons de trouver la valeur optimale de ce seuil.

3 Détermination du seuil optimal

Le seuil est considéré optimal dans le sens où il minimise la probabilité de mauvaise détection P_{md} d'une colonne théoriquement dépendante (une colonne appartenant à l'ensemble $\mathcal{I}$). Connaissant la loi de probabilité de ϕ, nous sommes en mesure de déterminer le seuil optimal.

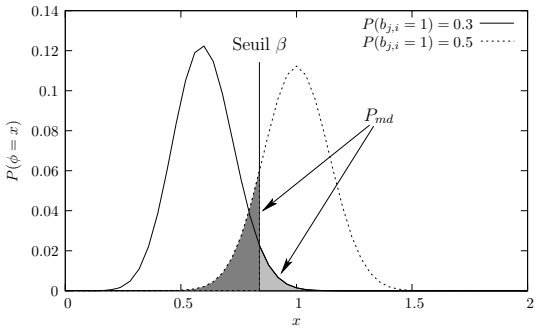


FIG. 4 – Définition de P_{md}

La figure 4 nous présente un exemple de densités de probabilités[1] de $\phi_S(i)$, celle de gauche correspond à une colonne appartenant à l'ensemble $\mathcal{I}$ et celle de droite correspond à une colonne indépendante. De plus nous avons représenté un seuil possible ainsi que la probabilité de mauvaise décision, P_{md}, associée. P_{md} correspond à la somme de 2 probabilités : la probabilité d'avoir $\phi_S(i) > \beta$ alors que la colonne i est une colonne dépendante, et la probabilité d'avoir $\phi_S(i) < \beta$ alors que la colonne i est une colonne dépendante.

Par conséquent le problème est de déterminer le seuil qui minimise P_{md}, ce seuil dépendant *a priori* de m_B, P_e et de p_j. Ainsi il est possible de définir le seuil β^* comme ci-dessous :

$$\beta^* = \arg\min_{\beta} P_{md}(m_B, P, \beta) \tag{5}$$

En utilisant les équations (3) et (4), on obtient l'expression suivante :

$$P_{md}(m_B, P, \beta) = \sum_{i=0}^{\lfloor m_B.\beta \rfloor} \binom{i}{2m_B} (0.5)^{2m_B} + \sum_{i=\lfloor m_B.\beta \rfloor + 1}^{2m_B} \binom{i}{2m_B} P^i (1-P)^{2m_B - i} \tag{6}$$

La résolution de ce problème d'optimisation permet d'obtenir la valeur théorique du seuil optimal β^*.

[1]Ces densités de probabilités sont normalement discrètes, néanmoins par un souci de clarté nous avons préféré les représenter sous forme continue

3.1 Propriétés du seuil optimal β^*

Cette section illustre l'influence des paramètres m_B et P sur β et en particulier sur le seuil optimal β^*.

Dans un premier temps, nous étudions l'influence de β sur P_{md} et cela pour différentes valeurs de P_e. La figure 5 présente P_{md} en fonction de β où l'on a fixé $m_B = 50$ et P est calculé pour $p_i = 6$ et pour différents P_e (cf. eq.(4)).

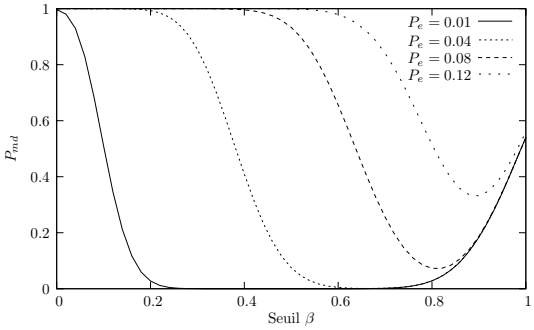


FIG. 5 – $P_{md}(m_B, P, \beta)$ en fonction de β avec $m_B = 50$, et P calculé avec $p_i = 6$

Comme le montre la figure 5, lorsque le canal est peu sévère (P_e faible), la valeur du seuil a peu d'influence sur les performances de bonne détection : en effet pour $P_e = 0.01$ si le seuil $\beta \in [0.2, 0.8]$, la valeur de P_{md} est proche de zéro. Si l'on connaît P_e et p_i, il est tout à fait possible de déterminer le seuil optimal à partir de ces courbes.

Maintenant, nous illustrons figure 6 la dépendance du seuil optimal β^* par rapport à la variable p_i. On remarque que plus p_i est grand plus le seuil optimal β^* est grand. En d'autres termes plus p_i est grand et plus difficile sera la détection.

Enfin, nous illustrons l'influence de m_B sur β^*. m_B n'a que peu d'influence sur β^* et cela peut facilement être expliqué. En effet si m_B augmente, alors la variance de la densité de probabilité des B_i diminue. Cela n'influence donc pas la position du seuil. m_B possède donc un effet important sur la valeur de P_{md} comme le montre la figure 7. Sur cette courbe sont représentées les valeurs de P_{md} lorsque que l'on choisit le seuil optimal β^*. Cette figure montre en particulier que P_{md} baisse significativement avec m_B, par exemple pour $P_e = 0.1$, $P_{md} = 0.35$ pour $m_B = 25$, et on trouve que $P_{md} = 0.06$ lorsque $m_B = 100$. Nous proposons maintenant une procédure permettant d'estimer le taux d'erreur binaire du canal (*i.e.* P_e) et ainsi d'obtenir une estimation du seuil optimal β^*.

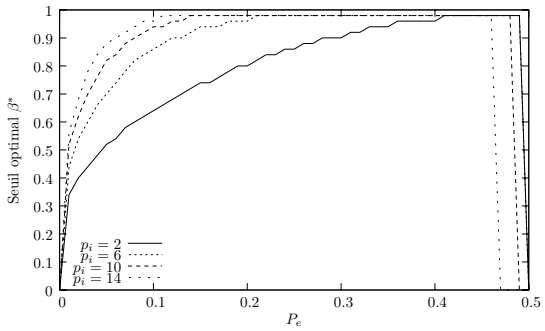


FIG. 6 – Le seuil optimal β^* en fonction de p_i

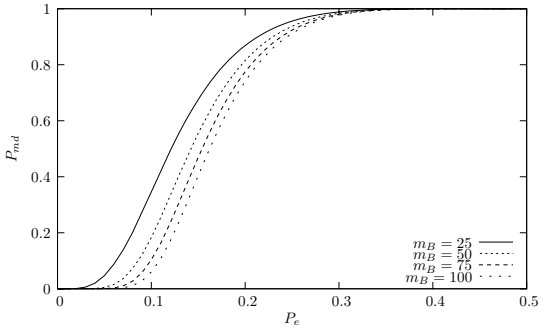


FIG. 7 – P_{md} en fonction de P_e lorsque $\beta = \beta^*$

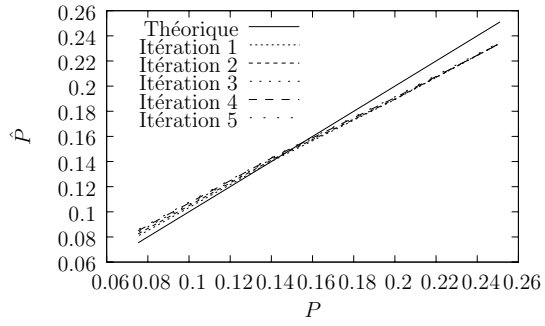


FIG. 8 – Estimation of P

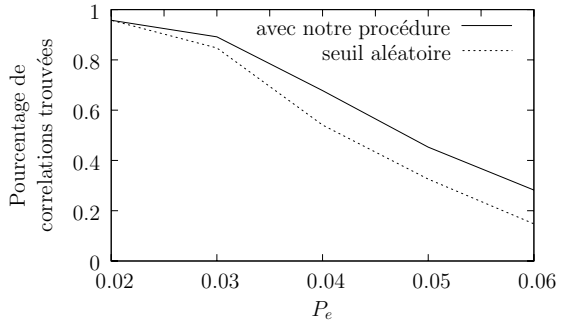


FIG. 9 – Gain de performance avec notre procédure

3.2 Choix du paramètre β

Nous avons vu que le paramètre dépend de p_i, P_e et m_B. m_B est connu du récepteur, cependant p_i, P_e sont inconnus. Donc nous sommes *a priori* incapables de fixer le seuil au début de l'algorithme à sa valeur optimale. Nous proposons cependant une procédure itérative compatible avec l'algorithme itératif initial permettant d'adapter la valeur du seuil.

Après une itération de l'algorithme initial, nous sommes éventuellement en mesure de trouver une valeur de n_a pour laquelle on détecte des colonnes dépendantes. Cette valeur correspond à la taille de l'entrelaceur. Soit $\mathcal{K}$ l'ensemble de ces colonnes détectées. Grâce à la matrice A_2, on obtient les p_i pour ces colonnes (p_i représente le nombre de colonnes impliquées pour obtenir la colonne i de $L(S, d)$). Ensuite en utilisant la valeur des $\phi_S(i)$ pour $i \in \mathcal{K}$, il est possible d'estimer P :

$$\hat{P} = \frac{1}{\text{Card}(\mathcal{K})} \sum_{i \in \mathcal{K}} \frac{\phi_{\hat{S}}(i)}{2}.$$

L'équation (4) permet alors d'estimer P_e. Enfin la figure 6 nous permet d'estimer la valeur du seuil optimal.

3.3 Validation du modèle et performances de notre nouvel algorithme.

Tout d'abord, nous validons expérimentalement la loi de probabilité suivie par $\phi(.)$ (*i.e.* équation (4)). Nous avons utilisé un entrelaceur pseudo aléatoire de taille 56 et un code correcteur de Hamming (7,4) (*i.e.* $p_i = 4$). Le nombre de bits interceptés est fixé à 10000 bits ($m_B = 122$ pour $n_a = 56$). 3000 tirages de Monte-Carlo ont été réalisés où pour chaque essai la séquence émise, les erreurs et l'entrelaceur ont été tirés aléatoirement. La figure 8 représente la moyenne de $\hat{P}$ ainsi que sa valeur théorique asymptotique. Nous constatons que les valeurs obtenues par simulation sont proches des valeurs théoriques.

Enfin, nous illustrons l'amélioration apportée par notre procédure d'adaptation du seuil par rapport à notre méthode initiale. Le contexte de simulation est le même que précédemment. Le seuil initial est tiré aléatoirement entre 0.3 et 0.7. Afin de quantifier le gain apporté par notre procédure, on suppose que la taille de l'entrelaceur a été correctement estimée (*i.e.* au moins une corrélation a été détectée). La figure 9 présente le pourcentage de corrélations trouvées après 20 itérations de notre algorithme de détection. Le nombre de corrélations trouvées avec notre procédure de seuil adaptatif est plus grand : en effet pour un TEB de 4% et après 20 itérations, nous trouvons environ 14% de corrélations de plus que notre algorithme initial. Nous obtenons les mêmes performances que notre algorithme initial mais seulement après 13 itérations soit un gain de 7 itérations. Avec notre nouvel algorithme, nous serons donc en mesure de mieux reconstruire le motif de l'entrelaceur (*i.e.* plus de correlations disponibles), de mieux trouver le début de la trame entrelacée.

4 Conclusion

Les performances de notre premier estimateur aveugle des paramètres d'un entrelaceur dépendent de notre capacité à choisir un seuil de manière appropriée. L'étude théorique menée dans cet article nous permet de définir un seuil optimal qui minimise la probabilité de mauvaise détection d'une corrélation due au code correcteur d'erreurs. Nous avons montré qu'il existe un seuil optimal mais que malheureusement ce seuil dépend de paramètres inconnus du récepteur (par exemple le TEB). Cependant, nous proposons une procédure permettant d'estimer ces paramètres et ainsi d'adapter le seuil pour qu'il converge vers sa valeur optimale. Ainsi nous pouvons espérer atteindre les meilleures performances attendues avec cet algorithme. De plus cette procédure nous permet d'estimer le taux d'erreur du canal de propagation.

Références

[1] G. Burel and R. Gautier, "Blind estimation of encoder and interleaver characteristics in a non cooperative context," *IASTED - CIIT*, November 2003.

[2] G. Sicot and S. Houcke, "Blind detection of interleaver parameters," in *Proc. of ICASSP*, Philadelphia (USA), 2005.

Analyse du Facteur de Crête dans le contexte Radio Logicielle.

Sidkiéta ZABRÉ, Jacques PALICOT

IETR/Supelec, Equipe SCEE
Avenue de la Boulaie BP 81127 35511 Cesson-Sévigné Cedex, France
sidkieta.zabre@supelec.fr,jacques.palicot@supelec.fr

Résumé – Dans le nouveau contexte de la Radio Logicielle, les signaux traités corespondent à une somme de porteuses modulées dans une large bande. Par conséquent, ils auront de grandes variations de puissance. C'est précisément ce phénomène que nous étudions dans cet article. Dans un premier temps, nous proposons une borne supérieure du facteur de crête. Ensuite, grâce à une analogie forte entre les signaux OFDM et Radio Logicielle, nous proposons une vision porteuse par porteuse de ce facteur de crête. Cette nouvelle vision, permettra d'aborder avec une nouvelle approche les techniques de réduction du facteur de crête.

Abstract – In this paper, in the new context of Software Radio, we propose an upper bound of the creast factor. Furthermore, thanks to an analogy with OFDM signals we propose a carrier by carrier interpretation of the creast factor. This new vision opens a new way of creast factor reduction methods.

1 Introduction

Nous avons assisté au cours des deux dernières décennies, à une prolifération de standards de radio communications aux spécifications différentes. De fait, la plupart des systèmes sont actuellement incompatibles entre eux. Les systèmes de télécommunications devront donc faire l'objet à l'avenir de développements importants pour aboutir à des terminaux universels capables de communiquer dans tout environnement. Ceci implique les notions d'interopérabilité et de re-configurabilité. La Radio Logicielle (RL) qui inclut ces notions pourrait donc être une solution [1].

Il est clair dans ce contexte multi-standards, que les signaux résultants seront très large bande et présenteront une grande variation d'amplitude. Or cette caractéristique les rend très sensibles aux non-linéarités des composants analogiques, en particulier celles de l'amplificateur de puissance.

Le problème d'amplification des signaux à enveloppes non constantes à forte dynamique est un problème récurent. Il s'est posé dans les années 80 avec le satellite et dans les années 90 avec l'OFDM, et se posera à nouveau pour les futures modulations BFDM, IOTA, MC-CDMA. Cependant il sera encore plus complexe dans le contexte Radio Logicielle car les signaux à amplifier seront à très large bande et à très grande dynamique. Bien entendu, trouver des techniques efficaces et bien adaptées au problème posé requiert à la fois une compréhension très fine des phénomènes mis en jeu et des outils pour les caractériser. C'est précisément l'objet de cet article.

Nous allons dans un premier temps définir et modéliser les différentes sortes de signaux susceptibles d'être rencontrés en Radio Logicielle. Nous caractériserons la notion de dynamique dans le nouveau contexte Radio Logicielle, en partant des résultats connus pour les modulations OFDM. Dans une étude précédente [6] nous avions constaté que la définition du facteur de crête est sujette à de multiples interprétations. En effet, nous trouvons, par exemple, les appellations PMEPR (Peak to Mean Envelop Power Ratio), PAPR (Peak to Average Power Ratio), CF (Crest Factor), PF (Peak Factor)[3], [5], [4]. Dans cette étude [6] nous avons proposé une appellation qui permet d'homogénéiser toutes ces différentes notations. C'est cette appellation que nous utiliserons tout au long de cet article: il s'agit du Power Ratio (PR). Dans la section 3, nous proposons une borne supérieure du PR. De cette borne, nous retrouvons les bornes déjà connues comme dans le cas de l'OFDM. En rompant avec la démarche jusque là adoptée pour l'analyse du Power Ratio de l'OFDM, nous proposons une vision porteuse par porteuse du calcul de ce paramètre, dans la section 4. Cette nouvelle vision nous permet d'envisager de nouvelles techniques du traitement des non-linéarités.

2 Modélisation des signaux "Radio Logicielle "

De par le concept de la RL, il est à prévoir différentes sortes de signaux : signaux mono-canal, signaux mono-standard, signaux multi-standards etc. Il est donc primordial d'être très clair sur ces différents types de signaux, car l'objectif étant de trouver des techniques pour réduire le PR, on comprend que si certaines techniques s'avéreraient très efficaces pour certains signaux, il n'en sera plus de même pour d'autres. Pour ce faire, nous proposons des définitions avec une description analytique de ces différents signaux.

- Multiplex de porteuses modulées: Cette expression sera utilisée pour un signal contenant un certain nombre N de porteuses modulées. L'expression analytique d'un tel signal s'écrit :

$$x(t) = \sum_{k=1}^{N} (fem_k(t) * m_k(c(t)))e^{2i\pi f_k t} \qquad (1)$$

où $fem_k(t)$ est le filtre de mise en forme relative à la fréquence porteuse f_k et $m_k(t)$ la modulation correspon-

dante. Notons que le signal transmis n'est autre que la partie réelle du signal $x(t)$.

- Signal Multiporteuses : C'est un Multiplex de porteuses modulées dont l'écart entre les différentes porteuses est constant.

$$x(t) = \sum_{i=k}^{N} (fem_k(t) * m_k(c(t)))e^{2i\pi(f_o+(k-1)\delta)t} \quad (2)$$

- Signal Composite ou Signal Multi-standards : C'est un signal (très) large bande comprenant plus d'un standard de télécommunication. Bien entendu, un signal composite est un multiplex de porteuses modulées, mais n'est généralement pas un signal multiporteuses.

$$x(t) = \sum_{k=1}^{S} S_k(t) \quad (3)$$

avec S le nombre de standards et $S_k(t)$ est l'équation du standards k donné par l'équation (1).

On a donc:

$$x(t) = \sum_{k=1}^{S} \sum_{p=1}^{P_k} (fem_{k,p}(t) * m_{k,p}(c(t)))e^{2i\pi f_{k,p}t} \quad (4)$$

où P_k est le nombre de porteuses dans le standard k.

- Signal Monostandard : Il s'agit d'un cas particulier d'un signal composite.

- Signal OFDM: C'est un cas particulier d'un signal multiporteuses.

3 Calcul d'une borne supérieure du Power Ratio d'un signal multi-standards

Comme déjà mentionné, le signal multi-standard est un signal très large bande contenant au moins un standard. Il peut contenir un ou plusieurs canaux de un ou plusieurs standards. C'est donc un signal très complexe. Il est donné par la relation (4), que nous rappelons ci-dessous:

$$x(t) = \sum_{i=1}^{S} \sum_{p=1}^{P_i} r_{i,p}(t)e^{2j\pi f_{i,p}t} \quad (5)$$

La puissance instantanée du signal $P(t) = |x(t)|^2$ s'écrit alors:

$$\begin{aligned} |x(t)|^2 &= (\sum_{i=1}^{S} \sum_{p=1}^{P_i} r_{i,p}(t)e^{2i\pi f_{i,p}t})(\sum_{i=1}^{S} \sum_{p=1}^{P_i} r^*_{i,p}(t)e^{-2i\pi f_{i,p}t}) \\ &= \sum_{i=1}^{S} \sum_{p=1}^{P_i} |r_{i,p}(t)|^2 + \\ &\quad \sum_{i=1}^{S} \sum_{p=1}^{P_i} \sum_{p' \neq p} r_{i,p}(t)r^*_{i,p}(t)e^{2i\pi(f_{i,p}-f_{i,p'})t} + \\ &\quad \sum_{i=1}^{s} \sum_{i' \neq i} \sum_{p=1}^{P_i} \sum_{p' \neq p} r_{i,p}(t)r^*_{i',p'}(t)e^{2i\pi(f_{i,p}-f_{i',p'})t} \end{aligned} \quad (6)$$

Ainsi, on a:

$$\begin{aligned} max_t|x(t)|^2 \leq\ & max_t(\sum_{i=1}^{S} \sum_{p=1}^{P_i} |r_{i,p}(t)|^2) + \\ & + max_t(\sum_{i=1}^{S} \sum_{p=1}^{P_i} \sum_{p' \neq p} r_{i,p}(t)r^*_{i,p}(t)e^{2i\pi(f_{i,p}-f_{i,p'})t}) \\ & + max_t(\sum_{i=1}^{S} \sum_{i' \neq i} \sum_{p=1}^{P_i} \sum_{p' \neq p} r_{i,p}(t)r^*_{i',p'}(t)e^{2i\pi(f_{i,p}-f_{i',p'})t}) \end{aligned}$$

En majorant les différents termes, on montre que:

$$max_t(|x(t)|^2) \leq (\sum_{i=1}^{S} P_i^2 + \sum_{i=1}^{s} \sum_{i'=1} P_iP_{i'})max_t(|r_{i,p}(t)|^2) \quad (7)$$

La puissance moyenne de l'enveloppe du signal multi-standard s'écrit alors:

$$\begin{aligned} E(|x(t)|^2) &\geq E(\sum_{i=1}^{S} \sum_{p=1}^{P_i} |r_{i,p}(t)|^2) \\ &\geq \sum_{i=1}^{S} \sum_{p=1}^{P_i} E(|r_{i,p}(t)|^2) \end{aligned} \quad (8)$$

On déduit ainsi des relations précédentes une limite supérieure du PR d'un signal multi-standard.

$$PR \leq (\sum_{i=1}^{S} P_i^2 + \sum_{i=1}^{s} \sum_{i' \neq i} P_iP_{i'}) \frac{max_t(|r_{i,p}(t)|^2)}{\sum_{i=1}^{S} \sum_{p=1}^{P_i} E(|r_{i,p}(t)|^2)} \quad (9)$$

3.1 Cas particulier du DVB-T

Nous dérivons, dans ce qui suit, cette borne dans le cas particulier du signal mono-standard DVB-T.

L'expression du signal analytique est donnée par la relation (pour un symbole OFDM par canal):

$$S(t) = \sum_{p=1}^{P} \sum_{k=0}^{N-1} C_{k,p}e^{2i\pi f_{k,p}t} \quad (10)$$

où P désigne le nombre de canaux, N le nombre de sous-porteuses par canal et $C_{k,p}$ les symboles numériques issus d'une modulation numérique.

Par identification avec la relation donnée par (5)et pour $S = 1$, la relation (10) s'écrit:

$$S(t) = \sum_{p=1}^{P} r_p(t)e^{2i\pi f_p t} \quad (11)$$

avec $r_p(t) = \sum_{k=0}^{N-1} C_{k,p}e^{2i\pi(f_{k,p}-f_p)t}$

On déduit de la relation (9), la borne du PR d'un signal mono-standard DVB-T:

$$PR \leq P^2 \frac{max_t(|r_p(t)|^2)}{\sum_{p=1}^{P} E(|r_p(t)|^2)} \quad (12)$$

En majorant le numérateur de la borne ci-dessus, on trouve l'inégalité suivante:

$$PR \leq NP \frac{max_{k,p}|C_{k,p}|^2}{E_{k,p}(|C_{k,p}|^2}) \quad (13)$$

Lorsque P=1, on retrouve la borne déjà connue dans le cas particulier du signal OFDM pour un symbole:

$$PR \leq N \frac{max_k |C_k|^2}{E(|C_k|^2)} \quad (14)$$

Autrement dit $(9) \leq (14)$ donc la borne trouvée par (9) sera plus fine que celle donnée classiquement par (13).

Nous illustrons les bornes obtenues dans les relations (12) et (13) à travers la figure (1). Le PR du signal est donné en trait continu, la borne donnée par (12) est en $(*)$ et celle donnée par (13) en o. Dans cette simulation, chaque canal DVB comprend 2048 sous-porteuses, chacune étant modulée par une MAQ16.

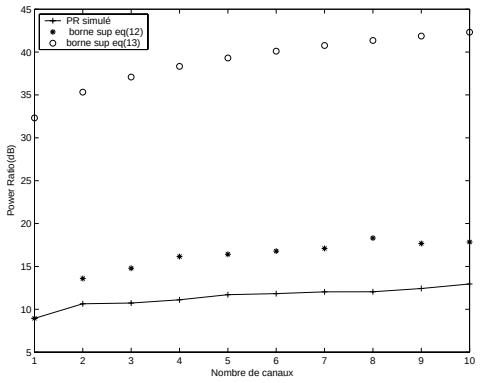


Figure 1: Borne supérieure du PR:cas DVB, pour un symbole par canal.

4 Analyse du PR

La différence entre le PR théorique donné par la borne classique pour un symbole et le PR simulé est très grande. Pour diminuer cette différence nous proposons une nouvelle approche, à savoir l'analyse du PR sur un grand nombre de symboles.

4.1 Nouvelle approche du PR de l'OFDM: Calcul sur plusieurs symboles

Soit le signal complexe bande de base OFDM :

$$S(t) = \sum_{j=1}^{N_s} \sum_{k=0}^{N-1} C_{j,k} \Psi_{j,k}(t) \quad (15)$$

ou encore

$$S(t) = \sum_{j=1}^{N_s} \sum_{k=0}^{N-1} C_{j,k} e^{2i\pi f_k (t - jT_s)} g(t - jT_s) \quad (16)$$

On peut écrire le signal $S(t)$ comme somme de signaux sur le nombre de symboles considérés, comme ceci:

$$S(t) = \sum_{j=1}^{N_s} S_j(t) \quad (17)$$

avec

$$S_j(t) = \sum_{k=0}^{N-1} C_{j,k} e^{2i\pi f_k (t - jT_s)} g(t - jT_s) \quad (18)$$

Les signaux $S_j(t)$ étant à support disjoint, on montre aisément que:

$$|S(t)|^2 = \sum_{j=1}^{N_s} |S_j(t)|^2 \quad (19)$$

En considérant un grand nombre de symboles et de par la définition du PR [6], nous pouvons écrire que le PR du signal OFDM est:

$$PR_{N_s} = \frac{\max_{t \in [0, N_s T_s]} |S(t)|^2}{\frac{1}{N_s T_s} \int_0^{N_s T_s} |S(t)|^2 dt} \quad (20)$$

Nous pouvons écrire que:

$$\int_0^{N_s T_s} |S(t)|^2 dt = \sum_{j=1}^{N_s} \int_{(j-1)T_s}^{jT_s} |S_j(t)|^2 dt \quad (21)$$

Comme,

$$Max_t |S(t)|^2 = Max_t (\sum_{j=1}^{N_s} |S_j(t)|^2) = Max_j (Max_t (|S_j(t)|^2)$$

et

$$\frac{1}{T_s} \int_{(j-1)T_s}^{jT_s} |S_j(t)|^2 dt = E(|S_j(t)|^2) \quad (22)$$

Nous pouvons donc établir:

$$PR_{N_s} = \frac{Max_j (\max_t |S_j(t)|^2)}{\frac{1}{N_s} \sum_{j=1}^{N_s} E(|S_j(t)|^2)} \quad (23)$$

D'après [3], $\max_t |S_j(t)|^2 = Max_n |c_{j,n}|^2$ où $c_{j,n} = \sum_{k=0}^{N-1} C_{j,k} e^{2ik\pi \frac{n}{N}}$. On en déduit:

$$PR_{N_s} = \frac{Max_j (\max_k |c_{j,k}|^2)}{\frac{1}{N_s} \sum_{j=1}^{N_s} E(|c_{j,k}|^2)} \quad (24)$$

Alors, avec les notations de [6] la borne du PR échantillonné à T_s:

$$PR_{N_s} \leq N \frac{Max_j (\max_k |C_{j,k}|^2)}{\frac{1}{N_s} \sum_{j=1}^{N_s} E(|C_{j,k}|^2)} \quad (25)$$

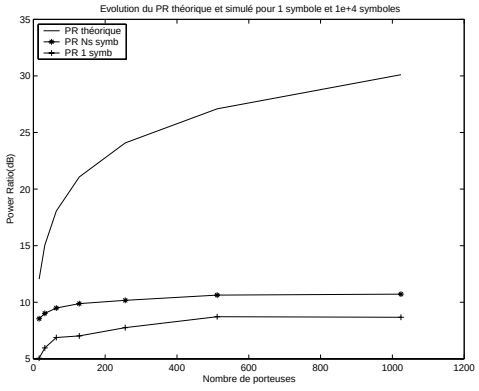


Figure 2: Évolution du PR simulé pour 1 et 10000 symboles.

La figure 2 montre bien l'intérêt de calculer ce PR sur plusieurs symboles. En effet, sur celle-ci, le PR pour un symbole OFDM de 128 porteuses croît de environ 7 dB à environ 10 dB, quand le nombre de symboles passe de 1 à 10000, même si l'on reste encore loin de la borne qui, rappelons le, est une borne très grossière.

4.2 Nouvelle approche du PR de l'OFDM: Vision porteuse par porteuse

Nous restons dans le contexte où le PR est calculé sur un grand nombre de symboles.

Cette vision du signal OFDM sur un grand nombre de symboles nous a amenée à proposer la Fig3, sur laquelle il est facile de constater que chaque sous porteuse transporte N_s symboles numériques étalés sur N_s temps symboles OFDM: $N_s T_s$.

Le signal sur la porteuse f_k peut donc s'écrire comme:

$$P_k(t) = \sum_{j=1}^{N_s} C_{j,k} e^{2i\pi f_k (t-jT_s)} g(t-jT_s) \tag{26}$$

De même, le j^{ime} symbole OFDM peut s'écrire comme:

$$S_j(t) = \sum_{k=0}^{N-1} C_{j,k} e^{2i\pi f_k (t-jT_s)} g(t-jT_s) \tag{27}$$

En faisant la somme des N signaux vus porteuse par porteuse on obtient le signal:

$$P(t) = \sum_{k=0}^{N-1} \sum_{j=1}^{N_s} C_{j,k} e^{2i\pi f_k (t-jT_s)} g(t-jT_s) \tag{28}$$

Ce signal n'est rien d'autre que le signal obtenu en sommant les N_s symboles $S_j(t)$ et qui est donné par la relation:

$$S(t) = \sum_{j=1}^{N_s} \sum_{k=0}^{N-1} C_{j,k} e^{2i\pi f_k (t-jT_s)} g(t-jT_s) \tag{29}$$

Par conséquent, le PR obtenu par "vision" porteuse par porteuse est exactement le même que le PR obtenu symbole par symbole(donné par (25)). Il va s'en dire qu'on a aussi la même borne du PR échantillonné à T_s:

$$PR_{N_s} \leq N \frac{Max_k(\max_j |C_{j,k}|^2)}{\frac{1}{N_s}\sum_{j=1}^{N_s} E(|C_{j,k}|^2)} \tag{30}$$

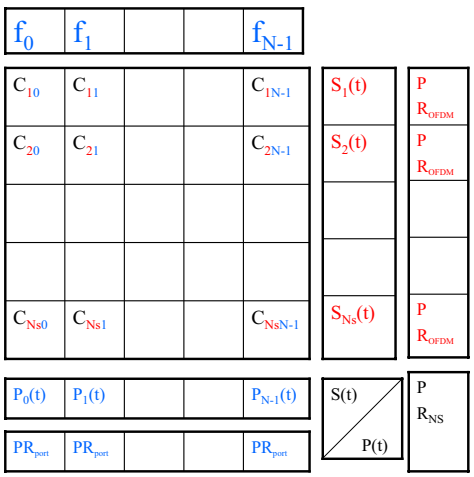


Figure 3: Représentation symbole-porteuse d'un signal OFDM.

4.3 Vision porteuse par porteuse du PR en RL

Par analogie avec ce qui vient d'être vu pour l'OFDM et sous les hypothèses d'espacement régulier entre les porteuses et d'orthogonalité temporelle (à TS) et fréquentielle (à 1/TS) alors nous pouvons mettre le signal monostandard GSM sous la forme de la figure 3 vérifiant les équations 28 et 29. Les visions classique et porteuse du PR seront donc identiques. C'est exactement ce que prouve le résultat de la figure 4. Ceci corrobore les travaux de [2] montrant que l'on peut créer et démoduler les signaux GSM dans le domaine fréquentiel.

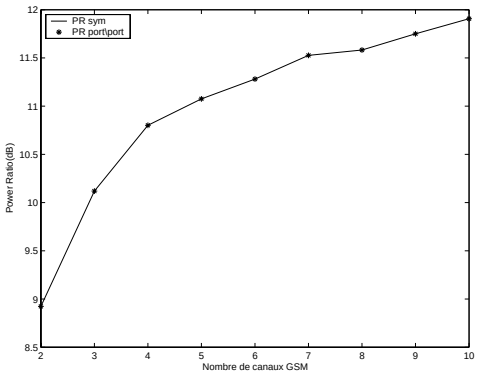


Figure 4: Calcul du PR d'un signal monostandard GSM: vision port /port et par symbole .

5 Conclusion

Partant du constat qu'un signal Radio Logicielle n'est rien d'autre qu'un signal multiporteuses, c'est à dire un multiplex de porteuses modulées, et utilisant l'analogie forte avec le signal OFDM, nous avons trouvé une borne supérieure pour le Power Ratio, puis nous avons proposé une vision porteuse par porteuse de ce paramètre. La généralisation de cette vision porteuse du PR à tout type de signal Radio Logicielle, lorsque les porteuses sont obtenues par Transformée de Fourier sur une durée égale à Ts est en cours d'étude. Cette vision rendra plus facile le calcul du PR, car généralement les puissances maximale et moyenne seront plus faciles à calculer de cette façon. Elle ouvre aussi la voie à de nouvelles interprétations des phénomènes mis en jeu ainsi qu'à de nouvelles méthodes de diminution de ce Power Ratio.

References

[1] J.Mitola, "The software Radio Architecture", IEEE Communications Magazine, May 95, pp. 26-38.

[2] Tim Hentschel, "Channelization for Software Defined Base Station", Annals of Telecom, 57, n°5-6, May and June.

[3] Xuefou Zhou, James Cffery, "A new distribution bound and reduction scheme for OFDM PAPR", http://www.ececs.uc.edu/ jcaffery/ofdm_papr.pdf

[4] Richard Van Nee, Ramjee Prasad, " OFDM for wirelless multimedia communications", universal personal communications, Artech House publishers, Chapter 6.

[5] Nati Dinur, Dov Wulich, "Peak-to-Average Power Ratioin High-Order OFDM", IEEE transactions on communications, Vol 49, No.6, June 2001, pp 1063-1072.

[6] Jacques Palicot and Yves Louët, "Power Ratio definitions and analysis in single carrier modulations", EUSIPCO September 2005.

Critères multi-échelles d'évaluation de la segmentation

Sylvie Philipp-Foliguet[1], Laurent Guigues[2]

[1]ETIS, CNRS UMR 8051, ENSEA, 6 avenue du Ponceau, 95014 Cergy-Pontoise Cedex, France

[2]CREATIS, CNRS UMR 5515, INSERM U 630, INSA, 7 rue Jean Capelle, 69621 Villeurbanne Cedex, France

philipp@ensea.fr, laurent.guigues@creatis.insa-lyon.fr

Résumé – Ce papier s'intéresse à la définition de nouveaux critères d'évaluation de la segmentation, dans le cas où l'on ne dispose pas de vérité-terrain. L'avantage de ces critères est de pouvoir comparer des images segmentées en fonction du niveau de détail recherché. Ces critères sont comparés aux principaux critères existants : les critères de Levine-Nazif [5], de Liu-Yang [6] et de Borsotti [2].

Abstract – This paper defines new criteria for segmentation evaluation when no ground truth is available. These new criteria are able to compare segmented images with respect to the level of detail needed by the application. They are compared with the main existing criteria : Levine-Nazif [5], Liu-Yang [6] and Borsotti [2] criterion.

1 Introduction

Devant le foisonnement de méthodes développées depuis plusieurs décennies pour la segmentation des images, le problème de l'évaluation est devenu crucial. Définir une bonne segmentation demeure difficile d'autant que la solution dépend du but recherché. La segmentation n'est pas une fin en soi et l'idée de qualité absolue, indépendante de toute application, est certainement vide de sens. Nous pensons donc que la seule façon réaliste d'aborder l'évaluation est de l'envisager comme une mesure de l'adéquation d'un résultat à un besoin donné. Or une des caractéristiques essentielles d'un besoin en segmentation - si ce n'est la caractéristique essentielle - est le niveau de détail attendu : l'application requiert-elle une description fine ou à grands traits de la scène ?

Nous considérons le cas où l'on ne dispose pas de vérité-terrain et où l'on cherche à évaluer la qualité relative de différentes segmentations. Tous les critères proposés à ce jour pour l'évaluation sans vérité-terrain se présentent sous la forme d'une unique mesure de qualité. Nous proposons pour notre part de *rendre explicite la dépendance de la qualité d'une segmentation à l'échelle requise par l'application*, en développant des critères qui sont des *fonctions* d'un paramètre réel qui se comporte comme un paramètre d'échelle.

2 Critères d'évaluation multi-échelles

D'une façon générale, le problème de segmentation d'image peut être envisagé comme un problème de modélisation par morceaux d'une image : modélisation constante, polynomiale, gaussienne, lisse par morceaux ... chaque région correspondant à un "morceau" du modèle. Une fois une classe de modèle sélectionnée (p.ex. constant par morceaux), la recherche du meilleur modèle peut toujours se formuler comme un problème d'optimisation : rechercher une partition P de l'image en régions et sur chaque région R de P le modèle M_R qui minimisent une certaine énergie totale $E(P)$. L'énergie doit évidemment prendre en compte la qualité de la modélisation en incorporant un terme $E_D(P)$ mesurant la distance entre le modèle et l'image. Toutefois, si l'on se contente d'une énergie d'adéquation du modèle aux données, la segmentation optimale est systématiquement la sur-segmentation absolue ou une partition proche. Par exemple, si l'on considère des modèles constants par morceaux, alors le modèle qui comporte une région par pixel, de valeur la valeur du pixel, est une solution exacte, de distance nulle à l'image. Pour obtenir un résultat utile, il faut alors incorporer un terme énergétique $E_C(P)$, dit de "simplicité", qui pénalise les segmentations trop fines, c'est-à-dire les modèles trop "complexes". Si l'on considère des modèles indépendants sur chaque région, on aboutit alors à des énergies qui prennent la forme générale :

$$E_k(P) = \sum_{R \in P} E_D(R) + k \times E_C(R) \qquad (1)$$

où k est un paramètre réel qui règle la contribution relative des deux termes énergétiques. Notons qu'outre contrôler la finesse de la solution, l'énergie de "simplicité" E_C permet de contrôler la régularité géométrique de la solution, en privilégiant par exemple les régions aux frontières peu tortueuses.

Dans ce cadre, le choix d'une segmentation relève d'un compromis entre adéquation aux données et simplicité du modèle et il n'y a pas intrinsèquement de meilleure solution : certaines applications peuvent avoir besoin d'un modèle précis, qui sera donc complexe, d'autres au contraire peuvent avoir besoin d'une description grossière, à grand traits, de l'image. Si l'énergie E_C est une fonction croissante avec la finesse de la partition, alors le paramètre k de l'équation (1) permet de contrôler la finesse de la

solution, c'est-à-dire se comporte comme un *paramètre d'échelle* : comme nous l'avons vu, si $k = 0$ on trouve un modèle très morcelé qui s'adapte parfaitement à l'image, à l'inverse, pour k assez grand, l'image est modélisée par une seule région. Suivant cette idée, il a été proposé dans [4, 3] de ne pas se contenter d'une seule solution de (1), pour une valeur de k fixée a priori, mais de rechercher une famille de segmentations sous la forme d'une séquence $\{P_k\}_{k\in R^+}$ de partitions de finesse décroissante avec k. On montre que c'est équivalent à rechercher une hiérarchie indicée dont les ensembles représentent les régions appartenant aux minima de (1) et dont l'indice représente la plus petite échelle pour laquelle une région de la hiérarchie appartient à une solution du problème de minimisation. Un algorithme efficace pour trouver une séquence de solutions localement optimales, dans un sens précis, a été proposé [4, 3].

Nous inspirant de ce travail, nous nous intéressons ici à un problème complémentaire du problème de segmentation : celui de l'évaluation de la qualité de résultats de segmentation.

Nous reprenons pour celà l'expression énergétique à deux termes (1), attache aux données $E_D(P)$ et simplicité $E_C(P)$, et proposons de caractériser une segmentation P par la fonction $E_k(P)$ qui à k associe $E_D(P) + kE_C(P)$. Il s'agit d'une fonction affine, croissante si E_C est positive. Si l'on dispose de deux segmentations P_1 et P_2 d'une même image, deux cas se présentent : soit $E_k(P_1) < E_k(P_2)$ pour tout k (ou inversement) auquel cas P_1 est systématiquement meilleure que P_2, soit il existe k_0 tel que $E_{k_0}(P_1) = E_{k_0}(P_2)$, auquel cas une des deux segmentations est meilleure aux petites échelles et l'autre meilleure aux grandes échelles. Cette analyse se généralise au cas où l'on dispose d'un nombre arbitraire n de segmentations. On montre facilement que l'ensemble des échelles k pour lesquelles une segmentation est meilleure que toutes les autres est un intervalle, éventuellement vide, qui représente la gamme d'échelles sur laquelle cette segmentation est la plus pertinente. Notre approche permet donc d'ordonner les segmentations en échelle et d'évincer les segmentations qui ne sont pertinentes pour aucune échelle.

3 Comparaison de différentes énergies

Afin d'évaluer la pertinence de notre approche, nous avons envisagé différentes formes d'énergie.

Soit R_i une région contenant A_i pixels notés $(X_1, X_2, ..., X_{A_i})$ et soit X_p^j la j-ème composante couleur du pixel X_p. Soit μ^j la moyenne de la composante j et V la matrice de variance / covariance des X_p de terme général : $V(j,k) = \frac{1}{A_i} \sum_{p=1}^{A_i} \left(X_p^j - \mu^j\right)\left(X_p^k - \mu^k\right)$. Nous notons λ_j la j-ème valeur propre de V.

Le premier modèle considéré est un modèle constant par morceaux et la distance modèle-image, i.e. l'énergie E_D, est mesurée en norme L_2. Pour une région donnée, le vecteur qui minimise la distance à l'image est bien entendu le vecteur moyen, et la distance est $Q(R_i) = A_i \cdot Trace(V) = A_i \sum \lambda_j$.

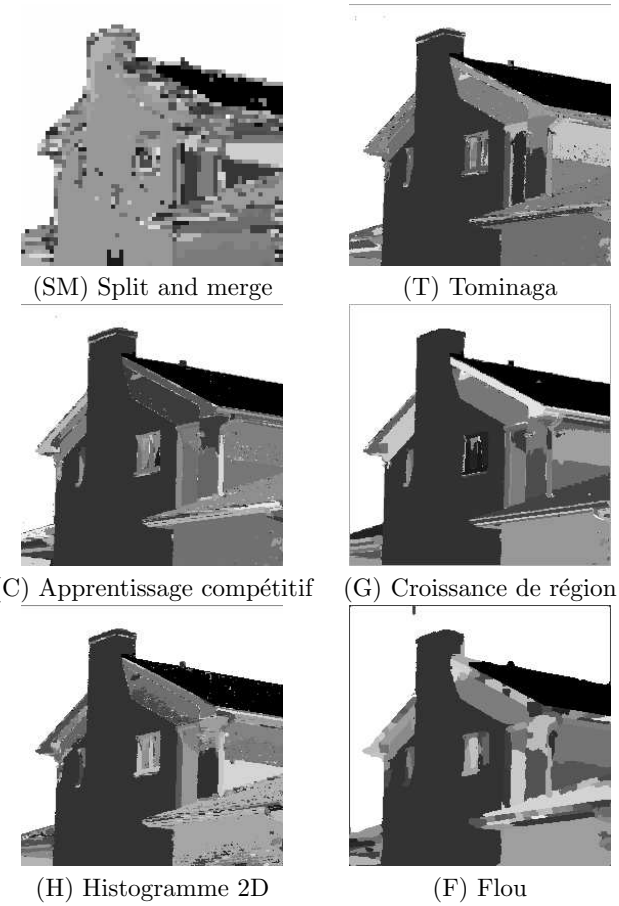


FIG. 1: *6 résultats de segmentation de l'image Maison*

Le deuxième modèle envisagé est un modèle probabiliste, qui suppose que l'échantillon des couleurs d'une région a été obtenu par tirage i.i.d. selon une loi gaussienne. L'énergie considérée est alors l'opposé du log de la vraisemblance des observations sachant le modèle. Les estimateurs optimaux pour la moyenne et la variance/covariance de la gaussienne sont alors les estimateurs empiriques et l'énergie d'une région s'écrit à une constante près : $G(R_i) = A_i \log(\det V) = A_i \sum_{j=1}^{3} \log(\lambda_j)$.

D'autres formes proches des deux formes précédentes peuvent être utilisées, nous emploierons dans nos tests $D(R_i) = A_i \det V = A_i \prod_{j=1}^{3} \lambda_j$.

La normalisation de l'énergie interne est différente pour les deux expressions : pour des images codées sur $2n$ valeurs par composante, n^2 est un majorant des valeurs de variance et covariance. La normalisation s'effectue donc en divisant par $n^2 \times 3 \times A \times 100$ pour l'énergie Q, par $n^6 \times 3 \times A \times 10000$ pour l'énergie D et par A pour l'énergie G.

En ce qui concerne l'énergie de simplicité, nous prenons simplement la longueur totale de tous les contours, comme dans le modèle de Mumford et Shah [8]. La normalisation

s'effectue par division par le nombre de pixels de l'image.

Nous avons comparé ces critères sur des résultats de segmentation obtenus par différents algorithmes. La figure 1 montre 6 résultats de segmentation obtenus par split and merge (SM), Tominaga (T), apprentissage compétitif (C), croissance de région (G) et classification d'histogramme 2D (H) (cf. [1]) et une méthode floue (F) [9] .

Nous calculons séparément les deux termes d'énergie (attache et simplicité) et nous comparons les différentes expressions de l'énergie d'attache. Le but est de confronter les résultats fournis par les 3 formes de l'énergie d'attache avec notre perception visuelle.

Si l'on représente dans le plan les couples (E_D, E_C) obtenus pour les 6 segmentations (Fig. 2), on constate que Split and merge est nettement discriminé par l'énergie Q, et moins bien par les autres formes d'énergie d'attache. La méthode floue est séparée des autres résultats pour les énergies Q et D, ainsi que par l'énergie contour.

Comme nous l'avons vu, le paramètre k de l'équation (1) est lié à la résolution ou au niveau de détail recherché. Les algorithmes qui donnent peu de régions ou des régions aux frontières régulières sont favorisés par l'énergie de simplicité et défavorisés par l'énergie d'attache aux données.

Dans la figure 2d, nous comparons les résultats en fonction du paramètre d'échelle k : G est toujours meilleur que C, T, SM et H, puisque les courbes E_k pour ces 4 segmentations sont toujours au-dessous de celle pour G. On peut conclure que si l'on recherche une segmentation grossière ($k > 25$) de l'image Maison, F fournit la meilleure segmentation, et que pour une segmentation plus précise, G donne le meilleur résultat.

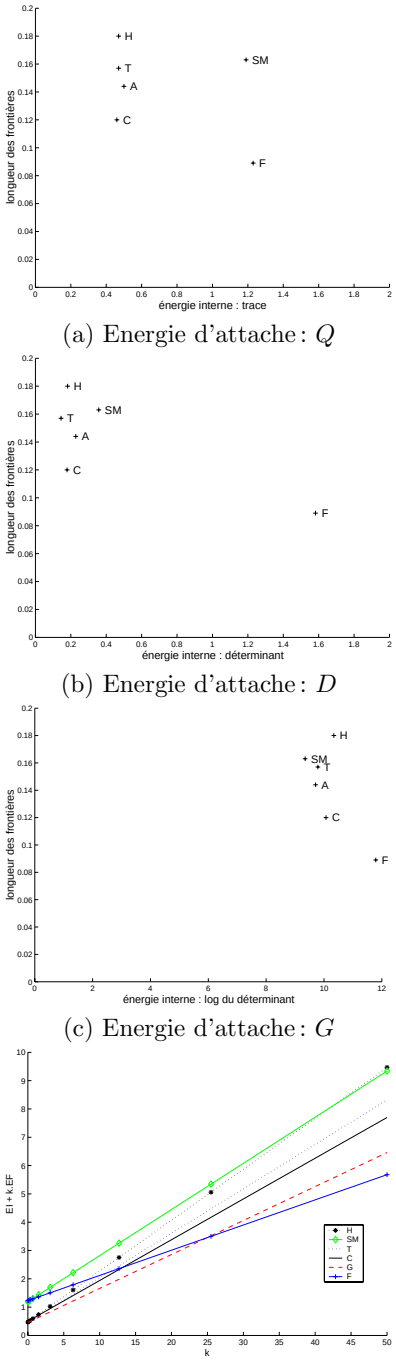
4 Comparaison avec différents critères existants

Nous avons comparé les critères proposés avec 3 critères habituellement utilisés : Levine-Nazif [5], Liu-Yang [6] et Borsotti [2], pour les 6 résultats de segmentation automatique de l'image Maison (cf. table 1) et pour les 5 segmentations manuelles de l'image "Tulipe" issues de la base de Berkeley [7](cf. figure 3 et table 2). Pour tous ces critères, la meilleure segmentation est celle de plus faible score.

	SM	T	C	G	H	F
nombre de régions	776	1375	1057	654	1384	139
Levine-Nazif	116	78	65	49	70	**31**
Liu-Yang	3.2	0.40	0.37	**0.25**	0.39	0.47
Borsotti	0.4	29	8	1.1	24	**0.1**
Energie $k = 10$	2.82	2.04	1.94	**1.66**	2.28	2.12
Energie $k = 100$	17.5	16.2	14.9	12.5	18.5	**10.2**

Table 1 : Comparaison des critères pour les 6 résultats de segmentation automatique de l'image Maison.

Les résultats sur l'image Maison mettent en évidence que le critère de Borsotti est très sensible aux toutes petites régions (un ou deux pixels). Il classe en premier la segmentation floue de même que Levine et Nazif. Le critère d'énergie classe en premier le résultat G pour un niveau de détail fin ($k = 10$) et le résultat F pour une résolution grossière ($k = 100$), comme montré dans la Fig. 2.



(a) Energie d'attache : Q

(b) Energie d'attache : D

(c) Energie d'attache : G

(d) E_k en fonction de k avec l'énergie d'attache Q

FIG. 2: *Résultats pour les 6 segmentations de l'image Maison.*

Pour l'image Tulipe, les 3 critères Levine-Nazif, Liu-Yang et Borsotti préfèrent la segmentation (a), qui est une caricature en 4 régions de l'image. Ces 3 critères privilégient donc les segmentations grossières. Comme le montre la figure 4, pour notre critère multi-échelles, la segmentation (e) est jugée la meilleure pour les petites échelles; aux échelles moyennes c'est la segmentation (c), et aux grandes échelles c'est la segmentation (a). Les segmentation (b) et (d), elles, ne sont jamais jugées pertinentes. On constate en effet visuellement qu'elles sont

hétérogènes en terme de niveau de détail : globalement grossières mais très détaillées en certains lieux précis.

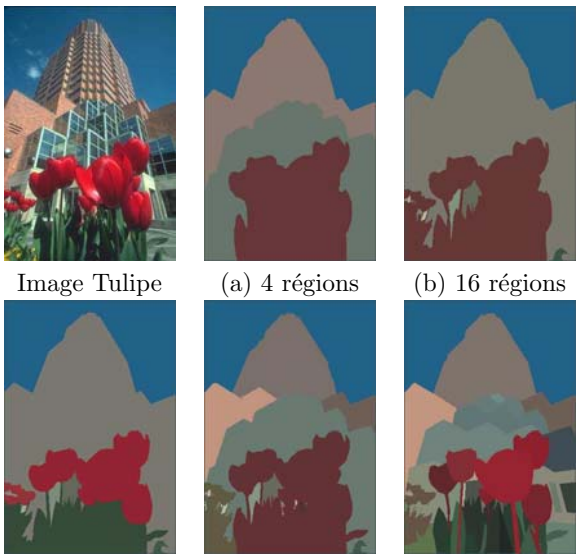


Image Tulipe (a) 4 régions (b) 16 régions

(c) 20 régions (d) 41 régions (e) 79 régions

FIG. 3: *5 segmentations manuelles de l'image "Tulipe" de la base de Berkeley.*

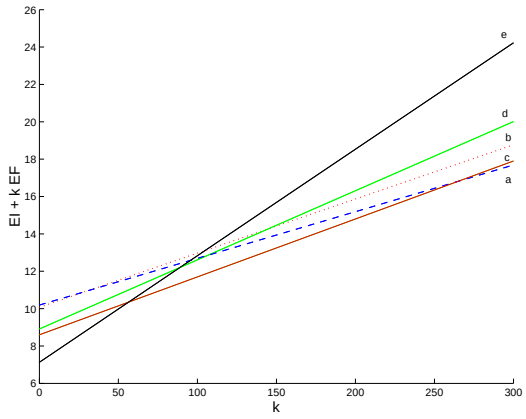


FIG. 4: *E_k en fonction de k (énergie d'attache Q) pour les 5 segmentations manuelles de la figure 3.*

	a	b	c	d	e
nombre de régions	4	16	20	41	79
Levine-Nazif	**2.76**	4.56	6.31	8.78	14.9
Liu-Yang	**0.06**	0.10	0.11	0.20	0.46
Borsotti	**0.1**	0.17	0.17	0.26	0.32
Energie $k = 10$	10.45	10.36	8.91	9.29	**7.70**
Energie $k = 100$	**12.7**	13	11.73	12.65	12.8

Table 2 : Comparaison des critères pour les 5 segmentations manuelles de l'image Tulipe.

5 Conclusion

La plupart des critères pour évaluer des résultats de segmentation prennent en compte d'une part les distances entre les couleurs d'une région et la couleur moyenne de la région et d'autre part une certaine complexité au travers du nombre de régions. Nous pensons que l'évaluation ne peut se faire qu'en fonction d'un but, dont un des éléments est le niveau de détail recherché.

C'est pourquoi nous avons proposé des critères d'évaluation liés au niveau de détail et qui prennent en compte à la fois la complexité de la segmentation et la proximité des régions extraites avec l'image originale. Le premier aspect est mesuré par la longueur des contours, qui représente à la fois le nombre de régions (liée à la résolution) et la régularité des contours. Le second aspect est l'attache aux données, pour laquelle nous avons envisagé trois expressions différentes. Après comparaison, il semble que le plus efficace parmi ces critères soit le critère qui correspond à un modèle constant par morceaux. Pour mesurer la complexité de la segmentation, des critères plus sophistiqués que la simple longueur des contours peuvent être employés.

De plus ces critères sont très simples à calculer, et s'adaptent à tout type d'images, monochromes ou multispectrales, avec ou sans contours entre les régions.

Remerciements : Les auteurs remercient L. Macaire pour les images segmentées de Maison. Si le nombre de régions est différent entre [1] et nos résultats, c'est probablement parce que nous comptons les régions 4-connexes alors qu'elles sont comptées en 8-connexité dans [1].

Références

[1] A. Trémeau, C. Fernandez-Maloigne, and P. Bonton. *Image numérique couleur*. Dunod, Paris, 2004.

[2] M. Borsotti, P. Campadelli, and R. Schettini. Quantitative evaluation of color image segmentation results. *Pattern Recognition Letters*, 19:741–747, 1998.

[3] L. Guigues. *Modèles multi-échelles pour la segmentation d'images*. PhD thesis, Université de Cergy-Pontoise, 2003.

[4] Laurent Guigues, Hervé Le Men, and Jean-Pierre Cocquerez. Scale-sets image analysis. In *Proc. of IEEE Int. Conf. on Image Processing (ICIP'03), Barcelona, Spain*, September 2003.

[5] M.D. Levine and A.M. Nazif. Dynamic measurement of computer generated image segmentations. *IEEE Trans. on PAMI*, 7(25):155–164, 1985.

[6] J. Liu and Y.-H. Yang. Multiresolution color image segmentation. *IEEE Trans. on PAMI*, 16(7):689–700, 1994.

[7] D. R. Martin. *An empirical approach to grouping and segmentation*. PhD thesis, University of California, Berkeley, USA, 2002.

[8] D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and asociated variational problems. *Comm. Pure Appl. Math.*, 42:577–685, 1989.

[9] S. Philipp-Foliguet, M. B. Vieira, and M. Sanfourche. Fuzzy segmentation of color images and indexing of fuzzy regions. In *First Europ. conf. on Colour in Graphics, Imaging and Vision*, pages 507–512, Poitiers, France, 2002.

Segmentation pyramidale localisée dans un ruban fermé

Pascal BERTOLINO[1], Jérémy HUART[1], Guillaume FORET[2]

[1]Laboratoire des Images et des Signaux,
BP 46, 38402 Saint Martin d'Hères, France

[2]Equipes Traitement des Images et du Signal,
BP 44, 95014 Cergy-Pontoise, France

pascal.bertolino@inpg.fr, jeremy.huart@inpg.fr

Résumé – Nous montrons comment un processus de segmentation peut être utilisé localement sur des régions d'intérêt d'une image, délimitées grossièrement par l'utilisateur. L'algorithme de la pyramide irrégulière convient particulièrement bien à cette approche : une propagation concurrente d'étiquettes de *racines* dans un graphe de régions permet de classifier les pixels à l'intérieur ou à l'extérieur de l'objet considéré.

Abstract – This paper shows how a segmentation process can be locally used within regions of interest roughly delineated by the user, in a still image. The graph pyramid (aka irregular pyramid) algorithm is used to perform a concurrent root label propagation that classifies each pixel inside or outside a given object.

1 Introduction

La segmentation d'images naturelles précise et efficace peut rarement être effectuée automatiquement ; c'est pourquoi le recourt à la segmentation interactive [1, 2] est souvent nécessaire dans de nombreuses applications : son but est d'extraire un ou plusieurs objets d'intérêt en proposant à l'utilisateur potentiel un outil simple et pratique à utiliser fournissant des résultats facilement exploitables.

Les contours actifs [3, 4] sont des méthodes très classiques qui nécessitent une initialisation avec un contour grossier à l'intérieur ou à l'extérieur de l'objet. Malheureusement, les snakes sont généralement très sensibles aux minima locaux et assez difficiles à paramétrer. La ligne de partage des eaux [5] quant-à-elle, fournit une sur-segmentation de l'image. Dans [6], les auteurs proposent un traitement interactif où les marqueurs sont sélectionnés par des clics avec la souris. Dans [1], la LPE fournit une segmentation multi-échelle dans laquelle l'utilisateur peut sélectionner des régions d'intérêt. Dans [7], après un seuillage automatique, les contours des régions sont représentés par des courbes paramétriques. Leurs points de contour peuvent être édités par l'utilisateur. Les ciseaux intelligents ont été développés par [2] pour la composition. La méthode détecte un contour d'objet potentiel proche du pointeur de souris et trouve le meilleur chemin partant d'un pixel de départ.

Dans notre approche, une croissance de régions est appliquée uniquement près du bord de l'objet indiqué grossièrement par l'utilisateur.

2 Pyramide d'image irrégulière *vs* pyramide locale

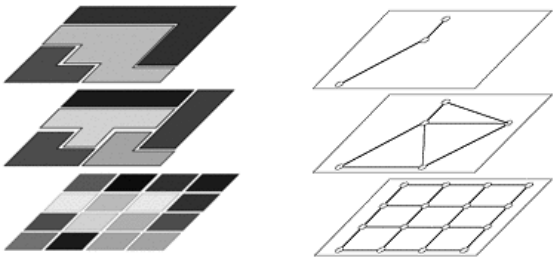
La pyramide irrégulière [8] [9] est un outil de segmentation puissant. Le principe de cette méthode est d'initialiser un graphe d'adjacence, où chaque sommet correspond à une région monopixelique de l'image. Utilisant un algorithme de type parallèle, les régions voisines similaires peuvent fusionner, entraînant une décroissance du nombre de sommets (figure 1.a). Les régions et sont similaires ($_i \sim _j$) si la distance entre leur couleur moyenne, dans l'espace YUV, est plus faible qu'un certain seuil : $_i \sim _j$ si $(\quad (_i) \quad (_j)) \quad$. Le traitement, qui porte sur toute l'image, est itéré jusqu'à stabilisation du nombre de régions.

Dans la pyramide locale que nous proposons, seul un nombre réduit de pixels de l'image sont identifiés à des sommets, tandis que les composantes connexes restantes sont identifiées à sommets (figure 1.b et c), et sont des *racines* (régions qui appartiendront au résultat final) tel que le fond par exemple. Cette focalisation sur une zone particulière de l'image nécessite une interaction humaine, et peut être réalisée à la souris en marquant grossièrement le contour de la zone à l'aide d'un ruban fermé (figure 2.a).

3 Propagation des étiquettes

Une fois dessiné, le ruban induit trois composantes connexes, soit trois étiquettes : l'extérieur, l'intérieur et le ruban lui-même, supposé contenir le vrai contour de l'objet (figure 2.b). Les pixels de l'extérieur appartiennent au fond (le premier sommet racine) tandis que ceux de l'intérieur appartiennent à l'objet d'intérêt (un second sommet racine). Tous les pixels formant le ruban représentent une zone indéfinie. Ces derniers seront segmentés afin de fusionner avec l'une ou l'autre des racines.

A la base de la pyramide locale, le graphe d'adjacence inclut un sommet pour chaque racine et un sommet pour chaque pixel de la zone indéfinie (figure 1.c). A chaque racine est associée une étiquette unique. Les sommets (i.e. pixels) de la zone indéfinie restent non étiquetés. Une fine couche de pixels de part et d'autre du ruban sont séparés des racines et constituent autant de sommets portant l'étiquette de la racine correspondante (figure 2.d). Ce sont ces deux couches qui vont permettre

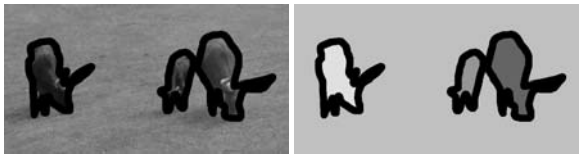


(a) Exemple de pyramide de graphe construite sur une image 4 × 4 : empilements de partitions et de graphes

(b) Exemple d'une initialisation de pyramide locale

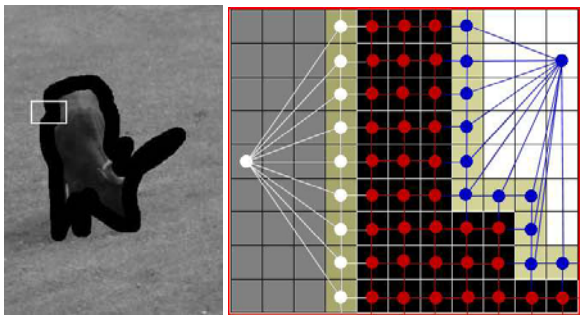
(c) Graphe correspondant à (b)

FIG. 1 – *Pyramide d'image vs pyramide locale*



(a) Initialisation de régions d'intérêt

(b) Etiquetage initial des racines induites par (a). Les zones noires n'ont pas d'étiquette



(c) Le rectangle est zoomé dans (d)

(d) Graphe d'adjacence de la zone marquée dans (c)

FIG. 2 – *De l'initialisation manuelle d'une zone d'intérêt à la définition d'un graphe d'adjacence où sera effectuée la segmentation*

(en fusionnant avec d'autres pixels) de propager les étiquettes des deux racines dans la zone indéfinie. Durant la construction de la pyramide locale, les fusions s'effectuent toujours selon le critère de similarité, mais respectent également des règles qui assurent la propagation cohérente des étiquettes à travers la zone indéfinie.

Le regroupement de sommets [8] se fait localement selon une phase de décimation (des sommets disparaissent) et une phase de rattachement (les pixels des sommets disparus se rattachent au sommet voisin survivant le plus similaire). Le nombre de sommets se regroupant en un nouveau sommet n'est pas fixé *a priori*. Les phases de décimation et de rattachement nécessitent que la similarité soit évaluée entre tous couples de sommets. Si sommets (≥ 2) sont similaires, ils pourront éventuellement être fusionnés en une seule et même région, selon les phases décimation/regroupement. Ainsi, la notion de similarité est capitale ; rajouter de nouvelles contraintes dans les règles de construction de la pyramide peut simplement être vu comme modifier le critère de similarité. Pour prendre en compte ces nouvelles contraintes, on modifie le critère vu en section 2 et on rajoute des règles concernant la propagation des étiquettes. Par la suite, ( ) est une région portant l'étiquette et est une région non étiquetée ; $\sim$ est le symbole de similarité.

- **Règle de similarité :** $_i(\ _i) \nsim\ _j(\ _j)$ si $_i \neq\ _j$ ou si (($_i$) ($_j$)) $\geq$. Deux régions portant deux étiquettes différentes ne peuvent pas fusionner, ceci pour éviter la fusion de différents objets.
- **Règle de propagation 1 :** Si $_i \sim\ _j(\ ) \ _i \cup\ _j(\ ) =\ _k(\ )$. Une région non étiquetée fusionnant avec une région étiquetée donne naissance à une région portant l'étiquette .
- **Règle de propagation 2 :** Si $_i(\ ) \sim\ _j(\ ) \ _i(\ ) \cup\ _j(\ ) =\ _k(\ )$. La fusion de plusieurs régions d'étiquette donne naissance à une région portant l'étiquette .
- **Règle de propagation 3 :** Si $_i \sim\ _j \ _i \cup\ _j =\ _k$. Des régions non étiquetées fusionnant entre elles donnent naissance à une région non étiquetée.

Lorsque plus aucune fusion n'est possible et qu'il reste éventuellement des sommets non étiquetés, la règle de similarité est relâchée : $_i(\ _i) \nsim\ _j(\ _j)$ ssi $_i \neq\ _j$, $_i(\ _i) \sim\ _j(\ _j)$ sinon. La partition finale comporte alors autant d'objets que de racines.

4 Résultats

Il est naturel que la méthode fonctionne bien dans le cas d'objets homogènes sur un fond homogène.

Beaucoup plus intéressant, la méthode fonctionne bien dans l'hypothèse où les objets ont une colorimétrie hétérogène sur un fond relativement homogène (ou inversement), même lorsque le signe du gradient s'inverse à la frontière entre le fond et l'objet. Les dégradés sont aussi bien pris en compte. Cette approche est intéressante notamment lorsque les objets ont une forme complexe.

Les résultats présentés ne sont pas post-traités, et sont obtenus avec le même seuil de similarité . La figure 3.b montre comment il est souvent impossible avec des méthodes simples (ici un seuillage) d'extraire correctement un objet, soit à cause de sa texture, de son ombre portée ou du manque de contraste localement. On peut remarquer que le contour du dos de l'animal (figure 3.a) n'est pas du tout contrasté, ni en couleurs, ni en niveaux de gris.

Les résultats (figures 3 et 4) montrent que notre méthode est peu sensible à la façon dont l'utilisateur initialise le traitement :

en effet, une variabilité importante de l'épaisseur du ruban, de son positionnement (plutôt intérieur ou extérieur), de sa forme (plus ou moins régulière), fournit des résultats très similaires. Le ruban peut être tracé soit à l'aide de segments (figures 3.c,e, 4.a,c), soit à main levée (figure 4.e,g).

Les figures 3.d et f présentent deux résultats obtenus avec deux épaisseurs différentes (10 et 40 pixels). Le ruban de la figure 3.e est obtenu avec 10 clics de la souris seulement ; en revanche, il fournit des détails précis de la tête et des jambes de l'animal.

Dans la figure 4.a, la segmentation est obtenue avec un ruban positionné plutôt dans l'objet alors que dans la figure 4.c, il est plutôt dans le fond. Bien que ces résultats ne soient pas exactement identiques, ils diffèrent assez peu.

Dans les figures 4.e et g, on a varié la forme du ruban : l'utilisateur peut effectivement dessiner des "pâtés" qui contiennent de nombreux détails de contour afin de les extraire correctement, comme le montre la jambe arrière dans la figure 4.g.

Enfin, la méthode présentée peut être utilisée telle quelle, sans modification, avec un nombre quelconque de racines et un ruban fermé peut avoir une topologie quelconque et être connexe à un nombre quelconque de racines. comme le montre l'exemple de la figure 5. L'approche proposée garantit un nombre d'objets final égal au nombre de racines. Il est donc important que le ruban soit fermé pour que l'intérieur puisse se distinguer de l'extérieur.

5 Discussion

Nous proposons une méthode bien adaptée aux images complexes où le contour d'un objet peut successivement prendre des configurations très différentes : gradient positif, négatif, contours multiples. La figure 6 montre un résultat obtenu avec notre application ExtraK'Obs[1] pour un objet qui se distingue assez difficilement du fond.

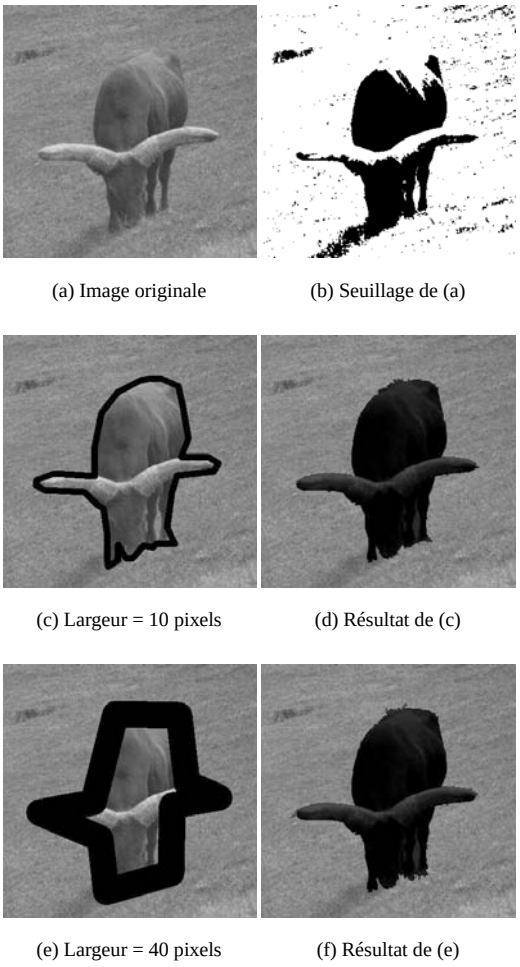
La texture du fond et/ou de l'objet est bien entendu un problème important auxquelles doivent faire face la majorité des méthodes de segmentation et notamment celle que nous présentons. La (re-)connaissance de présence de textures (objet ou fond) et de contours pourrait servir à favoriser certaines fusions lors de la propagation concurrente d'étiquettes. Il semblerait par exemple judicieux de favoriser la propagation dans les zones peu texturées qui pourraient alors s'étendre jusqu'à être stoppées par des zones plus texturées.

La localisation de la frontière finale n'est pas liée géométriquement à la localisation et à l'épaisseur du ruban. Ces deux paramètres peuvent néanmoins influencer le résultat final dans le sens où une localisation et une largeur mieux adaptées limiteront les erreurs de segmentation (faux contour, fuites, ...). La frontière finale correspond localement à la frontière entre deux régions qui ont crû puis acquis une étiquette. Ainsi la croissance de régions non encore étiquetées est un élément essentiel que nous devons étudier de manière plus approfondie.

6 Conclusion

Nous avons montré dans ce travail préliminaire une méthode associée à une interaction qui fournit à un utilisateur un outil permettant une extraction facile et précise d'objets. Maintenant, pour améliorer ces premiers résultats, nous envisageons deux pistes différentes : tout d'abord générer de façon automatique les zones d'incertitude par un pré-traitement des images. Ensuite, nous voulons intégrer d'autres critères permettant de régulariser localement les contours détectés pendant la segmentation.

1. *http : //www.lis.inpg.fr/pages_perso/bertolino/software.php*



(a) Image originale (b) Seuillage de (a)

(c) Largeur = 10 pixels (d) Résultat de (c)

(e) Largeur = 40 pixels (f) Résultat de (e)

FIG. 3 – *Une variabilité importante de l'épaisseur du ruban entraîne une faible variabilité des résultats*

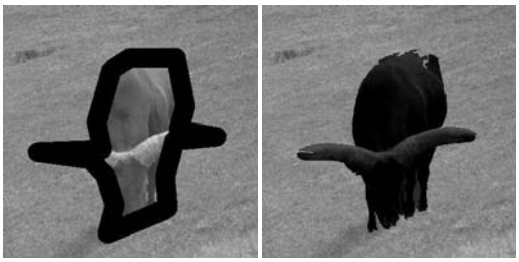
Références

[1] F. Zanoguera, B. Marcotegui, and F. Meyer, "A toolbox for interactive segmentation based on nested partitions," in *IEEE International Conference on Image Processing, ICIP'99*, Kobe, Japan, 1999.

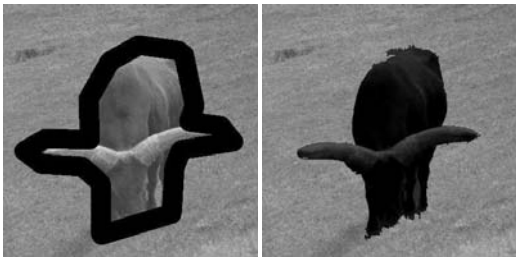
[2] Eric Mortensen and Wiliam A. Barrett, "Intelligent scissors for image composition," in *Proc. of the ACM SIGGRAPH 95 : Computer Graphics and Interactive Techniques*, Los Angeles, August 1995, pp. 191–198.

[3] M. Kass, A. Witkins, and D. Tersopoulos, "Snakes: Active contour models," *Computer Vision Graphics and Image Processing*, pp. 321–331, 1988.

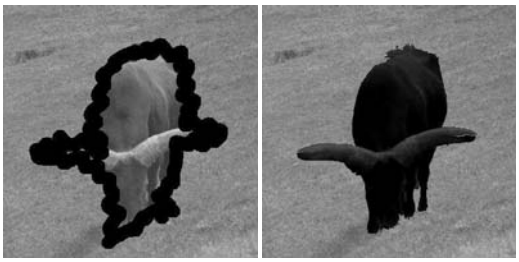
[4] A.K. Jain, Y. Zhong, and M.P. Dubuisson-Jolly, "Deformable template models: A review," *Signal Processing*, vol.



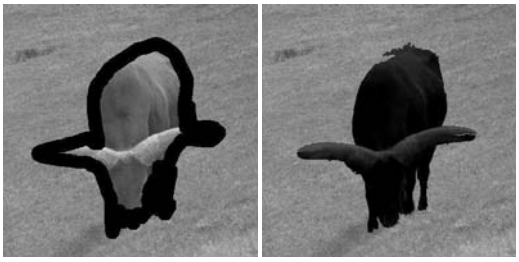
(a) Ruban intérieur (b) Résultat de (a)



(c) Ruban extérieur (d) Résultat de (c)



(e) Ruban bruité (f) Résultat de (e)



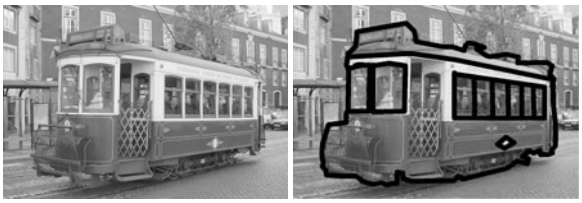
(g) Ruban irrégulier (h) Résultat de (g)

FIG. 4 – *Une variabilité importante du positionnement et de la régularité du ruban entraîne une faible variabilité des résultats*

(a) Image originale (b) Initialisation

(c) Partition obtenue

FIG. 5 – *Segmentation de plusieurs régions d'intérêt*



(a) Image originale (b) Initialisation

(c) Partition obtenue

FIG. 6 – *Exemple de résultat obtenu avec l'application Extra-K'Obs*

71, no. 2, pp. 109–129, 1998.

[5] S. Beucher and F. Meyer, "The morphological approach to segmentation: the watersheds transformation in e.r. doughterty (ed.)," in *Mathematical Morphology in Image Processing*, New-York: Dekker, 1992, pp. 433–481.

[6] J. Cutrona and N. Bonnet, "Two methods for semi-automatic segmentation based on fuzzy connectedness and watersheds," in *IASTED International Conference on Visualisation, Imaging and Image Processing VIIP' 2001*.

[7] M. Jackowski, A. Goshtasby, and M. Satter, "Interactive tools for image segmentation," in *SPIE's International Symposium on Medical Imaging*, San Diego, USA, 1999.

[8] A. Montanvert, P. Meer, and A. Rosenfeld, "Hierarchical image analysis using irregular tessellations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 307–316, April 1991.

[9] J.M. Jolion and A. Montanvert, "The adaptive pyramid a framework for 2d image analysis," *CVGIP Image Understanding*, vol. 55, no. 3, pp. 339–348, 1992.

Extraction de formes par approche agent dans des IRM

Y. JACQUELET[1], J.-J. MONTOIS[1], M. L. NDIAYE[1], A. KINIE[1]

[1]LTSI INSERM U642, antenne de Saint Malo, IUT de Saint Malo, rue de la croix Désilles, BP 195, 35409 Saint Malo Cedex

{yann.jacquelet, jean-jacques.montois, abel.kinie}@univ-rennes1.fr, mamadou.ndiaye@etudiant-univ-rennes1.fr

Résumé – *La reconnaissance des formes au niveau relationnel est de nature combinatoire et peut donc être modélisée par un problème de satisfaction de contraintes. L'espace de recherche associé est issu de la segmentation d'IRM et du filtrage des régions obtenues. La transposition du réseau sémantique dans l'espace comportemental agent permet l'intégration de l'expérience du spécialiste guidant ainsi le système vers les solutions pertinentes. Les performances de l'éco-résolution sont comparées à celles de trois autres recherches stochastiques classiques. La comparaison met en évidence un nombre moyen et maximum de déplacements très restreint.*

Abstract *– The shape relational recognition process is a combinatorial problem. It may be formalised with the constraint satisfaction problem theory. The Regions resulting from MRI segmentation get filtered in order to restrict the size of the associated search space. The transposition of the semantic network in a behavioural space enables the specialist to practice his skills at extracting anatomical shapes and leads the system to the suitable solutions. The performances of the eco-resolution framework are compared with three other stochastic algorithms. The results show that the required moves are few in number, both in the average and in the worst case.*

1. Introduction

Dans les systèmes de vision, la segmentation est sous-contrainte, comme de nombreux autres problèmes. Les techniques récentes (« snakes », champs de markov, « level-set ») mettent plus fréquemment en œuvre des processus d'optimisation permettant de re-contraindre le problème par l'apport d'informations *a priori* (de type photométrique, géométrique ou relationnel) sur les objets de la scène. Les opérateurs de traitement semblent évoluer vers une plus grande adaptabilité et le problème de segmentation vers un problème de contrôle.

L'évaluation de la qualité des régions extraites par un processus de segmentation demeure un problème ouvert. Le couplage entre opérateurs de reconnaissance de formes et de segmentation peut permettre de sélectionner efficacement les régions qui représentent un réel intérêt. Sur ce principe, nous avons construit un système de segmentation exploitant les pyramides irrégulières de type combinatoire, qui permettent de produire des segmentations de l'image à des niveaux d'analyse différents. Nous nous intéressons à la localisation d'objets de scène multiples possédant chacun leurs propres *a priori*. L'étape de reconnaissance se décompose en deux parties, la première servant à filtrer la base de régions suivant des gabarits photométriques et géométriques, la seconde cherchant à satisfaire les contraintes relationnelles d'un réseau sémantique basé sur des relations d'orientations spatiales. Le système doit permettre d'intégrer l'expérience ainsi que les intuitions du spécialiste tout en aboutissant à une solution vérifiant le réseau sémantique.

Les approches de type systémique en vision [1] conviennent à ce type de problème, par nature combinatoire. En segmentation, dans [2] l'accent est porté sur les différentes sortes de coopérations entre les opérateurs de traitement d'images utilisés afin d'améliorer le résultat final. Une autre approche [3], privilégiant les aspects organisationnels, transpose le graphe d'adjacence en réseau d'accointances. Ce dernier apporte une plus grande souplesse pour l'intégration d'une coopération région-contour.

Il existe d'autres approches agents qui ont fait leurs preuves dans des domaines différents. En particulier, l'éco-résolution [4] propose une approche réactive pouvant s'adapter à de nombreuses situations. L'état initial du problème est réorganisé itérativement de manière à faire émerger la solution. Dans le jeu de taquin [5], les recherches heuristiques (toutes dérivées de l'algorithme A*) ne permettent pas de résoudre des jeux de taille > 99, du fait de la croissance exponentielle des espaces de recherche. L'éco-résolution, par sa réactivité, a permis de proposer des solutions sous-optimales pour des jeux de taille ≤ 899. En physique [6], la compréhension de la structure d'un empilement de matériaux granulaires à l'équilibre statique nécessite de résoudre des problèmes de satisfaction de contraintes dont la taille (plusieurs centaines de milliers de variables et autant de contraintes associées) est trop importante pour les solveurs classiques. L'adaptation de l'éco-résolution à ce problème a permis de fournir un simulateur dont le temps d'exécution est linéaire en fonction du nombre de grains. Dans [7], c'est la génération d'interfaces graphiques qui est l'objet de l'application. Au-delà des bonnes performances du système, l'auteur met en lumière la capacité de la méthode à intégrer l'expérience de l'utilisateur tout en adaptant ses propositions, de manière à les rendre opérationnelles.

L'éco-résolution, par ses performances prometteuses ainsi que sa capacité à intégrer les intuitions de l'utilisateur dans le choix de l'initialisation, semble être un bon candidat pour notre système. Dans cette optique, le problème combinatoire est formalisé au §2. La solution proposée au §3 s'appuie sur ce même formalisme pour décrire notre modélisation en éco-résolution. Cette approche comportementale est évaluée au sein d'un groupe de trois autres recherches stochastiques classiques au §4. Enfin, la conclusion et les perspectives vis-à vis de ces travaux sont présentées au §5.

2. Problème combinatoire

Notre système a été utilisé dans le cadre d'une étude médicale permettant la discrimination entre différentes variantes de syndromes parkinsoniens.

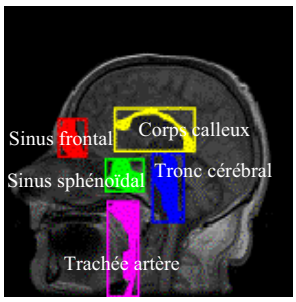


FIG. 1 : Les objets de la scène

La localisation du tronc cérébral (TC) est l'objectif prioritaire du système. Les autres structures voisines permettent de contraindre la recherche (corps calleux [CC], sinus frontal [SF], trachée artère [TA] et sinus sphénoïdal [SS]). (cf. figure 1)

A partir de la base IRM exploitée, il est possible d'apprendre les relations d'un réseau sémantique anatomique décrivant la scène observée. Connaissant les relations spatiales qui doivent nécessairement être vérifiées par les objets de la scène, nous recherchons alors une instance de ce réseau dans une des images de la base. Après avoir procédé à une reconnaissance des formes (critères photométriques et géométriques), il s'agit donc de confronter les solutions trouvées, aux connaissances relationnelles de la scène.

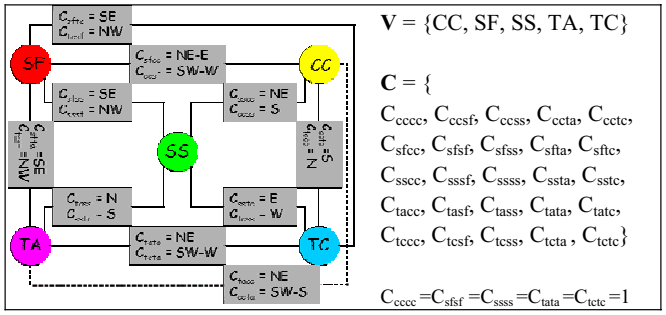


FIG. 2 : Problème de satisfaction de contraintes

Le problème à résoudre peut se formaliser comme un problème de satisfaction de contraintes (CSP) et le réseau sémantique anatomique sous la forme d'un réseau de contraintes (cf. figure 2). Pour cela, il faut identifier le quadruplet *contraintes* × *variables* × *domaines* × *valeurs*. Les **contraintes** sont issues des prédicats d'orientation (N, S, E, W, …) du réseau sémantique dont les valeurs sont booléennes (1 pour satisfaite, 0 sinon). Les **variables** représentent les différents objets de la scène intervenant dans les équations des contraintes : CC, SF, SS, TA et TC. Les **domaines** regroupent l'ensemble des valeurs prises par les variables. Les **valeurs**, quant à elles, correspondent aux rectangles englobants des régions issues du filtrage de chacun des objets de la scène (d'après des critères photométriques et géométriques). On précise que chaque domaine possède une valeur virtuelle supplémentaire (-1) signifiant que la variable n'est pas encore affectée. L'illustration des variables et de leurs contraintes est présentée en figure 2, celle des domaines en figure 3.

Il est maintenant possible de décrire les images sous forme de graphe d'états. Un état est un vecteur donnant les valeurs des variables (ordre lexicographique), $x = (cc, sf, ss, ta, tc)$. La fonction $C(x)$ permet de lui associer une matrice de satisfaction des contraintes. On s'intéresse alors à la fonction de voisinage $V(x) = \{x' / nb(x_i \neq x_i') = 1, \forall i\}$. Par la suite on parlera de mouvement pour le déplacement d'un état x vers un autre état voisin x'. Le graphe ainsi défini est connexe ; tout état du graphe (dont on connaît les coordonnées) est accessible depuis n'importe quel autre, en 5 mouvements au plus. Trouver une solution pour le CSP revient donc à localiser un état x tel que $\sum\sum C_{ij}(x) = card(C(x))$ et dont aucune des coordonnées ne soit à -1. Le problème consiste à rechercher un état compatible avec toutes les contraintes mais dont on ignore les coordonnées. Afin de suivre l'évolution des mouvements du système, les états du graphe ont été regroupées en six catégories (T0 à T5) représentant le nombre d'agents dont la position est compatible avec tous les autres. Les états de satisfaction sont de type T5 dans cette nomenclature.

3. Modélisation agent

Nous effectuons une transposition de l'espace calculatoire des graphes vers un espace comportemental agent. Le réseau de contraintes se transforme alors en un réseau d'agents en accointances. L'idée est de renforcer « l'intelligence » et donc la réactivité des nœuds, permettant ainsi la définition de nouveaux algorithmes de parcours centrés sur eux. Ceci modélisent des points de vu locaux.

TAB. 1 : Modèle comportemental de l'éco-résolution

Méthode satisfaire(EcoAgent x) **Tant que** x ≠ ∅ • x⇨essayerEtreSatisfait() • x ← (x⇨*dépendance()*)	**Méthode** essayerEtreSatisfait() **Si** ¬*satisfait*() **Alors** • x⇨essayerFuir(*but()*),∀x∈*gêneurs()* • *êtreSatisfait()*
Méthode essayerFuir(Contrainte c) • p ← *placePourFuir(c)* • x⇨essayerFuir(p),∀x∈*gêneurs(p)* • *fuir (p)*	**Méthodes du domaine d'application :** *i. satisfait() & but() & êtreSatisfait()* *ii. gêneurs() & dépendance()* *iii. placePourFuir() & fuir()*

Le modèle éco-résolution caractérise des agents réactifs (mémoire limitée et décisions simples) dont le comportement se traduit soit par des agressions soit par des mouvements de fuite. Le modèle se compose de plusieurs méthodes (cf. tableau 1), certaines d'entre elles sont génériques (*satisfaire*, *essayerEtreSatisfait* et *essayerFuir*) et d'autres sont dédiées au domaine d'application (*satisfait*, *placePourFuir*, …).

Le travail de modélisation nécessite de répondre aux questions suivantes. Qu'est-ce qu'un agent, une place ou une contrainte ? Quand un agent est-il satisfait ? Pourquoi agresser ? Quels sont les gêneurs ? Comment fuir ? C'est ce que fait justement le tableau 2.

Les agents sont les variables du CSP, et les places correspondent à leurs valeurs. Dans ce schéma, le fait qu'un agent occupe une place implique la réservation d'une ou plusieurs autres (éventuellement aucune) dans les domaines de définition des autres agents (cf. figure 3).

TAB. 2 : Clés de la spécialisation du modèle

L'état courant	x
i⇨dépendance()	{i+1 si $1 \leq i < card(x)$ {∅ sinon
i⇨satisfait()	$\Leftrightarrow \sum C_{ij}(x) = card(x), \forall j$
i⇨but()	i
i⇨gêneurs()	$\{j / C_{ij}(x) \neq 1\}$
i⇨êtreSatisfait()	rien à faire
nb(e)	comptabilise le nombre d'événement e
$V_i(x)$	$\{x' / x_i \neq x_i'\}$
$E_i(p)$	$x' / x_j' = x_j, \forall j \neq i \wedge x_i = p$
G_{ic}	$\{p \in V_i(x) / C_{ic}(E_i(p)) = 1\}$
estRéservée(i,p)	$\Leftrightarrow \sum C_{ji}(E_i(p)) \neq card(E_i(p)), \forall j$
N_i	$\{p \in V_i(x) / \neg$estRéservée(i, p)$\}$
O_i	$\{p \in V_i(x) /$ estRéservée(i, p)$\}$
$M_i(P)$	arg max $\sum\sum C_{kj}(E_i(p)), \forall j, \forall k, \forall p \in P$
i⇨placePourFuir(c)	{ M_i ($N_i \cap G_{ic}$) si $N_i \cap G_{ic} \neq \varnothing$ { M_i ($O_i \cap G_{ic}$) sinon si $O_i \cap G_{ic} \neq \varnothing$ {-1 sinon
i⇨gêneurs(p)	$\{j / C_{ij}(E_i(p)) \neq 1\}$
i⇨fuir(p)	$x = E_i(p)$

Un agent est satisfait lorsqu'il occupe une place compatible avec tous les autres. Si ce n'est pas le cas, il agresse tous les gêneurs. Un agent *i* agressé doit fuir vers des places non réservées et satisfaisant la contrainte *c* de l'agresseur. Il choisit une place de la façon suivante : Il partitionne son voisinage en deux ensembles N_i (l'ensemble des places compatibles) et O_i (l'ensemble des places non compatibles). Puis il construit G_{ic}, l'ensemble des places vérifiant la contrainte *c* de l'agresseur. L'agent cherche la meilleure place au sens du critère M_i (cf. tableau 2) dans $N_i \cap G_{ic}$. Si ce n'est pas possible il recommence dans $O_i \cap G_{ic}$ et en dernier recours il opte pour la place *-1*.

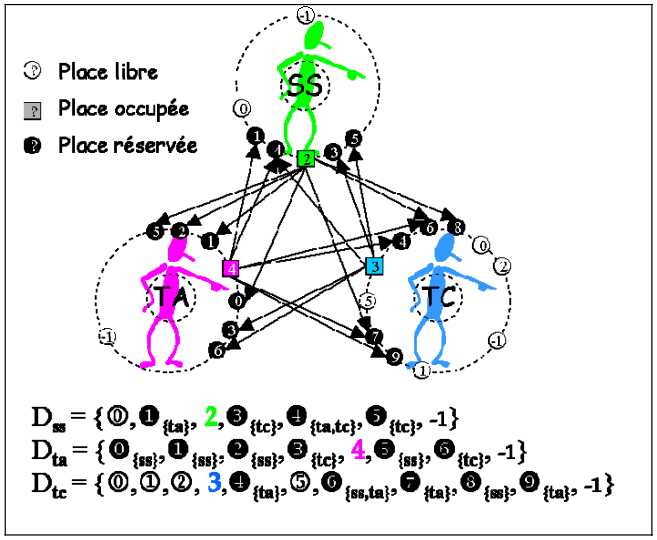


FIG. 3 : Trois éco-agents se contraignant mutuellement

4. Résultats

Nous disposons d'une base d'une cinquantaine de coupes IRM sagittales médianes (512x512) Spin écho T1 (5mm) au format DICOM. Cette base a été acquise auprès du CHU de Rennes qui est l'un des centres collecteurs pour le projet NNIPPS (« Neuroprotection and Natural history In Parkinson Plus Syndromes »). Ces coupes ont permis la construction des différents *a priori* du système. Dans notre étude, la caractérisation du comportement de l'éco-résolution, a été réalisée sur une IRM dont le problème combinatoire sous-jacent est de dimension moyenne. Elle possède 40656 états possibles : 10 régions pour le corps calleux, 10 pour le sinus frontal, 5 pour le sinus sphénoïdal, 7 pour la trachée artère et 6 pour le tronc cérébral. On relève 6924 états de type T5 (toutes les instances des objets de scène satisfont les contraintes du réseau sémantique anatomique) parmi lesquels 1808 seulement sont des états de satisfaction.

Les résultats de quatre stratégies différentes de recherche stochastique dans le graphe relationnel sont synthétisés dans le tableau 3: (i) **La première** stratégie consiste à implémenter un parcours en largeur exécuté à partir de l'état initial. L'énumération mécanique et systématique de toutes les possibilités détermine la borne inférieure des performances moyennes. Bien que la traversée s'arrête dès qu'un état de satisfaction est atteint, en pratique, de nombreux autres états intermédiaires auront été visités auparavant. (ii) **La seconde** stratégie consiste à explorer le voisinage d'un état selon le principe d'une marche aléatoire, le seul critère d'arrêt étant la rencontre d'un état de satisfaction. Cette stratégie permet de caractériser un niveau intermédiaire de performances, puisqu'elle ne privilégie aucun état particulier. Elle aboutit à un état de satisfaction par pur hasard. (iii) **La troisième** stratégie, consiste à se mouvoir uniquement vers des états de plus grande satisfaction. Cette stratégie de suivi de gradient pose la borne supérieure des performances. Lorsqu'un chemin complet de gradient existe, il permet d'arriver à destination avec le nombre minimum de mouvements. (iv) Enfin, **la dernière** stratégie évalue les performances d'un déplacement suivant les règles du modèle de l'éco-résolution.

Le protocole expérimental met en œuvre les différentes recherches stochastiques sur tous les états du graphe et répète cette opération une centaine de fois. Cette répétition tend à limiter l'influence des composantes aléatoires présentes dans les différentes stratégies. Quatre informations ont été retenues lors de ces parcours, il s'agit du nombre minimum, maximum et moyen de mouvements effectués, ainsi que le pourcentage d'erreur. Pour éviter toute circularité gênante, les stratégies sont limitées à 400 mouvements. Les résultats sont organisés suivant le type des états servant à l'initialisation. On note qu'il n'existe pas d'états de type T4 dans ce graphe.

L'espace des états n'étant pas convexe, une stratégie de suivi de gradient n'aboutit que dans 23% des situations. La réussite systématique du parcours en profondeur prouve, par ailleurs, qu'il existe toujours un état de satisfaction accessible depuis l'état initial considéré. Des tests complémentaires permettent de conclure à l'accessibilité des 1808 solutions du problème depuis les états de type T0 à T3. Le pourcentage d'erreur correspond aux états initiaux auxquels une recherche stochastique n'a pas permis d'associer un état de satisfaction. Dans le cas du suivi de gradient les états d'arrêt sont quelconques alors que dans le cas de l'éco-résolution, il

s'agit d'états de type T5 virtuels (au moins une des variables se voit attribuée la valeur -1). Seuls les états réels sont définis comme des états de satisfaction.

TAB. 3 : Résultats

100 essais	Parcours Largeur	Marche Aléatoire	Suivi Gradient	Eco Résolution
T0	Min :3	Min :1	Min :1	Min :1
	Max :197	Max :337	Max :4	Max :7
428.000	Moy :79,2	Moy :21,9	Moy :2,52	Moy :2,88
10,53 %	Err : 0 %	Err : 0 %	Err:75,7%	Err:$1,6E^{-3}$ %
T1	Min :2	Min :1	Min :1	Min :1
	Max :223	Max :313	Max :3	Max :7
1.150.800	Moy :67,0	Moy :21,7	Moy :1,78	Moy :2,45
28,3 %	Err : 0 %	Err : 0 %	Err:77,1%	Err :$6,1E^{-4}$ %
T2	Min :1	Min :1	Min :1	Min :1
	Max :223	Max :323	Max :2	Max :12
809.500	Moy :64,9	Moy :21,9	Moy :1,42	Moy :2,42
19,92 %	Err : 0 %	Err : 0 %	Err:83,5%	Err:$6,2E^{-4}$ %
T3	Min :1	Min :1	Min :1	Min :1
	Max :222	Max :299	Max :1	Max :11
984.900	Moy :36,5	Moy :20,4	Moy :1,00	Moy :1,95
24,22 %	Err : 0 %	Err : 0 %	Err:77,2%	Err : 0 %
T5	Min :0	Min :0	Min :0	Min :0
	Max :244	Max :302	Max :0	Max :7
692.400	Moy :36,3	Moy :15,7	Moy :0,00	Moy :1,30
17,03 %	Err : 0 %	Err : 0 %	Err:73,9%	Err: 0%
Total	Min :0	Min :0	Min :0	Min :0
	Max :244	Max :337	Max :4	Max :12
4.065.600	Moy :55,3	Moy :20,4	Moy :1,29	Moy :2,17
100 %	Err : 0 %	Err : 0 %	Err:77,7%	Err :$4,6E^{-4}$ %

Dans les cas moyens, le suivi de gradient permet d'atteindre les états de satisfaction en 1 ou 2 mouvements. Ces bonnes performances sont à relativiser car dans 77% des cas, la stratégie ne fonctionne pas. Pour y remédier, une stratégie mixte, couplant la marche aléatoire et le suivi de gradient, a été testée. Les performances de cette dernière tiennent plus de celle de la marche aléatoire que du suivi de gradient. La marche aléatoire nécessite en moyenne une vingtaine de mouvements pour atteindre un état de satisfaction. Plus du double est nécessaire pour que le parcours en largeur réussisse. L'éco-résolution révèle de très bonnes performances, très proches du suivi de gradient, puisque l'objectif est réalisé en 2 ou 3 mouvements (10 fois moins que la marche aléatoire).

Le nombre maximum de mouvements évalue les pires comportements des algorithmes étudiés. Lorsque le suivi de gradient fonctionne, ses performances sont meilleures que dans le pire cas théorique (connaissant les coordonnées de destination, il est possible de les atteindre en 5 mouvements au plus). L'énumération exhaustive du parcours en largeur prévient les cycles et permet de meilleures performances que la marche aléatoire. On observe une centaine de mouvements d'écart. Dans le cas de l'éco-résolution, le nombre maximum est systématiquement inférieur à 12 déplacements ce qui le rapproche des performances du suivi de gradient et le place loin devant le parcours en largeur (plus de 200 mouvements d'écart).

5. Conclusion & Perspectives

La stratégie d'éco-résolution montre de très bonnes performances à la fois en moyenne, mais aussi dans le pire cas. Elle suit de très près les performances du suivi de gradient tout en étendant considérablement le nombre d'états traités (diminution de son taux d'erreurs). Il est à noter que le nombre de tests de compatibilité relationnelle de formes entrepris entre chaque mouvement est nettement supérieur aux autres algorithmes. Le temps d'exécution s'en ressent nécessairement. Néanmoins, en moyenne, l'exécution de l'éco-résolution s'effectue en 3 ms. Dans l'absolu et vis-à-vis des besoins de notre application, ce temps est suffisamment court pour que le spécialiste puisse interagir immédiatement avec le système et tester d'autres configurations initiales lui permettant d'accéder à d'autres segmentations (cf .figure 4).

Les comportements actuels des éco-agents ne prennent pas en compte les problèmes de cycles qui étaient absents dans l'expérimentation présentée. Le marquage des places lors de la fuite des agents peut résoudre partiellement ce problème.

De plus, les d'erreurs sont des évènements de nature anormale et leur étude précise permettra de mieux comprendre comment les éviter. L'ensemble de la stratégie d'éco-résolution devra être ensuite conforté en l'expérimentant sur la totalité des IRM de notre base.

FIG. 4 : TC segmenté

Références

[1] J.- M. Jolion. *Sur la méthodologie de conception des systèmes de vision*. Chapitre du livre, *Les systèmes de vision*. Hermès (ICC), pp. 97-131, Paris, 2000.

[2] N. Richard, M. Dojat et C. Garbay. *Automated segmentation of human brain MR images using a multi-agent approach.* Artificial Intelligence in Medecine, vol. 30, n°2, pp. 153-176, 2004.

[3] E. Duchesnay, J.-J. Montois, Y. Jacquelet. *Cooperative Agents Society Organized as an Irregular Pyramid : a Mammography Segmentation Application.* Pattern Recognition Letters, vol. 24, n° 14, pp. 2435-2445, octobre 2003.

[4] J. Ferber. *Les Systèmes Multi Agents : vers une intelligence collective*. InterEditions (IIA), 1995.

[5] A. Drogoul, C. Dubreuil. *A Distributed Approach to N-Puzzle Solving*. Proceedings of the Distributed Artificial Intelligence Workshop, Seattle, 1993.

[6] L. Breton, N. Jussien. *Un CSP comme comportement d'agent. Application à la resolution d'équations en physique des milieux granulaires.* Journal Electronique d'Intelligence Artificielle, Vol. 3, 2004.

[7] D. Julien, Z. Guessoum, M. Ziane. *Building User Interfaces thanks to Eco-Resolution.* International Conference on Advance in Intelligent Systems – Theory and Applications, in cooperation with the IEEE Computer society, Luxembourg, November 2004.

Modèles GMM et algorithme de Brandt pour la correction de la segmentation de la parole par HMM

Safaa JARIFI[1], Dominique PASTOR[1], Olivier ROSEC[2]

[1]Département Signal et Communication
ENST Bretagne, Technopôle Brest-Iroise, 29238 Brest, France

[2]France Télécom, Division R&D TECH/SSTP/VMI
2, avenue Pierre Marzin, 22307 Lannion Cedex, France

safaa.jarifi@enst-bretagne.fr, dominique.pastor@enst-bretagne.fr
olivier.rosec@francetelecom.com

Résumé – On compare les performances de deux algorithmes de segmentation automatique. Le premier, nommé "HMM amélioré", affine la segmentation produite par les modèles de Markov cachés (HMM). Le deuxième est l'algorithme de Brandt qui vise, quant à lui, à détecter les ruptures de stationnarité. Le premier algorithme requiert la connaissance a priori de la phonétisation, le second non. Étant donné que l'algorithme de Brandt commet des insertions et des omissions, ce qui n'est pas le cas du HMM amélioré, on introduit une généralisation du taux de segmentation correcte (TSC) afin de comparer ces deux algorithmes. Les mesures expérimentales des TSCs permettent d'évaluer une limite supérieure des performances de l'algorithme de Brandt et suggèrent de combiner ces deux méthodes avec d'autres algorithmes adaptés à la séparation des classes acoustico-phonétiques.

Abstract – We compare the performance of two automatic segmentation algorithms. The first one is the so-called "refined HMM" and aims at refining the segmentation performed by Hidden Markov Models (HMM). The second is the Brandt's GLR (Generalized Likelihood Ratio) algorithm. It detects speech signal discontinuities. The first method assumes the knowledge of the phonetic sequence whereas in the Brandt'GLR method, this constraint is not used. The refined HMM yields no insertion and no omission in contrast with the Brandt's GLR method. Hence, we generalize the notion of correct segmentation rate (CSR) so as to compare these algorithms. The experimental results given in this paper exhibit an upper limit for Brandt's GLR method CSRs and suggest in combining the two methods with other algorithms adapted to the detection of boundaries between known acoustic classes.

1 Introduction

On s'intéresse à la segmentation automatique de grands corpus de parole spontanée et continue dédiés à la synthèse vocale. Dans l'approche classique, cette segmentation est automatisée par l'utilisation de modèles de Markov cachés (HMM) : après apprentissage de ces modèles sur l'ensemble du corpus de parole, la segmentation consiste à apposer les frontières de phones à l'aide d'un alignement forcé à partir d'un étiquetage phonétique supposé exact. Cette technique offre des résultats acceptables mais la précision de la segmentation peut parfois être insuffisante, entraînant des imperfections en sortie du système de synthèse. Des étapes de vérification, fastidieuses et coûteuses, demeurent indispensables. Dans le but de les réduire, nous cherchons à mettre en oeuvre des algorithmes de segmentation automatique dont la qualité approche celle de la segmentation manuelle. Dans cette perspective, ce papier a pour objectif de comparer deux algorithmes de segmentation automatique : le HMM amélioré et l'algorithme de Brandt.

Le premier, introduit dans [7] pour la segmentation d'un corpus chinois, consiste à affiner la segmentation produite par HMM en utilisant des modèles GMM (Gaussian Mixture Model) au voisinage des frontières phonétiques. Cet algorithme, parce qu'il est basé sur le HMM avec alignement forcé, produit un nombre de marques de segmentation égal au nombre de phonèmes présents dans la séquence phonétique.

Le second algorithme, originellement décrit dans [1, 2] et analysé dans [5], détecte les discontinuités locales de la stationnarité dans un signal de parole sans connaissance a priori sur la séquence phonétique de la phrase. Par conséquent, on n'échappe pas à la présence d'insertions et d'omissions dans la segmentation.

Pour ces raisons, on propose, après une description de chaque algorithme, une généralisation du taux de segmentation correcte (TSC). Cette généralisation a pour but d'évaluer les performances en termes de précision d'un algorithme de segmentation en prenant en compte ses insertions et ses omissions. La section 5 décrit l'ajustement des paramètres du HMM amélioré pour un corpus français et les TSCs des deux algorithmes en comparaison avec le HMM classique. Enfin, sur la base de ces résultats, on propose une approche qui vise à approcher la segmentation obtenue par HMM amélioré de la segmentation manuelle.

2 Le HMM amélioré

Le principe de cet algorithme est d'adjoindre à l'algorithme de segmentation par HMM un post-traitement utilisant des modèles GMM au voisinage des frontières phonétiques. Cette méthode s'effectue en deux étapes :

1. On associe un modèle GMM à chaque marque de segmentation d'un petit corpus d'apprentissage segmenté manuellement. En effet, pour accomplir cette association,

on crée d'abord un super-vecteur pour chaque frontière segmentée manuellement. Ce vecteur est obtenu en calculant les vecteurs acoustiques sur les trames autour de la frontière. La figure 1 représente un super-vecteur sur $(2N+1)$ trames. Chaque frontière B dépend du phonème X à sa droite et du phonème Y à sa gauche. On obtient ainsi un pseudo-triphone noté $X-B+Y$ (voir figure 2). Puisque la taille du corpus d'apprentissage est petite, il est alors difficile de bien apprendre un modèle GMM pour chaque frontière. Pour cette raison, une étape intermédiaire avant l'apprentissage consiste à rassembler les pseudo-triphones par classes grâce à la création d'un arbre de régression et de classification (CART). Ensuite, un GMM est appris pour chaque classe.

2. Pour une phrase de test donnée, on cherche une marque de segmentation autour de chaque frontière obtenue par HMM; la marque retenue est alors celle qui maximise la vraisemblance par rapport au modèle GMM associé à cette frontière.

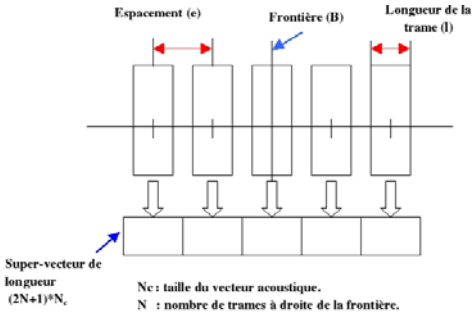


FIG. 1: *Exemple de super-vecteur pour une frontière donnée*

3 Algorithme de Brandt

L'algorithme de Brandt [5] est basé sur un critère statistique afin de décider s'il y a ou non une rupture de stationnarité dans le signal de parole. Il considère que le signal est une suite d'unités stationnaires puis modélise chaque unité par un modèle autorégressif (AR).

Soit alors $w = (y_n)$ une de ces unités. Chaque échantillon obéit donc à la loi de la forme : $y_n = \sum_{i=1}^{p} a_i y_{n-i} + e_n$ où p est l'ordre du modèle, supposé constant pour toutes les unités. Le $n^{\text{ième}}$ échantillon de la fenêtre est y_n et e_n est un bruit gaussian de moyenne nulle et de variance égale à σ^2. Par conséquent, chaque unité est associée à un vecteur de paramètres $\Theta = (a_1 = \ldots, a_p, \sigma)$.

Soit w_0 une fenêtre de longueur n. Le principe de l'algorithme de Brandt est de décider si w_0 doit être découpée en deux fenêtres w_1 et w_2 ou non. Cette décision se fait sur les vecteurs de paramètres Θ_1 et Θ_2 associés respectivement aux fenêtres w_1 et w_2. Ainsi, un changement entre Θ_1 et Θ_2 est détecté quand le rapport de vraisemblance généralisé (GLR) dépasse un seuil λ prédéfini. L'instant de ce changement est considéré comme l'instant de rupture de stationnarité.

4 Critères d'évaluation de performances

Les critères de comparaison des algorithmes de segmentation sont la probabilité d'insertion, la probabilité d'omission et le TSC. En particulier, le taux de segmentation correcte (TSC) mesure la précision d'une segmentation à une tolérance ϵ près par rapport à la segmentation de référence et ce en prenant en compte le nombre d'insertions et d'omissions.

TAB. 1: *TSCs de la segmentation produite par HMM*

ϵ	5	10	20	30
TSC	33.54	59.77	85.24	92.83

Soient $U = \{U_1, U_2, \ldots, U_n\}$ les instants des marques produites par un algorithme de segmentation automatique, et $V = \{V_1, V_2, \ldots, V_p\}$ les instants des marques de la référence. Nous construisons la liste $V_U = (V_{k_1}, \ldots, V_{k_n})$ tels que k_j est l'indice dans $\{1, \ldots, p\}$ et V_{k_j} la marque la plus proche de U_j. Donc chaque élément de U est relié à un élément de V. Les omissions sont les instants V_ℓ qui n'appartiennent pas à la liste V_U, où $\ell \in \{1, \ldots, p\}$. Pour les insertions, on regarde les éléments répétés dans V_U. Si V_U contient m fois la même marque V_ℓ par exemple, le nombre d'insertions autour de V_ℓ est $m-1$. Supposons que les m marques de U correspondantes aux m marques V_ℓ dans V_U sont $(U_j, \ldots, U_{j+m-1})$. La marque la plus proche, de ce dernier vecteur, de V_ℓ est considérée comme la marque non insérée. Les autres $m-1$ marques représentent les insertions autour de V_ℓ.

On applique cette technique pour tout ℓ tel que V_ℓ est répété dans V_U afin de localiser toutes les insertions. Ainsi, une segmentation dont le nombre de segments est le même que pour la référence, peut contenir des omissions et des insertions. De plus, nous définissons les probabilités d'insertion et d'omission comme suit : $P_i = \frac{n_i}{p+n_i}$ et $P_o = \frac{n_o}{n+n_o}$ où n_i est le nombre total d'insertions et n_o est le nombre total d'omissions. Pour calculer le TSC, on identifie les marques qui ne sont pas des insertions au sens donné précédemment. Ensuite, on compte le nombre de ces marques qui se situent à moins de ϵ ms de la marque manuelle correspondante. Enfin, on divise ce nombre par la somme du nombre de marques de la segmentation manuelle et du nombre d'insertions. Autrement dit :

$$TSC = \frac{100}{p+n_i} \sum_{j=1}^{n} I_{[0,\epsilon]}(|V_{k_j} - U_j|)$$

où $I_{[0,\epsilon]}(x)$ est l'indicatrice de l'intervalle $[0, \epsilon]$. Un algorithme est donc considéré comme d'autant plus précis que son TSC est proche de 1.

5 Résultats

On utilise un corpus français de 7350 phrases prononcées par un seul locuteur masculin. Les signaux de parole résultants sont échantillonnées à 16 kHz et la phonétisation sur un corpus a été vérifiée manuellement.

5.1 HMM amélioré

L'objectif de cette section est d'ajuster les paramètres du HMM amélioré sur un corpus français. Nous partons des valeurs de paramètres ajustés sur un corpus chinois dans [7] et analysons si ces valeurs restent encore valables sur le corpus étudié. Seuls les paramètres d'apprentissage sont ajustés. La zone de recherche de la marque de segmentation autour de celle du HMM est fixée à 60 ms avec un pas de segmentation de 5 ms. Les TSCs dans cette section sont calculés pour

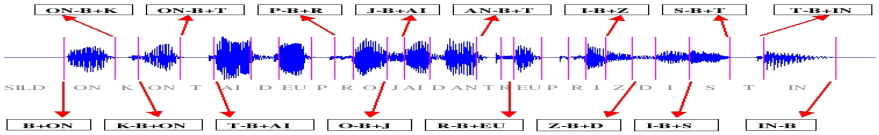


FIG. 2: *Exemples de pseudo-triphones*

TAB. 2: *TSC vs* (T, MTI)

T	ϵ	$MTI = 10$	$MTI = 20$	$MTI = 40$	$MTI = 100$
20	5	38.49	38.08	37.48	35.13
	10	64.55	64.29	62.94	60.13
	20	87.91	87.74	87.05	85.16
	30	94.46	94.44	93.93	93.45
100	5	38.55	38.12	37.48	35.13
	10	64.63	64.28	62.94	60.13
	20	87.98	87.73	87.04	85.16
	30	94.50	94.46	93.93	93.45
350	5	38.38	37.90	37.49	35.13
	10	64.32	64.01	62.94	60.13
	20	87.93	87.51	87.11	85.16
	30	94.35	94.42	93.93	93.45

TAB. 3: *TSC vs taille du corpus d'apprentissage*

Taille	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 20$	$\epsilon = 30$
200	38.27	63.97	87.63	94.31
300	38.55	64.63	87.98	94.49
600	41.26	67.10	88.78	95.03
800	41.42	67.76	88.87	95.15

TAB. 4: *TSC pour différentes valeurs du* (N, e)

e	*N*	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 20$	$\epsilon = 30$
0	0	33.95	58.70	83.70	92.27
10	2	39.01	64.42	86.13	93.41
30	2	38.55	64.63	87.98	94.49
30	3	37.16	63.83	88.06	94.58

des tolérances $\epsilon \in \{5, 10, 20, 30\}$ ms, puis ils sont comparés à ceux du HMM classique (voir tableau 1).

Les paramètres d'apprentissage sont : la taille du corpus d'apprentissage, le nombre de trames sur lesquelles on calcule le super-vecteur $(2N + 1)$, l'espacement entre les trames e, la taille de la trame (20 ms), le nombre de gaussiennes pour le GMM (ici égal à 1), le nombre de coefficients par vecteur acoustique (égal à 39), et les critères d'arrêt dans le CART. Ces derniers sont : le nombre minimum d'éléments par nœud final (MTI) et le seuil du log de vraisemblance (T) pour associer une question à un nœud du CART.

Dans le tableau 2, le couple (T, MTI) qui maximise le TSC est $(100, 10)$. Ce résultat est obtenu en fixant (N, e) à $(2, 30)$ et la taille du corpus d'apprentissage à 300.

Le tableau 3 montre l'influence de la taille du corpus sur les TSCs. Le couple (N, e) est encore fixé à $(2, 30)$ et (T, MTI) est égal à $(100, 10)$ d'après ce qui précède. Ce tableau montre que 300 phrases sont suffisantes pour avoir un bon TSC.

Enfin, grâce au tableau 4 nous vérifions que la valeur égale à $(2, 30)$ du couple (N, e) reste valable aussi pour le corpus français étudié. Ce choix semble être un bon compromis entre deux contraintes : la longueur du super-vecteur doit être suffisamment longue pour prendre en compte le plus d'information possible sur la transition entre les deux phonèmes ; cette longueur ne doit pas être trop longue afin de ne pas incorporer des informations qui ne sont pas liées directement à la frontière elle-même.

5.2 Algorithme de Brandt

Pour l'algorithme de Brandt, on choisit un ordre de modèle égal à 16 et un seuil égal à 30 comme dans [5]. En calculant les moyennes des probabilités d'insertion et d'omission de 4 tests utilisant 1200 phrases choisies aléatoirement, on trouve $P_i = 0.6$ et $P_o = 0.1$. la valeur élevée de P_i est due aux oscillations du GLR. En fusionnant cet algorithme avec un détecteur d'activité vocale (DAV) [3], la probabilité d'insertion diminue significativement (de 0.6 à 0.3). En effet, la plupart des insertions sont localisées dans les silences (voir figure 3).

5.3 Comparaison

Les résultats que nous obtenons en comparant ces algorithmes à l'aide des critères ci-dessus sont synthétisés par la figure 4. Cette figure est obtenue en moyennant 4 tests croisés sur un corpus de 1200 phrases. Chaque phrase contient en moyenne une vingtaine de phonèmes. Les 5 courbes représentées sur cette figure sont les suivantes :

- **La courbe "Brandt" :** les TSCs obtenus par l'algorithme de Brandt sont très faibles en raison de la valeur élevée de P_i.
- **La courbe "Brandt+DAV" :** elle est obtenue en calculant les TSCs de la segmentation de Brandt après avoir éliminé les insertions dans les silences grâce au DAV suggéré dans [3].
- **La courbe "HMM" :** elle est obtenue par HMM avec alignement forcé et considérée comme la courbe de référence dans notre comparaison.
- **La courbe "HMM amélioré" :** elle résulte de l'utilisation du HMM amélioré ; les TSCs sont meilleurs que ceux du HMM standard.
- **La courbe "Brandt idéal" :** elle correspond aux TSCs calculés à partir de l'algorithme de Brandt après avoir éliminé les insertions grâce au traitement décrit dans 4 et remplacé les omissions par les marques correspondantes de la segmentation du HMM amélioré ; les TSCs sont meilleurs que ceux du HMM amélioré, ce qui signifie que les marques de segmentation de Brandt qui ne sont pas des insertions sont plus proches des marques manuelles que celles des autres algorithmes étudiés ici. En éliminant les insertions de l'algorithme de Brandt, il est théoriquement possible d'atteindre les TSCs de "Brandt idéal". Cependant, nous pensons qu'il est a priori très

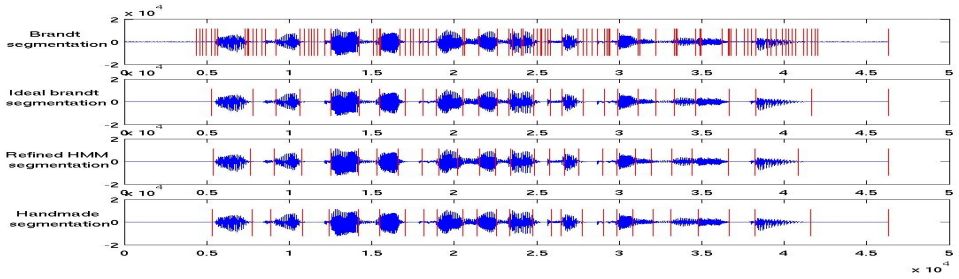


FIG. 3: *Segmentations manuelle, semi-automatique avec Brandt idéal, et automatiques avec l'algorithme de Brandt et HMM amélioré.*

difficile d'enlever les insertions dans les parties "parole" car ces insertions ne sont pas suffisamment bien localisées autour des marques de segmentation manuelle (figure 3).

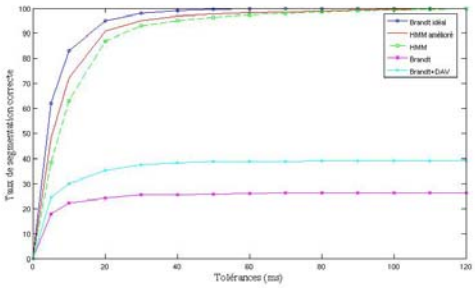


FIG. 4: *TSCs pour différents algorithmes*

6 Approche dédiée à la synthèse vocale

D'après ce qui précède, on peut songer compléter l'algorithme de Brandt par des algorithmes permettant de réduire le nombre d'insertions et d'omissions et espérer ainsi approcher les performances de "Brandt idéal". La détection de voisement, l'analyse multi-résolution, la détection parole/non parole sont autant d'approches susceptibles de détecter des insertions et des omissions de Brandt. La difficulté à la laquelle on se heurte alors est que le contrôle des insertions et des omissions de la segmentation de Brandt est reporté sur les algorithmes complémentaires qui, eux aussi, commettent des erreurs de segmentation.

Il est donc plus raisonnable de partir d'une segmentation sans insertion ni omission (comme le HMM amélioré, voire le HMM classique) et d'essayer d'affiner cette segmentation à l'aide d'algorithmes dédiés. La méthode que nous proposons d'étudier est alors la suivante. On part de la segmentation obtenue par HMM amélioré. Pour chaque marque de cette segmentation, on calcule le GLR. Si ce GLR dépasse un certain seuil, alors on conserve la marque de segmentation du HMM amélioré. Sinon, on recourt à des algorithmes de segmentation acoustique adaptés aux classes acoustico-phonétiques des segments à identifier. Ces classes sont connues grâce à la phonétisation dont on dispose. De nombreux exemples d'algorithmes de segmentation acoustique sont décrits dans la littérature [6]. Parmi ces exemples, on s'intéressera particulièrement à la méthode dite de "Mallat-Modulation" décrite dans [4]. Cette méthode est une extension de l'algorithme de Mallat, adaptée à la détection d'énergie autour de fréquences arbitraires.

7 Conclusion

Après avoir décrit et comparé expérimentalement deux algorithmes de segmentation automatique, nous avons déduit une méthode de segmentation qui conjugue l'absence d'insertion et d'omission du HMM amélioré avec la précision des marques de la segmentation de Brandt, lorsque celles-ci ne sont pas des insertions. Notre travail futur consistera alors à mesurer les performances de cette méthode par rapport au HMM amélioré.

8 Remerciements

Nous remercions les reviewers pour leurs suggestions constructives que nous allons en prendre compte dans nos travaux.

Références

[1] A.V. Brandt, *Detecting and estimating parameters jumps using ladder algorithms and likelihood ratio test*, Proc.ICASSP, pp.1017-1020, Boston, MA, 1983.

[2] A.V. Brandt, *Modellierung von Signalen mit Sprunghaft Veränderlichem Leistungsspektrum durch Adaptive Segmentierung*, Doctor-Engineer Dissertation, 1984, München, RFA (in German).

[3] S. Jarifi, D. Pastor, O. Rosec, *Jump and silence/speech detection for automatic continuous speech segmentation*, International Symposium on Image/Video Communications, Brest, France, 2004.

[4] C. Lemoine *Recherche de traits acoustiques de la parole bruité par analyse multi-résolution*, Thèse, Université de Bordeaux I, 1998.

[5] R.A. Obrecht, *A New Statistical Approach for the Automatic Segmentation of Continuous Speech Signals* , IEEE Transactions on Acoustics, Speech and Signal Processing, Vol.36(1), pp.29-40, 1988.

[6] D.T. Toledano and L.A. Hernández Gómez and L. Villarubia Grande, *Automatic Phonetic Segmentation*, IEEE Transactions on Speech and Audio Processing, Vol.11(6), pp.617-625, 2003.

[7] L. Wang and Y. Zhao and M. Chu and J. Zhou and Z. Cao, *Refining Segmental Boundaries for TTS Database Using Fine Contextual-Dependent Boudary Models*, Proc. ICASSP, pp.641-644, Montreal, Canada, 2004.

Segmentation en Caractères Individuels dans des Images de Scènes Naturelles

Céline MANCAS-THILLOU, Matei MANCAS, Bernard GOSSELIN

Faculté Polytechnique de Mons, Laboratoire de Théorie des Circuits et Traitement du Signal
Bâtiment Multitel - Initialis, 1, avenue Copernic, 7000, Mons, Belgique
celine.thillou@tcts.fpms.ac.be, matei.mancas@tcts.fpms.ac.be
bernard.gosselin@tcts.fpms.ac.be

Résumé – Grâce à l'essor des appareils miniatures surmontés de caméras basse résolution, parfois même un peu gadgets, de nouveaux challenges sont apparus pour analyser les images de scènes naturelles. Le texte présent dans ces images est difficilement reconnu par les logiciels OCR actuels en raison des nombreuses dégradations. Nous détaillons dans cet article une méthode de segmentation de caractères collés présents dans des images photographiques. Le principal but est de segmenter ces caractères en composants individuels pour faciliter et améliorer leur reconnaissance. Les premiers résultats de notre méthode éprouvée sur une base de données publique sont également présentés.

Abstract – Embedded applications are growing everyday with the multiplicity of tiny devices and cameras. Hence, new challenges appear in image processing with natural scene images. Characters present in this kind of images make off-the-shelf OCR softwares fail and we describe in this paper a method to handle touching characters in camera-based images. The main goal is to segment them into individual components to improve recognition rates, required in numerous applications. Using a public database, our first results are then detailed.

1 Introduction

La segmentation en unités distinctes et la reconnaissance de caractères ont été éprouvées depuis plus de deux décennies. Néanmoins, les systèmes commerciaux sont encore trop sensibles aux nombreuses dégradations dans les images. Les images de scènes naturelles, acquises ici par un appareil photo basse résolution comportent de nombreuses dégradations telles que le flou, la mauvaise illumination et la distorsion, absentes dans une acquisition par scanner, et une taille et police de caractères parfois extravagantes. Cet article traite principalement de la segmentation en unités distinctes des caractères collés. Dans les images de scènes naturelles, une majorité de ces caractères est présente si aucun traitement n'est effectué et la solution ne peut parfois pas être trouvée à l'étape précédente d'extraction de texte ou binarisation. Leur suppression est néanmoins indispensable pour obtenir des résultats de reconnaissance satisfaisants dans un environnement embarqué comme l'est l'acquisition des images de scènes naturelles et pour leurs nombreux usages dans les applications (aide aux personnes malvoyantes et aveugles, traducteur de poche...). En effet, les logiciels de reconnaissance actuellement disponibles sont lourds et très sensibles à un grand nombre de caractères collés.

Les algorithmes trouvés dans la littérature basent leurs tests sur du texte présent dans des flux vidéo principalement [9] en utilisant l'information présente dans les multiples trames disponibles. Cette information est manquante dans le cas d'images fixes, comme celles présentes dans notre base de données (issue de la compétition publique ICDAR 2003) illustrée avec quelques exemples dans Figure 1. La super-segmentation, exercée sur une analyse de documents acquis par scanner et qui a pour but de segmenter les caractères en primitives, a été testée

FIG. 1 – Exemples de la base de données publique Icdar 2003 [4]

pour donner des résultats satisfaisants lorsque la police des caractères n'est pas trop imaginative car alors le reconnaisseur de caractères peine à fusionner les différentes primitives. Des modèles [1] basés sur une étude préliminaire de la dégradation et des caractères présents ont donné des résultats encourageants mais la définition et application de ces modèles sont lourds en temps de calcul pour un usage embarqué et la diversité des polices accroît le nombre de modèles de manière exponentielle. Un état de l'art [2] basé sur les méthodes de segmentation de caractères acquis au moyen d'un scanner montre que ces techniques ne peuvent s'appliquer ici en raison des nouvelles dégradations qui sont absentes avec une numérisation par scanner. De plus, elles présentent trop de paramètres à définir, perdant la généricité, exigée par les images de scènes naturelles, qui, par

définition, sont aussi variées que possible. D'autres méthodes basées sur des modèles ou des marqueurs ont été utilisés [6] mais le plus délicat est d'approcher suffisamment le résultat pour initialiser les contours actifs ou autres marqueurs. En effet, si les marqueurs ou les modèles sont initialisés en utilisant une approximation trop éloignée de la réalité, le résultat ne sera pas satisfaisant.

Après cette brève description du travail déjà effectué, la première constatation est qu'il n'existe aucun algorithme de segmentation des caractères adapté aux images de scènes naturelles. Ainsi il convient de détailler notre méthode de segmentation des caractères collés et de présenter les premiers résultats obtenus. Enfin, nous discuterons les avantages de notre algorithme tout en présentant nos futurs travaux à ce sujet.

2 Méthode

2.1 Une approche haut niveau

Afin d'approcher le texte, nous utilisons une technique basée sur un débruitage effectué grâce à une décomposition en ondelettes suivie d'une classification non supervisée. Cette méthode d'extraction de texte (MET) a déjà été exposée lors d'articles précédents [8] et a prouvé sa généricité. Cette première approche permet d'avoir une boîte englobante précise des zones de texte et de voir le nombre de caractères collés diminuer pour n'être présent que dans 29.2% des images de notre base de données [4] grâce à un souci de segmentation future présent dans notre méthodologie.

Les principales étapes de MET sont :

- Débruitage avec les ondelettes de Daubechies en 16 niveaux. Conservation des hautes fréquences et de la plus basse fréquence.
- Discrimination binaire de la classe du fonds de l'image (simple, complexe) pour application de l'algorithme de seuillage global et rapide d'Otsu [7] dans les cas simples.
- Réduction du nombre des couleurs sans dégradation.
- Clustering K-moyennes non-supervisé des couleurs avec K=3.
- Combinaison éventuelle entre les clusters basée sur la distance entre les centroïdes. La combinaison s'effectue avec respect de la segmentation des caractères. Seuls les pixels ne connectant pas les caractères pré-existants sont ajoutés.

2.2 Complétion par une approche bas niveau

La deuxième étape qui est décrite dans cet article consiste dans l'analyse de chaque objet résultant de la méthode précédente. Ces objets sont soit des caractères uniques dans les images les moins dégradés, soit des caractères cassés dans de rares cas, soit enfin des caractères collés ce qui est l'erreur la plus courante. Dans cet article, nous traitons le problème des objets contenant des caractères collés tout en essayant de ne pas casser les objets contenant déjà un seul caractère. L'objectif prépondérant pour la suite des étapes est de réduire le taux de fausses alarmes pour ne pas alourdir les traitements à effectuer et dégrader l'image.

Nous avons besoin de localiser les informations à la fois spatialement afin de savoir où dans l'image se situe une éventuelle coupure entre les caractères et bien sûr fréquentiellement pour obtenir des informations sur les variations des niveaux de gris susceptibles de révéler les séparations entre caractères. Or la meilleure localisation simultanée dans l'espace image et fréquence est effectuée par un filtre de Gabor, qui applique un filtre de type cosinus possédant une certaine direction modulée par une fenêtre gaussienne.

Des études plus approfondies sur les images naturelles [3] ont montré qu'elles possédaient un spectre décroissant en $1/f$. Cela a permis d'introduire une échelle fréquentielle logarithmique au lieu de l'échelle linéaire afin de rendre compte de ce phénomène. La composante continue qui apparaît dans les filtres symétriques de Gabor lorsque la bande passante de ceux-ci est trop importante disparaît grâce à cette échelle ce qui donne plus de marge de manoeuvre pour la bande passante des filtres. Le filtre Log-Gabor présente aussi une similitude plus importante avec le système visuel humain qui montre une réponse symétrique sur une échelle de fréquences logarithmiques. Nous utilisons ici l'implémentation faite par Kovesi [5] de la convolution par un filtre Log-Gabor, ce qui nous permet de récupérer la phase φ de la convolution de l'image I. Le paramètre le plus important d'un filtre de Gabor est sa fréquence f_{Gabor}. Nous avons décidé de mettre ce paramètre directement en relation avec l'épaisseur moyenne des caractères. Pour la calculer, nous utilisons simplement le rapport entre le nombre de pixels constituant le squelette des objets de l'image I issus de la méthode MET et la somme de tous les points qui les composent (voir Figure 2-c) tels que :

$$f_{Gabor} = \frac{\sum pixels_{squelette}}{\sum pixels_I} \quad (1)$$

Le calcul des convolutions de deux filtres, l'un horizontal et l'autre vertical, avec l'image, est effectué (voir Figures 2-d et 2-e). La séparation des caractères en unités distinctes est majoritairement verticale ainsi, seul le résultat de la convolution de l'image par le filtre vertical sera pris en compte. Par la suite nous ne considérons plus que la valeur absolue de la phase de ce résultat notée I_φ (voir Figure 2-f).

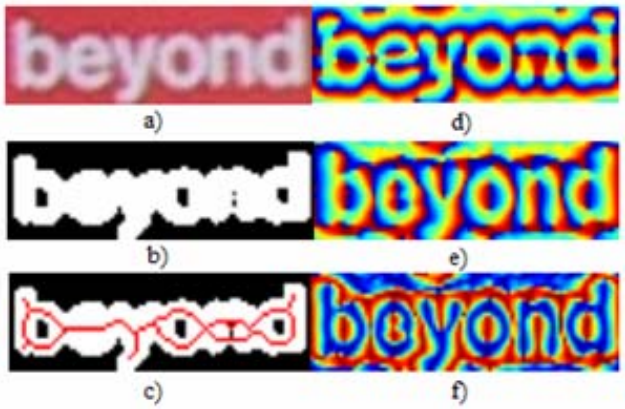


FIG. 2 – a) I, b) masque M, c) M et son squelette, d) et e) phases des filtres horizontal et vertical, f) valeur absolue de la phase e) I_φ

Enfin, nous multiplions le masque M par le résultat obtenu précédemment pour obtenir une image I_{masque} (voir Figure 3-a). Nous effectuons ensuite un seuillage S basé sur la méthode classique d'Otsu [7] qui va maximiser la variance inter-classes afin de trouver automatiquement ce seuil. Enfin, nous enlevons cette information inter-caractères éventuelle au masque M pour obtenir l'image finale I_{finale} par $I_{finale} = M - (I_{masque} > S)$ comme sur Figure 3-b.

FIG. 3 – a) I_{masque}, b) I_{finale}

3 Résultats

Sur la base de données d'exemples de mots segmentés [4], et après la méthode MET où 29.2% des images contenaient des caractères collés, l'algorithme de segmentation présenté dans cet article permet de diminuer le nombre de ces images de 76%, pour obtenir alors 7% d'images avec des caractères collés. De plus, seulement 2.3% de fausses alarmes sont constatées permettant de ne pas trop alourdir les traitements sur les caractères cassés dont le nombre n'augmente que légèrement. Un second résultat est illustré dans Figure 4 avec une comparaison avec l'algorithme de seuillage global Otsu [7] qui a également prouvé une grande généricité.

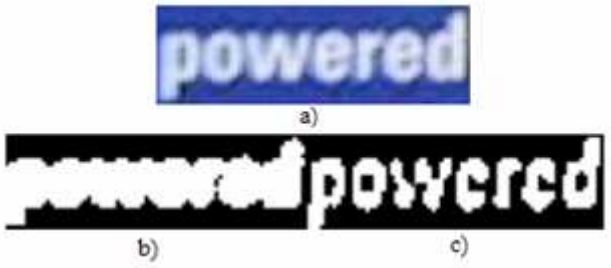


FIG. 4 – a) I, b) M obtenu par binarisation Otsu, c) I_{finale} par la méthode proposée

Sur Figure 5, un autre exemple est montré pour apprécier l'amélioration obtenue par rapport à la première approche MET. Ainsi cette approche en deux étapes permet d'avoir une bonne approximation de la segmentation des caractères.

Les résultats dépendent surtout de la bonne estimation de l'épaisseur des caractères. Celle-ci est, comme le montrent les premiers chiffres, assez satisfaisante dans 97.7% des cas. Elle est principalement erronée dans les images très bruitées où le masque M est imprécis.

FIG. 5 – a) I, b) M obtenu par la méthode MET, c) I_{finale} par la méthode proposée

4 Discussion et Conclusion

Un algorithme de segmentation de caractères collés adapté aux images de scènes naturelles a été présenté ainsi que les premiers résultats assez prometteurs. La méthode basée sur les filtres de Log-Gabor n'est pas gourmande en ressources de calcul puisque seulement deux convolutions (filtres horizontal et vertical) sont effectuées, ce qui permet ainsi un usage dans un environnement embarqué de style PDA.

Dans l'avenir, nous chercherons à améliorer la méthode d'estimation de l'épaisseur des caractères pour réduire encore davantage le taux de fausses alarmes. Dans cet article, un accent particulier a été donné aux caractères collés ainsi nos travaux actuels s'orientent sur la continuité des propriétés décrites dans le cas des caractères cassés même si leur nombre est très inférieur aux caractères collés dans les images de scènes naturelles.

Afin d'améliorer notre algorithme global, nous pouvons envisager deux évolutions principales. La première consisterait dans une étude bas niveau des paires d'éléments connectés issus de la première étape MET au lieu d'analyser les éléments connectés indépendamment les uns des autres. Cette approche permettrait d'éliminer les erreurs dues à une mauvaise segmentation présente dès la création du masque M. Enfin, un deuxième axe permettrait de compléter notre chaîne de segmentation avec un raffinement du résultat final en utilisant le dernier masque pour initialiser des méthodes de contours actifs ou de lignes de partage des eaux basées sur des marqueurs. Ainsi nous obtiendrions une bonne segmentation des caractères, étape fondamentale préalable à l'extraction de caractéristiques pertinentes pour une reconnaissance de caractères. Ces évolutions doivent être prises en compte avec à l'esprit le compromis qualité-temps de calcul. Les résultats seront améliorés mais avec un temps de calcul plus prohibitif. Ainsi l'étude du compromis en regard de l'application pour un usage embarqué est indispensable.

5 Remerciements

Les auteurs remercient le minitère de la Région wallonne en Belgique pour les subventions attribuées pour réaliser ce travail, qui est partie intégrante du projet Sypole.

Références

[1] M. Cannon, J. Hochberg et P. Kelly, *Quality assessment and restoration of typewritten document images*, Int. Jour. Document Analysis and Recognition, vol.2, pp. 80–89, 1999.

[2] R.G. Casey et E. Lecolinet, *A survey of methods and strategies in character segmentation*, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol.18, n.7, pp. 690–706, 1996.

[3] D.J. Field, *Relations between the statistics of natural images and the response properties of cortical cells*, Jour. of the Optical Society of America, vol.4, pp. 2379–2394, 1987.

[4] Robust Reading Competition Database : Consulté le 15 juin 2005 du site web :
http ://algoval.essex.ac.uk/icdar/RobustWord.html.

[5] P. Kovesi, *Image Features From Phase Congruency*, Videre : A Journal of Computer Vision Research, MIT Press, vol.1, n.3, 1999.

[6] H. Nishida, *Restoring high-resolution text images to improve legibility and OCR accuracy*, Proc. of the Electronic Imaging Conference of the International Society for Optical Imaging, pp. 136–147, 2005.

[7] N. Otsu, *A thresholding selection method from gray-level histogram*, IEEE Trans. on Systems, Man, and Cybernetics, vol.9, pp. 62–66, 1979.

[8] C. Thillou et B. Gosselin, *Color Binarization for Complex Camera-based Images*, Proc. of the Electronic Imaging Conference of the International Society for Optical Imaging, pp. 301–308, 2005.

[9] C. Wolf, J-M. Jolion et F. Chassaing, *Text localization, enhancement and binarization in multimedia documents*, Proc. of Int. Conf. on Pattern Recognition, vol.2, pp. 1037–1040, 2002.

Segmentation d'images multispectrales par arbre de Markov caché flou

Pierre LANCHANTIN[1], Fabien SALZENSTEIN[2]

[1]GET/INT, Département CITI,UMR-CNRS 5157,
9, rue Charles Fourier, 91011 Evry Cedex, France

[2]Laboratoire Iness, UPR CNRS 292
23, rue du Loess - BP 20 CR - F-67037 Strasbourg Cedex 2
Pierre.lanchantin@int-evry.fr, Fabien.Salzenstein@iness.c-strasbourg.fr

Résumé – Nous définissons un nouvel outil de segmentation statistique non supervisée, basé sur un modèle d'arbre de Markov caché flou. Notre modèle flou combine l'incertitude probabiliste des données observées avec les classes thématiques discrètes et continues qui représentent l'imprécision des données cachées. La technique de segmentation bayésienne mise en oeuvre correspond au critère MPM (Mode of Posterior Marginals). Notre approche permet d'une part le traitement d'objets contenant des structures diffuses comme c'est le cas en imagerie astronomique et d'autre part la prise en compte de données multi-bandes observées à différents niveaux de résolution et issues de capteurs corrélés. Nous validons notre modèle sur des images de synthèse et des images réelles multispectrales.

Abstract – A new hidden fuzzy Markov tree model is developed in an unsupervised way. Our fuzzy scheme combines the uncertainty of probabilities which models the observed data with discrete and continuous thematic classes which models the imprecision of the hidden data. Segmentation task is processed with Bayesian tools, such as the MPM (Mode of Posterior Marginals) criterion. Indeed, such fuzzy-based procedures seem to be a good answer for astronomical observations when patterns own diffuse structures and multiscale observations under dependence assumption. To validate our model, we perform the segmentation on synthetic images and raw multispectral data.

1 Introduction

Nous nous intéressons à de nouvelles approches d'analyse d'images basées sur la segmentation statistique et permettant l'identification de grands ensembles de données adaptées aux caractéristiques de l'imagerie astronomique. Ces objets sont généralement multirésolutions et multispectraux. Pour rendre compte de la complexité de l'information réelle et notamment de structures diffuses (nuages de Cirrus par exemple), les méthodes de segmentation fbues représentent une alternative aux techniques classiques. Les pixels y sont indexés au moyen d'ensembles discrets (classes thématiques homogènes) et/ou continus (régions intermédiaires). Ce type de classification, basé sur la théorie des ensembles fbus [1], fait correspondre à chaque pixel son degré d'appartenance aux régions thématiques. L'utilisation du "fbu" permet de modéliser l'imprécision des données cachées ou "vérité-terrain", tandis que les "probabilités" rendent compte de l'incertitude sur la classification des pixels [2].

Nous proposons d'enrichir ce modèle au moyen des arbres de Markov [3]. La segmentation fbue a fait appel à des techniques basées sur les champs [4, 5, 6], puis récemment sur les chaînes de Markov [7, 8]. Or, l'approche par arbres satisfait au caractère intrinséquement multirésolution des images astronomiques, et permet une grande flexibilité quant au type de données traitées (observations manquantes dans certaines régions de l'image par exemple). Par ailleurs nous traitons des données sous l'hypothèse de corrélation spectrales entre les bandes selon le modèle défini dans [9]. Les paramètres (lois d'attaches aux données supposées gaussiennes et paramètres de la distribution a priori) sont estimés au moyen de l'algorithme EM. L'étape de décision finale met en jeu le critère MPM [10] adapté au contexte du fbu.

Nous présentons dans un premier temps le modèle d'arbre de Markov caché fbu. Puis, nous détaillons la procédure d'estimation bayésienne du processus caché ainsi que l'algorithme d'estimation de type EM des paramètres. Enfin nous proposons deux expérimentations, l'une étant basée sur une image trispectrale sont le processus caché est un champ de Markov fbu et l'autre étant basée sur une jeu d'images astronomiques réelles dans le domaine de l'infrarouge.

2 Arbre de Markov flou

2.1 Loi du processus caché

Considérons un modèle à deux classes dures, notées "0" et "1". Une telle hypothèse est réaliste dans le contexte de l'imagerie astronomique s'il s'agit de détecter un objet à partir d'un fond homogène. Chaque site peut être alors défini sur l'intervalle $[0, 1]$. Soit S un ensemble fini d'indices et $S^1, \ldots, S^n$ une partition de S. Posons $N = Card(S)$ et, pour $1 \leq i \leq n$, $N^i = Card(S^i)$. Supposons que $N^1 = 1 \leq N^2 \leq \ldots \leq N^n$ et pour $1 \leq i \leq n-1$, associons à chaque $s \in S^i$ un sous-ensemble s^+ de S^{i+1}, appelé "enfants de s", de manière à ce que $(s^+)_{s \in S^i}$ soit une partition de S^{i+1}. L'unique élément de S^1 est appelé "racine". Par ailleurs, s^{++} désignera l'ensemble des descendants de s et pour tout $s \in S - S^1$, son unique "père" sera noté s^-.

Soit $X = (X_s)_{s \in S}$ un processus aléatoire à valeur dans

$[0,1]^N$. La loi de X_s est définie par une densité h_s selon la mesure $\nu = \delta_0 + \delta_1 + \mu$. Les masses de Dirac δ_0, δ_1 et la mesure de Lebesgue μ sur $]0,1[$, représentent respectivement l'information dure et l'information fbue [11]. La condition de normalisation implique : $h_s(0) + h_s(1) + \int_0^1 h_s(t)dt = 1$. La loi de X est définie par la loi initiale $h(x)$ et par la densité de transition $f(x_s|x_{s^-})$:

$$\pi(x) = h_s(x_1) \prod_{i=2}^{n} \prod_{s \in S^i} f(x_s|x_{s^-}) \quad (1)$$

2.2 Loi d'observation multispectrale

Soit une réalisation $\mathbf{y} = (y_s^{(1)}, y_s^{(2)}, \ldots, y_s^{(D)})_{s \in S}$ du vecteur aléatoire $\mathbf{Y} = (Y_s^{(1)}, Y_s^{(2)}, \ldots, Y_s^{(D)})_{s \in S}$. La variable $\mathbf{Y}$ représente l'observation d'une scène unique par D capteurs. Conditionnellement aux états x_s, les observations sont supposées spatialement indépendantes mais spectralement corrélées. La distribution de $\mathbf{y}_s = (y_s^{(1)}, y_s^{(2)}, \ldots, y_s^{(D)})$ conditionnellement à x_s est modélisée par une densité gaussienne multivariée

$$f_{x_s}(\mathbf{y_s}) = \frac{1}{(2\pi)^{D/2} (\det \mathbf{\Gamma}_{x_s})^{1/2}} . e^{(-\frac{1}{2}(\mathbf{y_s} - \mu_{x_s})^T \mathbf{\Gamma}_{x_s}^{-1} (\mathbf{y_s} - \mu_{x_s}))} \quad (2)$$

où $\mu_{x_s} = [\mu_{x_s}^{(1)}, \mu_{x_s}^{(2)}, \ldots, \mu_{x_s}^{(D)}]^T$ et $\mathbf{\Gamma}_{x_s} \in \mathbb{R}^D \times \mathbb{R}^D$ définissent respectivement le vecteur moyenne et la matrice de variance-covariance pour chaque état x_s.

Soient (μ_0, μ_1) et $(\mathbf{\Gamma}_0, \mathbf{\Gamma}_1)$ les vecteurs moyennes et les matrices de covariance relatifs aux classes dures "0" et "1". Nous proposons pour chaque site fbu x_s, le vecteur moyenne μ_{x_s} et la matrice $\mathbf{\Gamma}_{x_s}$ suivants [9]:

$$\mu_{x_s} = (1 - x_s)\mu_0 + x_s \mu_1 \quad , \quad \mathbf{\Gamma}_{x_s} = (1 - x_s)^2 \mathbf{\Gamma}_0 + x_s^2 \mathbf{\Gamma}_1$$

Finalement, la loi $f(x, \mathbf{y})$ du couple $(X, \mathbf{Y})$ s'écrit

$$f(x, \mathbf{y}) = h_s(x_1) f_{x_1}(\mathbf{y}_1) \prod_{i=2}^{n} \prod_{s \in S^i} f(x_s|x_{s^-}) f_{x_s}(\mathbf{y}_s) \quad (3)$$

3 Estimation du processus caché

Etant donné une observation $\mathbf{Y} = \mathbf{y}$, on souhaite estimer une réalisation $x \in [0,1]^N$. Le processus de décision finale, fondée sur une stratégie bayésienne utilise le critère MPM [10]. Étant donnée une fonction de perte $L^* : [0,1]^S \times [0,1]^S \to \mathbb{R}^+$ et un ensemble d'observations $\mathbf{Y} = \mathbf{y}$, la décision $\widehat{d}(\mathbf{Y})$ minimise l'espérance $E[L^*(X, \widehat{d}(\mathbf{Y}))]$. Le critère MPM revient à minimiser $E[L^*(X_s, \widehat{d}(\mathbf{Y}))]$ en chaque site s. Nous choisissons le coût "distance en valeur absolue" : $L(x_s, \widehat{x}_s) = |x_s - \widehat{x}_s|$, qui présente deux avantages: performance en terme de robustesse (bruit, hyperparamètres) et compatibilité avec la fonction de perte, "0-1" de la segmentation dure. Pour une réalisation $\mathbf{Y} = \mathbf{y}$ nous minimisons l'espérance conditionnelle :

$$E[L(X_s, \widehat{d}(\mathbf{Y}))|\mathbf{Y} = \mathbf{y}] \quad (4)$$

Le calcul de (4) nécessite la distribution a posteriori $h_s^{\mathbf{Y}}$ en X_s :

$$E[L(X_s, \widehat{d}(\mathbf{Y}))|\mathbf{Y} = \mathbf{y}] = h_s^{\mathbf{y}}(0).L(0, \widehat{d}(\mathbf{y})) + h_s^{\mathbf{y}}(1)L(1, \widehat{d}(\mathbf{y})) + \int_{t \in]0,1[} h_s^{\mathbf{y}}(t) L(t, \widehat{d}(\mathbf{y}))dt \quad (5)$$

La décision $\widehat{d}(\mathbf{y}) \in [0,1]$ en chaque site est celle qui minimise (5). Nous devons donc calculer au préalable les densités $h_s^{\mathbf{y}}(x_s)$. Pour cela, les transitions $f^{\mathbf{y}}(x_s|x_{s^-})$ sont d'abord calculées dans une passe montante sur l'arbre en fonction des probabilités backward $\beta(x_s) = f(\mathbf{y}_{s^{++}}|x_s)$

$$f^{\mathbf{y}}(x_s|x_{s^-}) = \frac{f(x_s|x_{s^-}) f_{x_s}(\mathbf{y}_s) \beta_s(x_s)}{\int_0^1 f(\omega_s|x_{s^-}) f_{\omega_s}(\mathbf{y}_s) \beta_s(\omega_s) d\nu(\omega_s)} \quad (6)$$

ces dernières étant calculées récursivement en posant

$$\begin{aligned} &\beta_s(x_s) = 1 \text{ pour } s \in S^n \\ &\beta_s(x_s) = \prod_{t \in s^+} \left(\int_0^1 \beta_t(x_t) f(x_s|x_{s^-}) f_{x_s}(\mathbf{y}_s) d\nu(x_s) \right) \\ &\qquad \text{pour } s \in S - S^n \end{aligned} \quad (7)$$

Pour $s = S^1$ (racine de l'arbre), la probabilité $h_1^{\mathbf{y}}(x_1)$ est calculée à partir de $\beta_1(x_1)$.

$$h_1^{\mathbf{y}}(x_1) = \frac{h_1(x_1) f_{x_1}(\mathbf{y}_1) \beta_1(x_1)}{\int_0^1 h_1(\omega_1) f_{\omega_1}(\mathbf{y}_1) \beta_1(\omega_1) d\nu(\omega_1)} \quad (8)$$

Les densités $f^{\mathbf{y}}(x_s, x_{s^-})$ sont alors calculées dans une passe descendante

$$f^{\mathbf{y}}(x_s, x_{s^-}) = h_{s^-}(x_{s^-}) f^{\mathbf{y}}(x_s|x_{s^-}) \quad (9)$$

On en déduit la valeur des marginales a posteriori $h_s^{\mathbf{y}}(x_s)$[12].

$$h_s^{\mathbf{y}}(x_s) = \int_0^1 f^{\mathbf{y}}(x_s, x_{s^-}) d\nu(x_{s^-}) \quad (10)$$

Par la suite, les intégrales ci-dessus sont calculées à partir de la discrétisation de l'intervalle $[0,1]$ en M valeurs équidistantes : $\{\varepsilon_0 = 0, \varepsilon_1 = \frac{1}{M}, \ldots, \varepsilon_i = \frac{i}{M}, \varepsilon_M = 1\}$

4 Estimation des paramètres

Nous proposons l'estimation des paramètres suivante. X est supposé stationnaire, sa loi donnée par $c_{ij} = f(x_s = j, x_{s^-} = i)$ ne dépendant pas de s. L'estimation des paramètres est effectuée par un algorithme EM [12, 13] de la manière suivante

1. initialisation : $\theta^0 = (c_{ij}^0, \mu_k^0, \mathbf{\Gamma}_k^0)_{0 \le i,j \le 1, 0 \le k \le 1}$
2. pour tout $q \in \mathbb{N}^+$, θ^{q+1} est calculée à partir de $\mathbf{y}$ et θ^q par

$$c_{ij}^{q+1} = \frac{1}{\sum_{i=2}^n N^i} \sum_{i=2}^{n} \sum_{s \in S^i} f^{\mathbf{y}}(x_s, x_{s^-}) \quad (11)$$

pour $k \in \{0,1\}$ (classes dures) : avec $A = \sum_{i=1}^n N^i$

$$\begin{aligned} &\hat{\mu}_k^{q+1} = \tfrac{1}{A} \textstyle\sum_{i=1}^n \sum_{s \in S^i} h_s^{\mathbf{y}}(k) \mathbf{y}_s \\ &\hat{\mathbf{\Gamma}}_k^{q+1} = \tfrac{1}{A} \textstyle\sum_{i=1}^n \sum_{s \in S^i} h_s^{\mathbf{y}}(k) \left(\mathbf{y}_s - \hat{\mu}_k^{q+1} \right) \left(\mathbf{y}_s - \hat{\mu}_k^{q+1} \right)^T \end{aligned}$$

pour $k \in]0,1[$ (classes fbues) :

$$\begin{aligned} &\hat{\mu}_k^{q+1} = (1-k)\hat{\mu}_0^{q+1} + k\hat{\mu}_1^{q+1} \\ &\hat{\mathbf{\Gamma}}_k^{q+1} = (1-k)^2 \mathbf{\Gamma}_0^{q+1} + k^2 \mathbf{\Gamma}_1^{q+1} \end{aligned} \quad (12)$$

5 Expérimentations

Nous avons testé notre algorithme sur une image trispectrale (Fig. 2(a)-(c)) dont la composante"vérité-terrain" est un champ de Markov fbu à 10 niveaux présenté Fig. 1. Les résultats de la

segmentation sont donnés Fig. 2(d). L'algorithme restitue correctement l'image d'origine avec un taux d'erreur de 4.58 %. Les paramètres des lois d'attache aux données estimés selon l'algorithme section 4 sont indiqués dans le tableau 2. L'information fbue est correctement restituée (44.96 % vs 53.96 %) et de ce point de vue, cette méthode donne de meilleurs résultats que l'approche multibandes par champ de Markov [9].

La Fig. 3(a)-(c) montre trois images qui se prêtent bien "visuellement" au modèle fbu. Ce sont des images astronomiques dans le domaine infrarouge à 60, 100 et 170 microns, observées à partir d'un niveau de résolution unique. Il s'agit de champs de haute altitude de notre galaxie pour lesquels on cherche à mesurer les fluctuations induites par la présence de galaxies extérieures à la voie lactée, non résolues dans le cadre d'une étude cosmologique. Des fluctuations, provoquées par la poussière interstellaire de notre galaxie, gênent l'observation. Cette poussière interstellaire (en bas à droite) possède une structure filamentaire que nous souhaitons mettre en évidence par une analyse fbue multibande. Le résultat d'une segmentation MPM sur 10 niveaux donne la structure en arbre multirésolution observée Fig. 3(d). Les paramètres des moyennes et variances obtenus pour chaque bande sont donnés tableau 3. Ces résultats montre également l'intérêt des méthodes par arbre, dans la mesure où il est possible de traiter des données dont l'observation manque dans certaines bandes, ou à certains niveaux de résolution. La méthode par arbre est également pertinente car elle permet d'étudier l'image pour certaines zones appartenant à la même classe (au lieu de toute l'image) et d'en extraire des paramètres cosmologiques par classes, dures ou fbues.

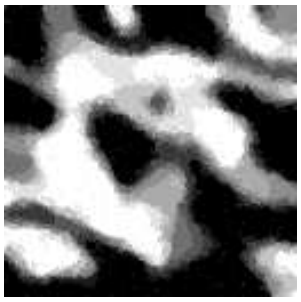


FIG. 1: *Image floue originale simulée par champ de Markov flous à 2 classes dures et 8 niveaux de flou. Le taux de flou est 53.96 %*

TAB. 1: *: Paramètres d'attache aux données de Fig. 2(a)-(c)*

$\hat{\mu}_0$	120	120	120
$\hat{\sigma}_0^2$	16	64	256
$\hat{\mu}_1$	146	146	146
$\hat{\sigma}_1^2$	16	64	256

TAB. 2: *: Paramètres estimés de Fig. 2*

$\hat{\mu}_0$	119.9	119.8	119.9
$\hat{\sigma}_0^2$	14.0	63.0	249.4
$\hat{\mu}_1$	146.0	145.9	145.8
$\hat{\sigma}_1^2$	13.8	64.0	244.0

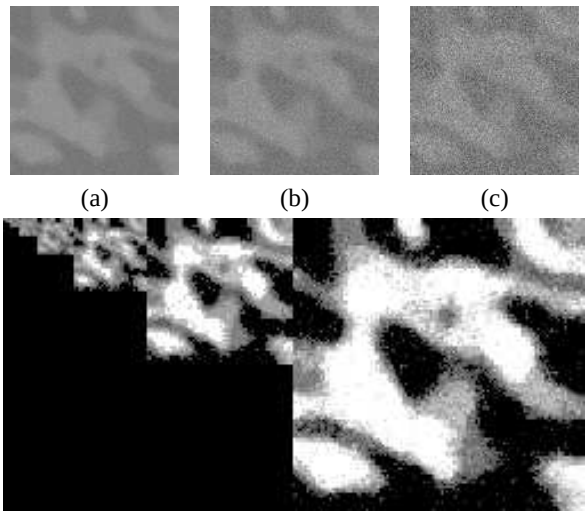


FIG. 2: *Segmentation de l'image de synthèse multispectrale par arbre de Markov caché flou. (a)-(c): les données multibandes; (d): résultats de la segmentation à 2 classes dures et 8 niveaux de flou. Taux de flou de l'image segmentée=44.61%, taux d'erreur=4.58%*

6 Conclusion

Au cours de cette étude, nous avons présenté la segmentation statistique non supervisée basée sur un nouveau modèle d'arbre de Markov fbu dans un cadre multispectral. Ce modèle s'appuie localement sur une densité de probabilité par rapport à la mesure constituée d'une composante continue et de composantes discrètes. Nous avons supposé les capteurs corrélés, en exprimant les lois du bruit par des gaussiennes multivariées dont les matrices des variances-covariances des classes fbues sont déduites linéairement des matrices de variances-covariances des classes dures. Les méthodes ont été testées sur des images de synthèse trispectrales puis appliquées à des données astronomiques. Notre approche par arbre de Markov multispectrale autorise une très grande flexibilité: elle est particulièrement adaptée aux données astronomiques multiéchelles et multibandes y compris lorsque des observations manquent à certains niveaux de résolution. Les premiers résultats encourageants montrent l'intérêt des méthodes non supervisées fbues dans le champ de l'astronomie, en restituant ainsi une gradation fbue entre les zones homogènes du ciel et permet de ne traiter qu'une région particulière de l'image. Nous envisageons d'enrichir le modèle en introduisant d'autres lois paramétriques [14] pour caractériser les observations.

Remerciements

Les auteurs remercient G. Lagache (Institut d'Astrophysique Spatiale d'Orsay) pour la fourniture des images astronomiques.

Références

[1] L. Zadeh, "Fuzzy sets," *Information and control*, vol. 8, pp. 338–352, 1965.

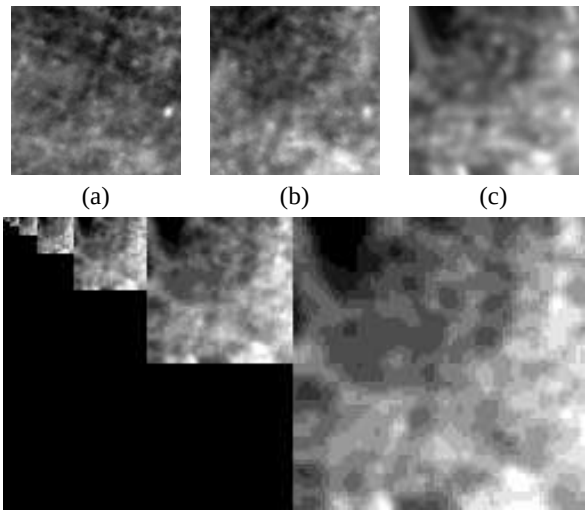


FIG. 3: *Segmentation d'une image astronomique multispectrale par arbre de Markov caché flou. (a)-(c): les données multibandes; (d): résultats de la segmentation à 2 classes dures et 8 niveaux de flou.*

TAB. 3: *: Paramètres estimés de Fig. 3(a)-(c)*

$\hat{\mu}_0$	51.6	34.0	14.0
$\hat{\sigma}_0^2$	301.5	175.0	52.9
$\hat{\mu}_1$	132.8	198.0	225.4
$\hat{\sigma}_1^2$	1019.7	1420	297.6

[2] M. H. Bloch I., "Fusion de données en traitement d'images : modèles d'informations et décisions," *Traitement du Signal*, vol. 11, no. 6, pp. 435–446, 1994.

[3] M. Crouse, R. Novak, and R. Baraniuk, "Wavelet-based signal processing using hidden markov models," *IEEE Transactions on Signal Processing*, vol. 46, no. 4, pp. 886–902, 1998.

[4] M. Kent J.T., "Spatial classification using fuzzy membership," *IEEE, Trans. On Pattern Analysis and Machine Intelligence*, vol. 10, no. 5, pp. 659–671, 1988.

[5] F. Salzenstein and W. Pieczynski, "Parameter estimation in hidden fuzzy markovian fields and image segmentation," *Graphical Models and image processing*, vol. 59, no. 4, pp. 205–220, 1997.

[6] S. Ruan, B. Moretti, J. Fadili, and D. Bloyet, "Fuzzy markovien segmentation in application of magnetic resonance images," *Computer Vision and Iimage Understanding*, vol. 85, pp. 54–69, 2002.

[7] S. Chevalier, M. Kaynak, A. Cheok, K. Sengupta, and K. Chung, "Use of a novel generalized fuzzy hidden markov model for speech recongnition," in *Proceedings of the IEEE Fuzzy Systems Conference (FUZZ-IEEE 2001)*, vol. 3, pp. 1207–1220, dec. 2001.

[8] C. Carincotte, S. Derrode, and J.-M. Boucher, "Chaînes de markov cachées fbues et segmentation non supervisée d'images," in *Logique Floue et Applications (LFA'04)*, (Nantes, France), 18-19 Novembre 2004.

[9] F. Salzenstein, C. Collet, and M. Petremand, "Champs de Markov fbus pour imagerie multispectrale - fuzzy Markov random fields for multispectral images," *Traitement du Signal*, vol. 21, no. 1, pp. 37–55, 2004.

[10] T. P. J. Maroquin, S.Mitte, "Probabilistic solution of ill-posed problems in computational vision," *Journal of the american statistical association*, vol. 82, pp. 76–89, 1987.

[11] H. Caillol, A. Hillion, and W. Pieczynski, "Fuzzy random Markov fields and unsupervised image segmentation," *IEEE Trans. Geosci. Remote Sensing*, vol. 4, pp. 801–810, 1993.

[12] J.-M. Laferté, P. Pérez, and F. Heitz, "Discrete Markov image modeling and inference on the quadtree," *IEEE Trans. on Image Processing*, vol. 9, no. 3, pp. 390 – 404, 2000.

[13] G. McLachlan and T. Krishnan, *EM Algorithm and Extensions*. Wiley, 1997.

[14] Y. Delignon, A. Marzouki, and W. Pieczynski, "Estimation of generalized mixtures and its application in image segmentation," *IEEE Transactions on Image Processing*, vol. 6, no. 10, pp. 1364–1375, 1997.

Segmentation non supervisée d'images non stationnaires avec champs de Markov évidentiels

D. BENBOUDJEMA et W. PIECZYNSKI

INT/GET, Dept. CITI, CNRS UMR 5157, 9 rue Charles Fourier, 91000 Evry Cedex

{dalila.benboudjema, wojciech.pieczynski}@int-evry.fr

Résumé – Fréquemment utilisés en traitement statistique d'images, les champs de Markov cachés (CMC) sont des outils puissants qui peuvent fournir des résultats remarquables. Cette qualité est principalement due à l'aptitude du modèle de prendre en compte des dépendances spatiales des variables aléatoires, même lorsqu'elles sont en très grand nombre, pouvant dépasser le milion. Dans un tel modèle le champ caché X est supposé markovien et doit être estimé à partir du champ observé Y. Un tel traitement est possible du fait de la markovianité de X conditionnellement Y. Ce modèle a été ensuite généralisé au champs de Markov couples (CMCouple), où l'on suppose directement la markovianité du couple (X,Y), qui offrent les mêmes possibilités de traitements que les CMC et permettent de mieux modéliser le bruit ce qui permet, en particulier, de mieux prendre en compte l'existence des textures. Par la suite, les CMCouples ont été généralisés aux champs de Markov triplet (CMT), où la loi du couple (X,Y) est une loi marginale d'un champ de Markov triplet $T=(X,U,Y)$, avec un champ auxiliaire U. Par ailleurs, la théorie de l'évidence peut permettre une amélioration des résultats obtenus par des traitements bayésiens dans certaines situations. Le but de cet article est d'aborder le problème de la segmentation non supervisée d'images non stationnaires en utilisant les champs de Markov évidentiels (CME), en exploitant, en particulier, un lien existant entre les CME et les CMT.

Abstract – The hidden Markov fields (HMF) are widely used in statistical image processing. Their success is mainly due to the fact that they are able to take into acount the spatial dependencies, even in a very large pixel set case. In such model the hidden field X which is assumed Markovian, must be estimated from the observed field Y. Such processing is possible because the distribution X conditional on the observed process Y remains markovian. HMF has been extended to the Pairwise Markov field (PMF) model, in which one directly assumes the markovianity of the couple (X,Y), and which offer similar processing capabilities and superior modelling ones. Afterwards, triplet Markov fields (TMF), which extend the PMF, have been proposed. In such model the distribution of the couple (X,Y) is the marginal distribution of a Markov field $T=(X,U,Y)$, where U is some auxiliary process. Otherwise, the theory of evidence can improve the bayesian prossesing results in some situations. The aim of this paper is to tackle the problem of the unsupervised non stationnary images using the evidential Markov fields (EMF) and show the connexion between such model and the TMF one.

1. Introduction

Les champs de Markov sont des outils probabilistes assez largement utilisés en traitement d'images [7, 11, 20]. Leur intérêt majeur est de pouvoir modéliser les dépendances spatiales des variables aléatoires, dont les réalisations modélisent les quantités observées ou recherchées, de façon relativement simple et se prêtant bien aux divers traitements. Divers modèles fondés sur les champs markoviens ont été proposés, parmi lesquels les champs de Markov cachés (CMC), très utilisés en segmentation d'images. Ceux-ci consistent à considérer deux champs stochastiques $X=(X_s)_{s\in S}$ et $Y=(Y_s)_{s\in S}$, avec S ensemble des pixels, dans lesquels chaque variable aléatoire X_s est à valeurs dans l'ensemble fini des classes $\Omega=\{\omega_1,...,\omega_k\}$ et chaque Y_s est à valeurs dans l'ensemble des nombres réels R. En segmentation d' images, le problème est d' estimer le champ inobservable directement (ou « caché ») $X=x$ à partir du champ observé $Y=y$. Dans le modèle CMC classique X est de Markov et la forme du bruit (loi de Y conditionnelle à X) est simple, ce qui implique la markovianité de X *a posteriori* (conditionnelle à Y). Cette dernière markovianité permet alors de mettre en œuvre les différents traitements bayésiens [7, 11, 20]. Cependant, cette même simplicité est difficile à justifier et peut poser problème dans certaines situations, comme la présence des textures [10]. Dans le modèle « champ de Markov couple » (CMCouple), plus général que CMC dans la mesure où X n'est plus nécessairement de Markov, on suppose directement la markovianité de $Z=(X,Y)$ [14, 15]. Le champ Y est alors markovien conditionnellement à X (ce qui autorise une modélisation plus complexe des bruits), et le champ X est markovien conditionnellement à Y (ce qui autorise les mêmes traitements que dans les CMC). Par la suite, les CMCouple ont été généralisés aux « champs de Markov triplet » (CMT) dans lesquels on introduit un troisième champs aléatoire $U=(U_s)_{s\in S}$, où chaque v.a. U_s est à valeurs dans l' ensemble$\Lambda=\{\lambda_1,...,\lambda_M\}$, et l'on suppose la markovianité du triplet $T=(X,U,Y)$ [13]. Le modèle CMT est alors plus général que les CMCouple car la loi de $Z=(X,Y)$, qui est une loi marginale de celle de $T=(X,U,Y)$, n'est plus nécessairement markovienne. Cependant, lorsque le cardinal de l'ensemble Λ n'est pas trop élevé, les CMT autorisent encore les traitements bayésiens permettant d'estimer $X=x$ à partir de $Y=y$ [4, 5, 13]. Notons que Λ peut avoir une signification concrète ou pas. A titre d'exemple, Λ peut modéliser l'existence des

sous-classes (voir [5]), ou encore l'existence des diverses stationnarités du champs des classes X (voir [4]).

Par ailleurs, la théorie de l'évidence [1, 17, 18] peut permettre une amélioration des résultats obtenus par des traitements bayésiens dans un certain nombre de situations [1]. Cette théorie est particulièrement bien adaptée à la fusion des informations fournies par des capteurs de nature différentes comme, à titre d'exemple, un capteur optique et un capteur radar dans une image satellitaire (voir [19]). Cependant, elle a été très peu utilisée, jusqu'à présent, dans le contexte des champs de Markov cachés [2, 6, 16].

L'objet de cet article est de montrer l'intérêt de l'introduction des fonctions de masse au niveau de la loi *a priori* d'un champs de Markov caché en segmentation non supervisée d'images non stationnaires. En procédant de manière proche à celle utilisée dans le contexte des chaînes de Markov cachées (voir [8, 9]), nous utilisons le fait que l'introduction d'un certain type de l'affaiblissement au niveau de la loi *a priori* aboutit, formellement, à un champ de Markov triplet. Le problème de l'estimation des paramètres dans les champs triplet étant proche de celui se posant dans les champs couple, nous adaptons la méthode proposée dans [3, 5]. Les résultats de diverses expérimentations, dont une est présentée en section 5, attestent de l'intérêt de la méthode proposée.

2. Théorie de l' évidence

La théorie de l' évidence a été introduite en 1967 par Dempster puis reprise par Shafer en 1976 [17]. Parfois décrite comme une généralisation de la théorie probabiliste classique, elle a l' avantage de pouvoir représenter de manière parfois plus complète l' incertitude sur un événement. Elle peut également permettre une amélioration des résultats obtenus par des traitements bayésiens dans un certain nombre de situations. Cela est en particulier vrai lorsque les paramètres du modèles sont mal connus, auquel cas on remplace certaines lois de probabilité, par exemple la loi *a priori* de X, par des « fonctions de masse » évidentielles. Une telle fonction de masse M est une application de $P(\Omega)$ (ensemble des parties de Ω) dans $[0,1]$ vérifiant $M(\emptyset)=0$ et $\sum_{A\in P(\Omega)} M(A)=1$. Notons qu'une fonction de masse M définit les « plausibilités » $Pl(A)=\sum_{A\cap B\neq\emptyset} M(B)$ et les « crédibilités » $Cr(A)=\sum_{B\subseteq A} M(B)$. Une des possibilité de définition d'une fonction de masse à partir d'une loi de probabilité est l'utilisation d'un « affaiblissement ». Cette démarche peut être utilisée pour la loi *a priori* de X, lorsque cette dernière est connue trop imparfaitement. La loi *a posteriori* de X est alors obtenue par la fusion de Dempster-Shafer de la masse évidentielle obtenue par affaiblissement avec une loi de probabilité définie par les observations $Y=y$. Cette démarche n'est pas immédiate dans le cas des modèles de Markov cachés car la fusion de Dempster-Shafer détruit la markovianité ; cependant, il a été montré que le résultat de cette fusion est la loi marginale d'un modèle « triplet », ce qui autorise les traitements d'intérêt [12]. En particulier, lorsque la loi *a priori* est non stationnaire et inconnue, son estimation donnera nécessairement une loi erronée et son affaiblissement peut éventuellement présenter un intérêt. En effet, cet intérêt a été montré dans le cas des chaînes de Markov cachées dans [8, 9], et nous montrons ci-dessous qu'il peut également exister dans le cas des champs de Markov.

3. Champs de Markov évidentiels

Soit S un ensemble de pixels, X un champ de Markov où chaque v.a X_s est à valeur dans un ensemble fini de classes $\Omega=\{\omega_1,...,\omega_k\}$, avec la loi $p(x)\propto\exp[-W(x)]$ donnée par l'énergie :

$$W(x)=\sum_{(s,t)\in C_H}\alpha_H(1-2\delta(x_s,x_t))+\sum_{(s,t)\in C_V}\alpha_V(1-2\delta(x_s,x_t)) \quad (1)$$

où C_H (respectivement C_V) est l'ensemble des cliques d'ordre deux horizontales (respectivement verticales), et $\delta(x_s,x_t)$ vérifie $\delta(x_s,x_t)=1$ pour $x_s=x_t$, et $\delta(x_s,x_t)=0$ pour $x_s\neq x_t$. Avec les hypothèses classiques portant sur le bruit, la loi du couple (X,Y) est donnée par:

$$p(x,y)=\gamma\exp[-W(x)+\sum_{s\in S} Log(p(y_s|x_s))] \quad (2)$$

Le modèle ainsi défini est un champ de Markov caché classique.

Plaçons nous maintenant dans le cadre évidentiel. Dans les champs de Markov évidentiels (CME), on introduit un champ $U=(U_s)_{s\in S}$ où chaque U_s est à valeurs dans l' ensemble des parties de Ω notée $\Lambda=P(\Omega)$. La fonction de masse M_1 est un champ de Markov évidentiel (CME) défini sur $(P(\Omega))^n$ (avec $n=Card(S)$) si:

$$M_1(A)=\gamma\exp-\left(\sum_{c\in C}\psi_c(A_c)\right) \quad (3)$$

où $A=(A_s)_{s\in S}$ et $A_c=(A_s)_{s\in c}$, l' ensemble des cliques C correspondant à un voisinage donné.

En posant $M_2\propto\prod_{s\in S} p(y_s|x_s)$ la probabilité sur Ω^n définie à partir du processus observé $Y=y\in R^n$ ($y=(y_s)_{s\in S}$ est fixé), la probabilité $M=M_1\oplus M_2$ est la probabilité conditionnelle $p(x|y)$, avec $p(x,y)$ la loi marginale d'un CMT dont l'énergie est donnée comme suit :

$$\begin{aligned}\varphi_c(t_c)&=\varphi_c(x_c,u_c,y_c)\\&=\begin{cases}\psi_c(u_c) & pour \quad Card(c)>1\\ \psi_c(u_c)+\log(p(y_s|x_s)) & pour \quad c=\{s\}\end{cases}\end{aligned} \quad (4)$$

En effet, $M=M_1\oplus M_2$ s' écrit

$$\begin{aligned}M(x)&=(M_1\oplus M_2)(x)\\&\propto\sum_{x\in u}\left[\exp\left(-\sum_{c\in C}\psi_c(u_c)\right)\prod_{s\in S}p(y_s|x_s)\right]\\&\propto\sum_{x\in u}\left[\exp\left(-\sum_{c\in C}\psi_c(u_c)+\sum_{s\in S}\log(p(y_s|x_s))\right)\right]\end{aligned} \quad (5)$$

Notons que dans la somme figurant dans (5) $x=(x_s)_{s\in S}$ est fixé et $u=(u_s)_{s\in S}$ varie dans l' ensemble $[P(\Omega)]^n$. Plus

précisément, l'énergie dans (5) est définie pour $(x,u)\in\Delta^n$ avec $\Delta\subset\Omega\times\Lambda$ défini par $\Delta=\{(\omega,A)|\omega\in A\}$.
A titre d'exemple, pour un champ de Markov à deux classes on a: $\Omega=\{\omega_1,\omega_2\}$, $\Lambda=P(\Omega)=\{\varnothing,\{\omega_1\},\{\omega_2\},\{\omega_1,\omega_2\}\}$, et le couple (X_s,U_s) prend ses valeurs dans $\{(\omega_1,\{\omega_1\}),(\omega_1,\{\omega_1,\omega_2\}),(\omega_2,\{\omega_2\}),(\omega_2,\{\omega_1,\omega_2\})\}$. La fonction de masse $M(x)$ devient:

$$M(x)=\sum_{u|(x,u,y)\in\Delta^n\times R^n}\left[\exp-\left(\sum_{c\in C}\varphi_c(u_c)-\sum_{s\in S}\log\left(p(y_s|x_s)\right)\right)\right] \quad (6)$$
$$=\sum_{u|(x,u,y)\in\Delta^n\times R^n}\exp-\varphi(x_c,u_c,y_c)$$

Ainsi on retrouve l'expression du champ Triplet $T=(X,U,Y)$ donné comme suit :

$$p(x,u,y)=\gamma\exp[-W(x,u)+\sum_{s\in S}Log(p(y_s|x_s))] \quad (7)$$

On peut ainsi dire que le résultat de la fusion de Dempster - Shafer est la loi conditionnelle $p(x|y)$ induite par un champ Triplet $T=(X,U,Y)$, avec le processus auxiliaire U à valeurs dans l'ensemble des parties $P(\Omega)$ de Ω. D' autre part, pour $x=u$ l'affaiblissement est réduit à néant, on retrouve l'énergie $W(x)$ classique du champ de Markov caché (équation (1)).

Remarque
Notons que la théorie de l'évidence peut également être utilisée au niveau des capteurs, comme proposé dans [2], l'introduction des champs de Markov triplet. Supposons que l'on a toujours deux classes et l'on souhaite utiliser le modèle classique donné par (1) et (2) en présence des nuages. Le champs de Markov $p(x)\propto\exp[-W(x)]$ n'est pas affecté par les nuages ; par contre, la probabilité $M_2\propto\prod_{s\in S}p(y_s|x_s)$ change de nature et tout se passe comme s'il y avait trois classes : ω_1, ω_2, et « nuages ». La troisième classe « nuages » peut alors être interprétée comme une classe n'apportant aucune information d'intérêt et assimilée à Ω. Finalement, la probabilité $M_2\propto\prod_{s\in S}p(y_s|x_s)$ fournie par les capteurs peut être interprétée comme une masse définie sur $\{\{\omega_1\},\{\omega_2\},\{\omega_1,\omega_2\}\}^n$ et l'on montre que sa fusion avec le champ de Markov $p(x)\propto\exp[-W(x)]$ reste un champ de Markov. Ce dernier apparaît alors immédiatement comme une extension du champs de Markov *a posteriori* classique et peut être classiquement utilisé à des fins de segmentation.

4. Apprentissage

Dans cette section nous décrivons brièvement une méthode, proche de celle proposée dans [3], d' estimation des paramètres du modèle. Elle utilise moindres carrés pour l' estimation des paramètres du champ de Markov (U,X), et des méthodes classiques utilisant les moments pour l' estimation des paramètres du bruit. Comme dans [3] et afin de rendre cette méthode automatique l'estimation conditionnelle itérative (ECI) est utilisée. L' ECI est une procédure générale d' estimation des paramètres dans le cas des données cachées [15]. Soit X et Y deux champs aléatoires dont la loi conjointe dépend d'un paramètres θ, et soit $\hat{\theta}=\hat{\theta}(X,Y)$ un estimateur de celui ci à partir des données complètes. Le champ X étant inobservable, l'idée de l'ECI est d'approcher $\hat{\theta}=\hat{\theta}(X,Y)$ par son espérance conditionnelle à $Y=y$. Or, l'espérance conditionnelle de $\hat{\theta}$ notée $E[\hat{\theta}|Y]$, dépend du paramètre θ. Ainsi, pour calculer $E[\hat{\theta}|Y]$, on est amené à considérer la valeur θ_q (valeur courante) du paramètre. La valeur suivante est donnée par $\theta_{q+1}=E_{\theta_q}[\hat{\theta}|Y]$. La procédure itérative ECI est donc :

- On se donne une valeur θ_0
- On calcule la valeur θ_{q+1} (valeur suivante) à partir de θ_q (valeur courante) et de $Y=y$ par:

$$\theta_{q+1}=E_q[\hat{\theta}(X,Y)|Y=y] \quad (8)$$

où E_q est l'espérance conditionnelle correspondant à θ_q.

$E_q[\hat{\theta}|Y=y]$ n'est pas explicitement calculable, mais la simulation des réalisations de X à partir de la loi *a posteriori* est possible. Alors on utilise l'approximation suivante :

$$\theta_{q+1}=\frac{1}{r}E_q\left[\hat{\theta}(x_1,y)+\cdots+\hat{\theta}(x_r,y)\right] \quad (9)$$

Où $x_1,x_2,\cdots,x_r$ sont les r simulations des réalisations de X selon la loi conditionnelle à Y.

5. Expérimentation

Considérons une image non stationnaire à deux classes illustrée en FIG. 1 (a) et sa version bruitée (b), avec un bruit blanc gaussien de variance 1 et de moyennes 0 et 2. L'utilisation de la méthode bayésienne Maximum des Marginales a Posteriori (MPM) fondée sur le modèle affaiblit triplet (5) donne 9.37% de taux d'erreur (image (d)), alors que celle fondée sur le modèle CMC classique (1), (2) donne 22.1% de taux d' erreur (image (c)). Dans les deux cas, la segmentation est non supervisée dans laquelle les paramètres ont été initialisés en utilisant la méthode des k - means et estimés par la méthode de type ECI décrite ci-dessus. Les paramètres ainsi obtenus sont représentés dans TAB. 1.

TAB. 1 : valeurs des paramètres estimés

Paramètres	CMC	CME
μ_1	0.89	-0.2
μ_2	2.05	1.99
σ_1	1.38	0.87
σ_2	0.99	0.99
α_{1H},α_{1V}	0.64, 0.64	0.24, 0.28
α_{2H},α_{2V}		0.35, 0.26
α_{3H},α_{3V}		0.00, 0.02

TAB. 1 montre que les paramètres du bruit estimés à partir du modèle évidentiel sont plus proches des paramètres réels que ceux estimés à partir des CMC classiques. Par ailleurs, aucune conclusion ne peut être tirée quant à l' estimation des paramètres d' interaction α_i étant donnée que leurs vraies valeurs sont inconnues. Nous pouvons cependant observer que l'utilisation des CME a permis l' amélioration du résultat

de segmentation non supervisée, et donc conjecturer que ce dernier modèle représente mieux l'image considérée.

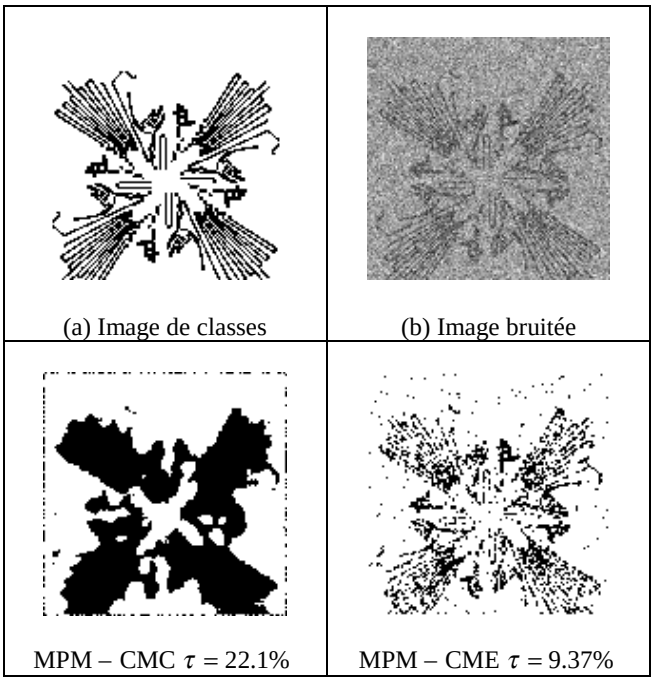


FIG. 1 : segmentation non supervisée

6. Conclusion

Dans cet article nous avons étudié la possibilité d' utiliser la théorie de l' évidence dans le cadre des champs de Markov cachés (CMC). A cet effet, nous avons proposé un modèle d'affaiblissement, aboutissant à un champ de Markov triplet, ainsi qu'une méthode d'estimation de paramètres proche de celle décrite dans [3]. L' intérêt de l' affaiblissement de la loi *priori* d'un champs de Markov caché en segmentation non supervisée d'images non stationnaires a également été validé par le biais d'un exemple.

Comme perspectives, nous pouvons envisager l' extension de ce modèle aux champs de Markov couple, ou l' utilisation d' un bruitage plus complexe. L'extension au casmultivarié peut également être envisagée.

Références

[1] A. Appriou, Probabilités et incertitude en fusion de données multisenseurs, *Revue Scientifique et Technique de la Défense*, No. 11, pp. 27-40, 1991.

[2] A. Bendjebbour, Y. Delignon, L. Fouque, V. Samson, and W. Pieczynski, Multisensor images segmentation using Dempster-Shafer fusion in Markov fields context, *IEEE Trans. on Geoscience and Remote Sensing*, Vol. 39, No. 8, pp. 1789-1798, 2001.

[3] D. Benboudjema and W. Pieczynski, Parameter estimation in pairwise Markov fields, *Advanced Concepts for Intelligent Vision Systems (ACIVS 04)*, Aug. 31-Sept. 3, Brussels, Belgium, 2004.

[4] D. Benboudjema and W. Pieczynski, Segmenting non stationary images with triplet Markov fields, International Conference on Image Processing (ICIP 2005), Genova, Italy, September 11-14, 2005.

[5] D. Benboudjema and W. Pieczynski, Unsupervised image segmentation using triplet Markov fields, *Computer Vision and Image Understanding*, to appear, 2005.

[6] S. Foucher, M. Germain, J.-M. Boucher, and G. B. Benié, Multisource classification using ICM and Dempster–Shafer theory, *IEEE Trans. on Instrumentation and Measurement*, Vol. 51, No. 2, pp. 277-281, 2002.

[7] S. Geman and D. Geman, *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images*, IEEE Trans. On PAMI, vol. 6, pp. 721-741, 1984.

[8] P. Lanchantin et W. Pieczynski, Chaînes et arbres de Markov évidentiels avec applications à la segmentation des processus non stationnaires, *Traitement du Signal*, Vol. 22, No. 1, pp. 15-26, 2005.

[9] P. Lanchantin and W. Pieczynski, Unsupervised restoration of hidden non stationary Markov chain using evidential priors, *IEEE Trans. on Signal Processing*, to appear, 2005.

[10] D. Melas and S. P. Wilson, Double Markov Random Fields and Bayesian Image Segmentation, *IEEE Trans. on Signal Processing*, Vol. 50, No. 2, pp. 357-365, 2002.

[11] P. Pérez. Markov random fields and images. *CWI Quarterly*, Vol. 11, No. 4, pp. 413-437, 1998.

[12] W. Pieczynski and D. Benboudjema, Triplet Markov fields and theory of evidence, submitted to *Image and Vision Computing Journal*, October 2003.

[13] W. Pieczynski, D. Benboudjema, and P. Lanchantin, Statistical image segmentation using Triplet Markov Fields, *SPIE's International Symposium on Remote Sensing*, September 22-27, Crete, Greece, 2002.

[14] W. Pieczynski and A.-N. Tebbache, Pairwise Markov random fields and segmentation of textured images, *Machine Graphics and Vision*, Vol. 9, No. 3, pp. 705-718, 2000.

[15] W. Pieczynski, Modèles de Markov en traitement d' images, *Traitement du Signal*, Vol. 20, No. 3, pp. 255-278, 2003.

[16] A. Sankar, A. Banerjee, N. Banerjee, S. Brahma, B. Kartikeyan, M. Chakraborty, and K. L. Majumder, Landcover classification in MRF context using Dempster-Shafer fusion for multisensor imagery, *IEEE Trans. on Image Processing*, Vol. 14, No. 5, pp. 634-645, 2005.

[17] G. Shafer, A mathematical theory of evidence, *Princeton University Press*, Princeton, 1976.

[18] P. Smets, Belief functions versus probability functions, In B. Bouchon, L. Saitta, and R. R. Yager (eds), Uncertainty and intelligent systems, *Berlin, Springer-Verlag*, pp. 17-24, 1988.

[19] F. Tupin, H. Maitre, and I. Bloch, A first step toward automatic interpretation of SAR images using evidential fusion of several structure detectors, *IEEE Trans. on GRS*, Vol. 37, No. 3, pp. 1327-1343, 1999.

[20] G. Winkler, *Image analysis, random fields and and Markov Chain Monte Carlo Methods: a mathematical introduction*, Springer, 2003.

Fusion de segmentations d'images microscopiques par SVM dans différents espaces couleur selon la théorie de l'évidence

Christophe CHARRIER, Gilles LEBRUN, Olivier LÉZORAY

Université de Caen-Basse Normandie, LUSAC EA-2607, Groupe Vision et Analyse d'Images
120, route de l'exode, 50000 Saint-Lô, France
{c.charrier,g.lebrun,o.lezoray}@chbg.unicaen.fr

Résumé – Dans cet article, un outil d'aide au diagnostic par segmentation automatique d'images microscopiques cellulaires est proposé. Cet outil repose sur la mise en place d'une méthode de segmentation automatique combinant (selon la théorie de l'évidence) les résultats issus des différentes segmentations par Machines à Support de Vecteurs–*Support Vector Machines (SVM)* –en intégrant des incertitudes et des redondances liés aux espaces couleur. Ces incertitudes sont obtenues par calcul d'une probabilité *a posteriori* qui est fonction des résultats générés par les SVMs. Une amélioration sensible de la qualité de la segmentation est finalement réalisée par la prise en compte de l'incohérence de plusieurs classifications des pixels.

Abstract – In this paper, a tool for diagnosis assistance by automatic segmentation of microscopic cellular images is introduced. This method is based on an automatic segmentation technique combining (with the Dempster-Shafer rule) the results obtained by Support Vector Machines (SVM) applied within different color spaces. This combination is performed by integrating uncertainties and redundancies for each one of the color spaces. These uncertainties are computed as *a posteriori* probabilities according to the SVM obtained results. An improvement of the final segmentation quality is performed by taking into account the inconsistencies of several pixel classification.

1 Introduction

L'analyse d'images microscopiques dans les laboratoires d'anatomie et de cytologie pathologiques tient une place très importante pour la détection des cellules anormales afin d'établir un diagnostic. Les cellules sont visuellement évaluées par un technicien lors d'une phase de screening. Cette analyse est d'un intérêt capital pour le pathologiste car son diagnostic dépend de la bonne reconnaissance des cellules anormales ou suspectes. Or le screening visuel est difficile et le faible nombre de cellules anormales comparativement aux nombre de cellules normales implique une grande concentration du technicien. A cause de ce facteur subjectif des faux-négatifs peuvent apparaître. Pour pallier ceci, une approche consiste à aider le technicien par un système de screening semi-automatique afin de confirmer l'analyse visuelle. Dans cette optique, nous désirons concevoir une méthode de segmentation des cellules d'images de microscopie, ces cellules étant constituées d'un cytoplasme et d'un noyau. Cependant leur configuration spatiale et colorimétrique est extrêmement variable, ce qui rend difficile leur extraction précise et une méthode de segmentation peut bien fonctionner sur certaines images et moins bien sur d'autres.

En général, un seul schéma de segmentation est appliqué afin d'obtenir des segmentations les plus fidèles possibles par rapport aux vérités terrains disponibles. Or améliorer les taux de classification, et donc la qualité de la segmentation induit inéluctablement la définition et le développement de protocoles de classification de plus en plus sophistiqués, et par la même de plus en plus sensibles à la moindre variation de la structure de l'image.

Une autre voie peut cependant être envisagée : celle de la combinaison des résultats obtenus par différents schémas. Ainsi nous proposons d'utiliser différents schémas de segmentation par classification de pixels dans différents espaces couleur afin d'obtenir plusieurs segmentations possibles d'une même image en trois classes : fond, cytoplasme, noyau. La classification est ainsi opérée séparément dans cinq espaces couleur à savoir RVB, $L^*a^*b^*$, YUV et deux systèmes de représentation colorimétrique antagonistes YC_bC_r et HSL. Ces différentes segmentations sont ensuite fusionnées afin d'obtenir une unique carte de segmentation.

2 Création et fusion des cartes de segmentation

La figure 1 illustre le synopsis du schéma de segmentation proposé. Une première étape de segmentation par classification des pixels est réalisée dans les cinq espaces couleur retenus. A l'issue de cette phase, et pour chacune des segmentations, une probabilité *a posteriori* p_i est calculée, ce qui permet d'obtenir un indice de confiance de la classification pour chaque pixel de l'image par rapport aux classes finales. Puis, dès lors qu'il y a divergence de classification pour un pixel, une étape de fusion est opérée en prenant en compte la probabilité *a posteriori*. Une carte de segmentation finale est ainsi générée.

2.1 Segmentation par classification des pixels

Parmi toutes les techniques de segmentation qui sont offertes, les segmentations par les SVM ont été retenues car une précédente étude a permis de mettre en exergue des niveaux de classification très satisfaisants [1].

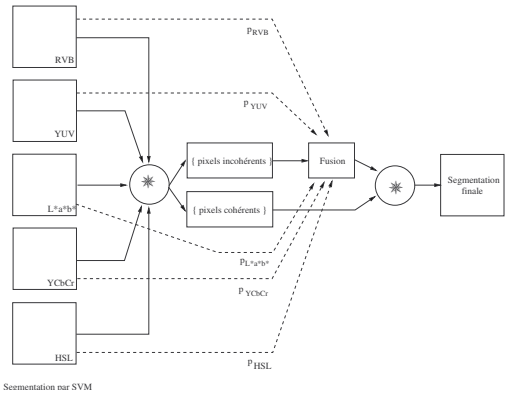


FIG. 1: *Synopsis du schéma proposé ou p_i représente la probabilité* a posteriori *calculée dans l'espace couleur i.*

2.1.1 Principe des SVM

La théorie de l'Apprentissage Statistique de Vapnik et de Chervonenkis [2] a conduit au développement d'une classe d'algorithme connu sous le nom de SVM. Ils permettent de réaliser des estimations en classification. Une des originalités de la méthode est de produire une fonction de décision qui n'utilise qu'un sous-ensemble de la base d'apprentissage. Les éléments de ce sous-ensemble sont nommés Vecteurs de Support ($\mathbf{SV}$).

Soit un ensemble d'apprentissage $A = \{(x_1, y_1), \ldots, (x_k, y_k)\}$ composé de k couples avec $x_i \in \mathcal{R}^n$ et $y_i \in \{-1, +1\}$. L'algorithme des SVMs projette les vecteurs x_i dans un espace de travail $\mathbf{H}$ à partir d'une fonction non linéaire $\phi : \mathcal{R}^n \rightarrow \mathbf{H}$. L'hyperplan optimal de séparation des deux classes dans l'espace $\mathbf{H}$ est ensuite recherché. Cet hyperplan $(\mathbf{w}, b)$ matérialise la frontière de séparation entre les deux classes. La classe y d'un nouvel exemple x est définie par $y = sign\,(\mathbf{w} \cdot \phi(x) + b)$. L'hyperplan est optimal s'il maximise la distance qui le sépare des exemples dont il est le plus proche. Cette distance est usuellement appelée marge du classifieur. Il a été démontré [2] que maximiser cette marge correspond à maximiser le "pouvoir" généralisateur du classifieur. En choisissant une fonction noyau $K(x_i, x_j) =< \phi(x_i), \phi(x_j) >$, il a également été montré que la résolution du problème dual correspondant produit une fonction de décision de la forme $f(x) = \sum_{x_i \in SV} \alpha_i y_i K(x_i, x) + b$ où $\{\alpha_i\}$ et b correspondent à la solution optimale du problème dual.

2.1.2 Calcul et utilisation des probabilités a postériori

Les SVMs étant des classifieurs binaires, la résolution d'un problème multi-classes est effectuée en le transformant en une combinaison de problèmes binaires [3]. Ainsi, dans notre cadre applicatif, trois fonctions de décision seront utilisées pour réaliser la classification en trois classes {noyau, fond, cytoplasme} correspondant à leurs discriminations en un contre tous. Cependant les SVMs ne permettent pas d'aboutir directement à des probabilités *a posteriori* de classification. Au lieu d'estimer les densités conditionnelles des classes $p(f|y)$, un modèle paramétrique est utilisé afin d'obtenir les probabilités *a posteriori* $p(y = 1|f)$ directement où f représente la valeur obtenue à la sortie des SVMs. PLATT [4] propose une méthode de calcul de ces probabilités à partir des paramètres obtenus par les SVMs. Ainsi les probabilités *a posteriori* sont déterminées à l'aide d'une fonction paramétrique sigmoïde par :

$$p(y = 1|f) = \frac{1}{1 + \exp(Bf + C)}, \tag{1}$$

dans laquelle les paramètres B et C sont estimés en utilisant la procédure du maximum de vraisemblance.

Un ensemble d'images segmentées constituant une vérité terrain des classes de pixels sert de base d'apprentissage. Chacune des trois fonctions de décision réalise un apprentissage par espace couleur, ce qui permet d'obtenir cinq cartes de segmentation distinctes avec pour chaque pixel de chaque carte une probabilité d'appartenance *a posteriori* p_i pour chaque classe. Cette probabilité correspond à un indice de confiance pour chaque classe et pour chaque espace couleur. L'objectif est ensuite de fusionner ces résultats afin d'aboutir à une classification finale, et donc une segmentation en trois classes. Parmi tous les outils qui s'offrent à nous dans ce domaine, nous décidons d'utiliser la théorie de l'évidence (aussi appelée théorie de Dempster-Shafer ou théorie des fonctions de croyance) [5], qui permet d'une part de traiter des informations incertaines, et d'autre part de combiner des informations provenant de plusieurs sources. Dans le cadre de cette théorie, plusieurs règles de décision sont définies pour nous permettre de sélectionner la classe à laquelle le pixel sera finalement affecté.

Afin de procéder à la fusion, une nouvelle carte de segmentation est générée par intersection des cartes obtenues dans les cinq espaces couleur. Seuls les pixels présentant des incohérences de classification seront traités par la théorie de l'évidence. Un "pixel incohérent" correspond à un pixel classé dans plus d'une classe lorsque l'union de toutes les cartes est réalisée.

2.2 Traitement des pixels incohérents

2.2.1 Elements de la théorie de l'évidence

Soit $\Omega = \{\omega_1, \ldots, \omega_N\}$, l'ensemble des N classes finales possibles pour un pixel incohérent, appelé cadre de discernement. Dans notre cadre applicatif $N = 3$ et Ω correspond aux trois classes $\{\omega_F, \omega_C, \omega_N\}$, respectivement pour le fond de l'image, le cytoplasme et le noyau. Au lieu de restreindre ses mesures à cet ensemble (comme le ferait la théorie des probabilités contrainte par son axiome d'additivité), la théorie de l'évidence s'étend sur l'ensemble puissance de Ω, noté 2^Ω, l'ensemble des 2^N sous-ensembles de Ω. On définit alors une fonction de masse m qui représente la croyance que l'on accorde aux différents états du système, à un instant donné. Cette fonction est également connue en tant que masse de croyance élémentaire. Elle est définie de 2^Ω dans $[0, 1]$ et vérifie

$$\sum_{A \subseteq \Omega} m(A) = 1 \text{ et } m(\emptyset) = 0 \tag{2}$$

où $m(A)$ représente la croyance que l'on place dans la proposition A. Il est à noter que A représente soit un singleton ω_n, soit une disjonction d'hypothèses. Ici, $m(A)$ quantifie la croyance que l'on accorde au fait que la classe cherchée appartienne au sous-ensemble A de Ω (et à aucun autre sous-ensemble de A). Les propositions A telles que $m(A) > 0$ sont appelées éléments focaux. De plus, dans le cas où l'ensemble des hypothèses est exhaustif et exclusif, la masse de l'ensemble vide $\emptyset$ est nulle.

Deux fonctions de masse initiales m_1 et m_2 représentant les informations respectives de deux sources différentes, peuvent être combinées selon la règle de Dempster [6] :

$$m(A) = \frac{\sum_{B\cap C=A} m_1(B)m_2(C)}{1-K}, \quad \forall A \in \Omega \quad A \neq \emptyset. \tag{3}$$

K est appelé le facteur de conflit et représente le désaccord entre les deux sources. Il correspond à la masse de l'ensemble vide $K = \sum_{B\cap C=\emptyset} m_1(B)m_2(C)$. Notons que la combinaison de Demspter, appelée aussi somme orthogonale et notée $m = m_1 \oplus m_2$ est commutative et associative.

Après combinaison, il reste à prendre une décision quant à l'élément le plus "probable" de Ω. Plusieurs règles de décision sont alors possibles, mais une des plus utilisées reste celle du maximum de probabilité "pignistique". Cette règle de décision présentée par SMETS [7] utilise la transformation pignistique, qui permet de transformer une masse m_i en une distribution de probabilité permettant la prise de décision telle que :

$$\omega^* = \text{Arg}\left\{\max_{\omega\in\Omega}\left[\sum_{\omega\in A\subseteq\Omega}\frac{m(A)}{|A|}\right]\right\}. \tag{4}$$

$|A|$ étant le cardinal de A.

2.2.2 Modélisation des fonctions de masse

L'une des difficultés réside dans le choix de la modélisation des fonctions de masse. Parmi les modélisations existantes, celle proposée par DENŒUX [8] a été utilisée car elle est basée sur la notion de distance aux voisins et permet d'intégrer divers critères de voisinage. Ainsi la masse $m(\{\omega_n\})$ est définie comme une fonction décroissante de la distance d entre le pixel à classifier et le barycentre de la classe :

$$\begin{cases} m(\omega_n) = \alpha \exp\left(-\gamma_n d^2\right) \\ m(\Omega) = 1 - m(\omega_n) \end{cases} \tag{5}$$

où $0 < \alpha < 1$ est une constante. Dans cette étude, α est déterminé à partir des probabilités *a posteriori* obtenues par les SVM pour la classe ω_n dans les différents espaces couleur : $\alpha = p_i.\ \gamma_n > 0$ dépend de la classe ω_n et est calculé par minimisation d'un critère d'erreur selon l'algorithme EM utilisé dans [9].

Les 5 fonctions de masses initiales $(m_i)\ i \in [1,\ldots,5]$ sont donc créées après calcul des k régions candidates à la fusion et avant décision du vote majoritaire. Ainsi pour une même carte de segmentation, un pixel qui a été initialement affecté à une seule des trois classes $\{\omega_F, \omega_C, \omega_N\}$ peut, en relâchant la contrainte d'affectation, être affecté à une classe parmi au plus les trois classes. La contrainte correspond simplement à la création d'un intervalle calculé à partir de la valeur maximale de la probabilité *a posteriori* p_i tel que $[p_i, p_i - \delta_i]$, avec δ_i fonction de l'espace couleur associé. Ainsi toutes les classes dont les probabilités *a posteriori* sont incluses dans cet intervalle sont considérées comme des classes candidates à l'affectation du pixel.

Les fonctions de masses permettent de prendre en compte une incertitude associée à chacune des cartes de segmentation. Ainsi, on regroupe dans le même élément focal les classes qui se trouvent très proches les unes des autres, et la prise de décision ne s'effectue qu'après avoir combiné les résultats des autres projections.

Dans cette étude deux formulations de la distance d ont été étudiées. Il s'agit 1) de la distance euclidienne, 2) de la distance MAHALANOBIS permettant de comparer la dispersion spatio-colorimétrique de deux nuages couleurs, définie par

$$d = \sqrt{(x-\mu_i)^{\mathrm{T}}\textstyle\sum_i^{-1}(x-\mu_i)} \tag{6}$$

où μ_i représente le vecteur moyenne et $\sum_i^{-1}$ la matrice de covariance inverse associée à l'hypothèse ω_n.

Lors de l'utilisation de la distance euclidienne seule, la formulation de la distance euclidienne pondérée préconisée en 1994 par la Commission Internationale de l'Eclairage (CIE) lors de l'utilisation de l'espace couleur $L^*a^*b^*$ a été utilisée. Même si l'expression du calcul de la distance entre couleurs doit s'exprimer dans un espace couleur autre que l'espace RVB, ces formules ne sont valables que pour des petites distances et leurs utilisations en dehors de ce contexte n'est pas validée par la CIE.

3 Résultats expérimentaux

3.1 Evaluation de la qualité de la segmentation

L'évaluation de la qualité d'un traitement est un domaine de recherche mobilisant de nombreuses équipes. Que les champs d'investigations concernent la compression ou la segmentation des images couleur, deux classes se distinguent en fonction de la présence d'une image de référence ou non. Dans le cas de la segmentation, le terme "vérité terrain" est utilisé en lieu et place d'image de référence.

Lorsque l'on dispose d'une vérité terrain, les mesures de qualité sont toutes basées sur la prise en compte de la taille des régions et du taux de recouvrement entre les régions R_i de l'image segmentée I et les régions V_j de l'image de référence V. Parmi toutes les métriques disponibles, la mesure définie par MARTIN a été retenue puisqu'elle permet de mesurer la dissimilarité entre la segmentation résultat et la segmentation de référence.

Cette mesure est basée sur l'erreur $E(s)$ calculée en chaque pixel s par :

$$E(s) = \frac{\text{card}(V_j \backslash R_i)}{\text{card}(V_j)}. \tag{7}$$

ou

$$E'(s) = \frac{\text{card}(R_i \backslash V_j)}{\text{card}(R_i)}. \tag{8}$$

La dissimilarité est ensuite fournie par l'erreur locale de cohérence fournissant une mesure de qualité :

$$\text{ELC}(I,V) = \frac{1}{h\times l}\sum_s \min\{E(s), E'(s)\} \tag{9}$$

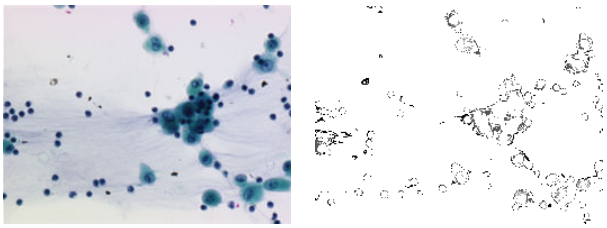
où h et l représentent respectivement la hauteur et la largeur en pixles de l'image V.

Il existe une erreur globale de cohérence qui n'a pas été retenue ici car cette dernière a le défaut de favoriser une sur-segmentation ou une sous-segmentation [10].

3.2 Résultats

La technique proposée est appliquée sur une base d'images médicales composée d'images microscopiques cellulaires. La

figure 2 présente les pixels inchoérents obtenus après intersection des cartes de segmentation dans les cinq espaces couleur (figure 2(b)). Les désaccords interviennent principalement au niveau du cytoplasme et des noyaux lorsque l'on compare les résultats avec l'image initiale (figure 2(a)).



(a) (b)

FIG. 2: *Pixels incohérents obtenus par intersections des diverses cartes de segmentation en fonction d'une image originale.*

Le tableau 1 présente les moyennes en pourcentages des pixels incorrectement et correctement classifiés ainsi que la moyenne en pourcentage des pixels présentant une incohérence de classification. On note qu'il subsiste un pourcentage de pixels (9.4%) pour lesquels il n'est plus possible de revenir sur la classification puisque ces derniers ont été associés à une classe erronée de manière unanime (selon la comparaison avec les images de vérité terrain). Ainsi, le taux de classification maximal que l'on puisse atteindre est fonction de ces affectations. Seuls les pixels pour lesquels subsistent un désaccord entre les classifieurs pourront subir une révision de classification, soit de 3.47% au maximum. Ainsi pour améliorer le schéma de classification, il est également nécessaire d'améliorer les pré-traitements, à savoir, les résultats obtenus par les SVMs.

Moyenne des classification par SVMs dans les différents espaces couleur		
correcte	incorrecte	incohérente
87.13 ± 6.12	9.4 ± 5.32	3.47 ± 1.13

TAB. 1: *Pourcentage des pixels classifiés correctement, incorrectement selon les vérités terrain.*

Le tableau 2 présente la moyenne des taux de classification correctes obtenus sur la base pour les SVMs et la méthode proposée ainsi que la mesure de qualité de la segmentation finale.

Méthode	moyenne TR	Moyenne qualité
SVM	87.13 ± 6.12	0.45 ± 0.04
SVM combinés (d_1)	87.67 ± 5.62	0.39 ± 0.03
SVM combinés (d_2)	88.12 ± 5.43	0.38 ± 0.04

TAB. 2: *Moyennes du taux de reconnaissance et de qualité selon le critère de Martin pour la segmentation par les SVMs, la combinaison des SVMs intégrant la distance euclidienne (d_1) et la distance de Mahalanobis (d_2).*

Les résultats obtenus démontrent que la combinaison des probabilités des pixels incohérents permet d'améliorer sensiblement la qualité de la segmentation basée sur un processus de classification des pixels. On note effectivement une augmentation du pourcentage de reconnaissance pour chacune des trois classes. L'incohérence se situe principalement sur les pixels appartenant au cytoplasme et aux noyaux. Le gain de qualité se portant essentiellement sur la classification de ces pixels, la qualité de la segmentation d'un point de vue cytologique est grandement améliorée.

4 Conclusion

Dans cet article, un outil d'aide au diagnostic d'images microscopiques cellulaires est présenté. L'algorithme sous-jacent repose sur une combinaison de segmentations réalisées dans divers espaces colorimétriques. A partir des cartes de segmentation ainsi obtenues, une probabilité *a posteriori* est calculée et permet ainsi de quantifier le degré de confiance que l'on accorde à la segmentation. Une combinaison, selon la théorie de Dempster-Shafer, est ensuite réaliser de manière à prendre en compte les incertitudes liées à chacune des classifications dans les divers espaces couleur. Les résultats permettent de montrer une amélioration des taux de classification.

Références

[1] G. Lebrun, C. Charrier, O. Lezoray, C. Meurie, and H. Cardot, "Fast pixel classification by SVM using vector quantization, tabu search and hybrid color space," in *the 11th International Conference on CAIP*, (Rocquencourt, France), 2005. to appear.

[2] V. N. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[3] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 3, pp. 415–425, 2002.

[4] J. Platt, "Probabilities for SV machines," in *Advances in Large-Margin Classifiers* (D. S. Peter J. Bartlett, Bernhard Schölkopf and A. J. Smola, eds.), Neural Information Processing Systems, MIT Press, 2000.

[5] G. Shafer, *A mathematical theory of evidence*. Princeton University Press, 1976.

[6] A. Dempster, "Upper and Lower Probablilities Induced by Multivalued Mapping," *Ann. Math. Statist.*, vol. 38, pp. 325–339, 1967.

[7] P. Smets, "Constructing the pignistic probability function in a context of uncertainty," *Uncertainty in Artificial Intelligence*, vol. 5, pp. 29–39, 1990. Elsevier Science Publishers.

[8] T. Denoeux, "A k-nearest neighbor classification rule based on dempster-shafer theory," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 25, no. 5, pp. 804–813, 1995.

[9] L. M. Zouhal and T. Denœux, "An evidence-theoritic k-NN rule with parameter optimization," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 28, pp. 263–271, 1998.

[10] S. Philipp-Foliguet, "Evaluation de méthodes de segmentation d'images couleur," in *Actes de l'école de printemps du GdR-ISIS, Opération Imagerie Couleur*, (Pau, France), 2001.

Fusion de critères pour l'évaluation de résultats de segmentation d'images

Sébastien CHABRIER[1], Hélène LAURENT[1], Christophe ROSENBERGER[1], Alain RAKOTOMAMONJY[2]

[1]Laboratoire Vision et Robotique - UPRES EA 2078
ENSI de Bourges - Université d'Orléans
10, Bd Lahitolle, 18020 Bourges Cedex, France
Tél : 02-48-48-40-64, Fax : 02-48-48-40-40

[2]Perception, Systemes, Information - FRE CNRS 2645
INSA de Rouen, Avenue de l'Université
76801 Saint Etienne du Rouvray, France
Tél : 02-32-95-97-03 Fax: 02-32-95-97-08
sebastien.chabrier@ensi-bourges.fr

Résumé – L'évaluation de résultats de segmentation est une tâche très difficile, et ce même pour un expert. L'objet de cet article est de proposer une mesure de qualité d'un résultat de segmentation, sans aucune connaissance *a priori*, en prenant en compte différents critères d'évaluation de la littérature. Nous proposons dans ce papier de fusionner les meilleurs d'entre eux par deux méthodes : une combinaison linéaire optimisée par un algorithme génétique et une méthode d'apprentissage reposant sur un algorithme de type Support Vector Machine. La mesure de Vinet est utilisée comme mesure objective pour comparer le comportement des deux approches envisagées. Nous présentons enfin les résultats que nous avons obtenus.

Abstract – Segmentation evaluation is a very difficult task even for an expert. The aim of this article is to propose a quality measure of a segmentation result, without any *a priori* knowledge, by taking into account different evaluation criteria from the literature. We propose in this paper to fusion the best ones by using two methods: a linear combination using a genetic algorithm and a learning method using a support vector machine algorithm. Vinet's measure is used as an objective measure to compare the behavior of the two proposed approaches. Some results are presented in the last section.

1 Introduction

La segmentation est une étape essentielle en traitement d'images dans la mesure où elle conditionne l'interprétation d'une image. De nombreux algorithmes ont ainsi été proposés durant les dernières décennies ([3], [5],...). Ils sont basés sur différentes approches : contour, région, texture, ... Devant la multitude des méthodes proposées, l'évaluation de la qualité d'un résultat de segmentation devient primordiale. Habituellement, l'efficacité d'un nouvel algorithme est illustrée par la présentation de quelques résultats de segmentation, ce qui n'autorise que des conclusions subjectives et qualitatives sur les performances de cet algorithme. Un critère d'évaluation de la qualité d'un résultat de segmentation peut alors présenter de nombreux intérêts. Il peut, par exemple, permettre de faciliter le choix du paramétrage d'une méthode de segmentation, ou encore autoriser la comparaison de différentes méthodes de segmentation de même niveau de précision, voire même, de définir de nouvelles méthodes de segmentation (par optimisation du critère).

De nombreux travaux traitent de ce sujet et plusieurs classes de critères d'évaluation ont été isolées ([9], [14]). Nous allons nous intéresser aux critères d'évaluation non supervisée. Ces critères ne nécessitent aucune connaissance *a priori* sur les résultats de segmentation à évaluer. Leur principe consiste à estimer la qualité d'un résultat de segmentation à partir de statistiques calculées sur chaque région, contour ou texture détectés. Nous avons orienté notre choix vers cette classe de critères par souci de généricité de l'évaluation.

Actuellement, aucun critère d'évaluation de segmentation ne se révèle satisfaisant dans tous les cas de figure. Comme le montre Brendan McCane [2], pour évaluer correctement des résultats de segmentation, il est pertinent d'utiliser le maximum de critères d'évaluation à notre disposition et de combiner leurs résultats.

Dans la première partie de cet article, nous présentons une brève définition de ce qu'est un "bon" résultat de segmentation d'une image, puis dans un deuxième temps, le choix des critères que nous allons fusionner. La troisième partie détaille les processus de fusion mis en place. Dans la quatrième partie, nous présentons les résultats obtenus après fusion.

2 Définition de la segmentation

Différentes définitions de la segmentation d'une image peuvent être trouvées dans la littérature. Haralick [5], Zhang [15] et Freixenet [4] résument bien les définitions existantes. D'un point de vue général, on retrouve le fait que l'image doit être décomposée en éléments (régions ou classes) et que le calcul de propriétés à l'intérieur de chacun de ces éléments doit offrir une forte stabilité et une

disparité maximale avec celle des éléments voisins.

Ces définitions mettent en avant le fait qu'un résultat de segmentation n'est pas dépendant d'une application donnée et qu'il est défini par des propriétés calculées sur les différentes régions d'une image. Ces remarques montrent qu'il est alors possible de comparer la qualité de résultats de segmentation sans aucune connaissance *a priori* à un même niveau de précision donné. En effet, cela n'aurait aucun sens de comparer des résultats de segmentation avec des niveaux de précision complètement différents.

3 Définition de nouveaux critères

Avant de fusionner différents critères d'évaluation, nous avons voulu déterminer les critères les plus pertinents. Pour ce faire, nous avons utilisé une base (BIS) de 300 images synthétiques composées de régions uniformes et texturées. Un exemple de création de telles images est présenté figure Fig. 1. Ces images ont été segmentées par 4 méthodes de segmentation et procurent donc 1200 résultats de segmentation. La mesure de Vinet [12] (taux de classification correcte) est le critère d'évaluation supervisée employé comme référence sur ces images synthétiques.

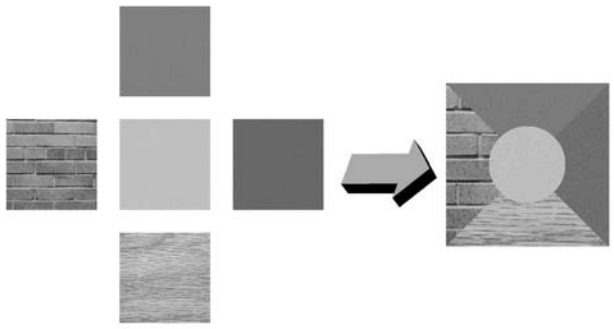


FIG. 1: Exemple de création d'une image mixte comportant 3 régions uniformes bruitées et 2 texturées pour la création de la base BIS.

Suite à l'étude comparative de critères d'évaluation non supervisée de résultats de segmentation présentée dans [8], les critères qui suivent le mieux l'évolution de la mesure de Vinet et que nous avons retenus pour la fusion sont : le critère de Borsotti [1], le critère de Zéboudj [13] et le critère de Rosenberger [11].

Les critères de Zéboudj et Borsotti donnent de bons résultats dans le cas d'images uniformes, et des résultats moyens dans le cas mixte (régions texturées et uniformes). Le critère de Rosenberger, quant à lui, donne de bons résultats sur les images texturées, mais se montre un peu plus faible sur les images uniformes que les critères de Zéboudj et de Borsotti.

Notre objectif est alors de fusionner ces trois critères afin d'obtenir un comportement plus proche de la mesure de Vinet. Deux approches sont envisagées : une fusion par combinaison linéaire des critères et une fusion par apprentissage. Ces approches permettra à terme d'évaluer le résultat de segmentation d'une image réelle en fusionnant les critères d'évaluation les plus pertinents obtenus à partir de la base d'images synthétiques.

3.1 Principe de la fusion par combinaison linéaire

Cette première approche consiste à effectuer une combinaison linéaire des trois critères que nous avons retenus lors de l'étude comparative. Nous allons alors définir un nouveau critère C_{Fus} s'écrivant comme la combinaison linéaire des trois précédents :

$$C_{Fus} = a * Zéboudj + b * Borsotti + c * Rosenberger \quad (1)$$

Afin d'identifier la similarité de comportement entre ce nouveau critère et la mesure de Vinet, nous allons utiliser le facteur de corrélation. Notre objectif sera donc de déterminer les paramètres optimaux (a, b, c) tels que la valeur absolue du facteur de corrélation entre les résultats de C_{Fus} et ceux de la mesure de Vinet sur la base BIS soit maximale. On posera comme contrainte $(a, b, c) \in [-1, 1]^3$. Ce choix est arbitraire car n'importe quel espace de recherche peut être utilisé : dans un autre espace de recherche, le vecteur résultat (a, b, c) serait le même à une constante multiplicative près, et le classement des résultats serait exactement le même. Cette combinaison induit alors un nouveau critère non normalisé s'écrivant comme un barycentre des 3 critères.

Ces critères n'étant pas exprimables simplement en une formule car représentant de nombreux calculs sur toute une image, il est difficile d'utiliser des méthodes d'optimisation telles que le simplexe pour obtenir le résultat optimal. Nous avons opté pour un outil ayant fait ses preuves dans un tel domaine d'application, à savoir les algorithmes génétiques [6]. Les algorithmes génétiques déterminent les solutions d'une fonction en simulant l'évolution d'une population jusqu'à la survie des individus les plus adaptés. Les survivants sont des individus obtenus par croisement, mutation et sélection d'individus de la génération précédente. Un algorithme génétique est défini en considérant cinq données essentielles :

1. la représentation de la solution ou génotype : un vecteur de paramètres (a, b, c),
2. la population initiale : un ensemble de vecteurs de paramètres (a, b, c),
3. la fonction d'aptitude : le facteur de corrélation entre C_{Fus} de l'équation (1) (calculé sur toute la base d'images avec un vecteur de paramètres donné) et les résultats de la mesure de Vinet sur cette même base,
4. les opérations sur les génotypes : mutation, sélection ou croisement des vecteurs de paramètres,
5. critère d'arrêt de l'évolution de la population : écart type de la fonction d'aptitude.

Une fois définies ces cinq données, l'exécution d'un algorithme génétique s'effectue en cinq étapes :

1. définition de la population initiale et calcul de l'aptitude de chaque individu,

2. sélection et mutation des individus de la population courante,
3. reproduction des individus restant de la population courante,
4. évaluation des individus de la population,
5. retour à l'étape 2 si le critère d'arrêt n'est pas satisfait.

Les coefficients (a, b, c) optimisant le facteur de corrélation entre C_{Fus} et la mesure de Vinet sont alors déterminés en utilisant cette méthode [6].

3.2 Principe de la fusion par apprentissage

Cette seconde approche consiste en la fusion de plusieurs critères par apprentissage. Soit un ensemble de couples $\{x_i, y_i\}_{i=1..\ell}$ avec $x_i \in \mathbb{R}^d$ un vecteur de d critères décrivant la qualité d'un résultat de segmentation d'une image donnée et y_i un indice de qualité du résultat de segmentation (dans notre cas la valeur de la mesure de Vinet discrétisée entre 1 et 10). Notre objectif est alors de construire à partir de la base d'apprentissage $\{x_i, y_i\}_{i=1..\ell}$ une fonction f capable de prédire l'indice de qualité d'un résultat de segmentation d'une nouvelle image x. Ainsi, notre idée est d'utiliser un apprentissage supervisé pour atteindre cet objectif mais aussi d'utiliser ce contexte pour fusionner différents critères et sélectionner les plus performants.

Ce problème d'apprentissage supervisé est un problème multiclasses avec un nombre de classes dépendant de la quantité d'indices de qualité (discrétisation en 10 classes de la mesure de Vinet pour ce qui concerne cette étude). Un problème multiclasse est résolu par une polychotomie basée sur une approche *un-contre-un* [7]. Notre algorithme d'apprentissage correspond donc à un Séparateur à Vaste Marge pour 2 classes. Ainsi, nous recherchons un hyperplan dans un espace $\mathcal{H}$ défini comme : $f(x) = \sum_{i=1}^{\ell} \alpha_i^\star y_i K(x_i, x) + b$ qui maximise la marge entre l'hyperplan et les données projetées x_i dans $\mathcal{H}$. Alors, les $\alpha_i^\star$ sont les solutions du problème d'optimisation suivant :

$$\begin{array}{l} \max_{\alpha_i} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j (K(x_i, x_j) + \frac{1}{C} \delta_{i,j}) \\ \text{avec} \sum_i \alpha_i y_i = 0, \quad 0 \leq \alpha_i \end{array} \quad (2)$$

où K est le noyau associé à $\mathcal{H}$, $\delta_{i,j}$ est le symbole de kronecker et C un paramètre de compromis entre la largeur de la marge et le nombre d'exemples situés au delà de la marge.

Il est à noter que tous les critères d'évaluation de résultats de segmentation sont implicitement fusionnés par le noyau K. En fait, grâce aux propriétés de $\mathcal{H}$, la fonction de décision peut aussi être écrite : $f(x) = \sum_i \alpha_i y_i K(x_i, x) + b$ ce qui devient $f(x) = \langle \sum_i \alpha_i y_i x_i, x \rangle + b$. Ainsi, chaque critère d'évaluation u d'un résultat de segmentation x est pondéré par $\sum_i \alpha_i y_i x_i^{(u)}$ dans la fonction de décision.

Un tel apprentissage combine implicitement plusieurs critères d'évaluation. Or, il est possible que seul un sous-ensemble de ces critères soit pertinent pour identifier l'indice de qualité du résultat de segmentation. Afin d'identifier ce sous-ensemble, nous avons réalisé une sélection de critères. Cette sélection de critères repose sur l'influence de chaque critère sur la marge dans la classification [10]. Ainsi, chaque critère a été pondéré par un facteur d'échelle σ et la sensibilité de la marge par rapport au critère u est reliée à $|\sum_i \sum_j \alpha_i \alpha_j y_i y_j \frac{\partial K(x_i, x_j)}{\partial \sigma_u}|$. Pour plus de détails sur cette procédure de sélection de variables, le lecteur peut se référer à [10].

4 Résultats expérimentaux

Dans le cas de la fusion par combinaison linéaire, nous avons utilisé les résultats de 3 critères sur la base d'images *BIS* et ses 1200 résultats de segmentation présentés précédemment : le critères de Rosenberger avec paramètres automatiques, le critère de Borsotti et celui de Zéboudj.

En ce qui concerne la méthode de fusion par apprentissage présentée, nous avons utilisé les valeurs de 22 critères (le critère de Rosenberger [11] avec 19 jeux de paramètres et une version avec calcul automatique de paramètres, le critère de Borsotti [1] et celui de Zéboudj [13]) sur la même base d'images. Dans ce cas, nous avons effectué une sélection de variables pour identifier les critères les plus performants.

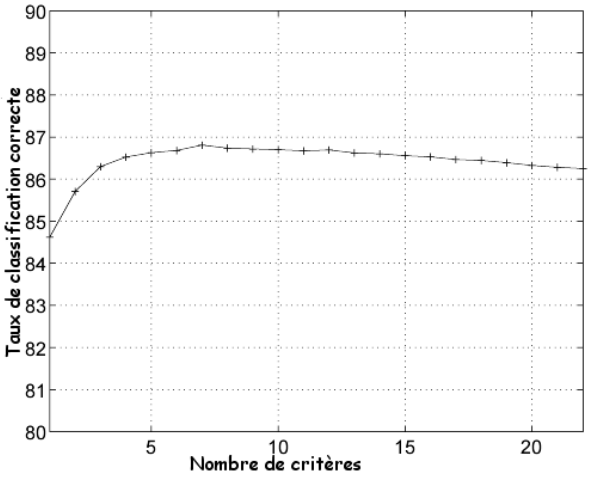


FIG. 2: Performances de l'apprentissage par SVM par rapport au nombre de critères d'évaluation utilisés pour prédire la mesure de Vinet.

Nous pouvons constater sur la figure Fig. 2 que les meilleures performances sont obtenues lorsque nous utilisons 7 critères (dont celui de Rosenberger automatique, celui de Borsotti et celui de Zéboudj). Nous avons alors sélectionné les critères correspondants et nous avons regardé la performance de notre critère fusionné pour différentes tailles de base d'apprentissage. Les résultats sont présentés dans la figure Fig. 3.

On peut remarquer que même avec très peu d'éléments dans la base d'apprentissage (1% de la base d'apprentissage) notre méthode de fusion de critères est capable de prédire avec précision la mesure de Vinet (78% de taux de classification correcte). Ainsi, nous pouvons nous contenter de quelques résultats de segmentation à comparer à une vérité terrain pour les utiliser comme base d'apprentissage.

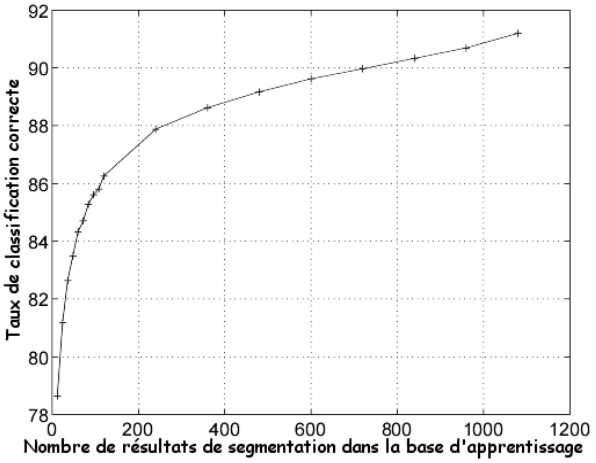


FIG. 3: Taux de classification correcte de la mesure de prédiction de la mesure de Vinet par rapport au nombre d'éléments de la base d'apprentissage.

Le tableau Tab. 1, présente les résultats obtenus par les méthodes présentées précédemment sur toute la base de résultats de segmentation.

TAB. 1: Optimisation du facteur de corrélation avec la mesure de Vinet par les deux procédés de fusion.

	Global	Mixte	Uniforme	Texturé
Meilleur résultat sans optimisation	0,7068	0,2638	0,9009	0,3554
Fusion par combinaison linéaire	0,7711	0,2934	0,9483	0,4195
Fusion par apprentissage	0,8624	0,4651	0,9784	0,5543

Comme nous pouvons le constater, la fusion par combinaison linéaire permet d'obtenir un gain de l'ordre de 10% par rapport au meilleur résultat obtenu avant fusion, tandis que nous obtenons un gain de l'ordre de 20% dans le cas de la fusion par apprentissage.

5 Conclusions et perspectives

De nos jours, le traitement d'images prend une importance grandissante. En vue d'interpréter des images acquises par un procédé quelconque, de nombreux algorithmes de segmentation ont été élaborés. L'évaluation de ces méthodes de segmentation est alors indispensable pour quantifier leurs performances.

La première méthode que nous avons présentée, permettant de fusionner par combinaison linéaire plusieurs critères d'évaluation de résultats de segmentation, donne un gain général de fiabilité de l'ordre de 10%, tandis que la deuxième méthode, par apprentissage, obtient un gain général de l'ordre de 20%.

Les perspectives d'une telle étude sont multiples. Elles concernent de simples applications comme l'évaluation de nouvelles méthodes de segmentation ou le paramétrage optimal d'un algorithme de segmentation, mais peut aussi permettre la définition d'un nouvel algorithme de segmentation reposant sur l'optimisation du critère développé.

6 Remerciements

Les auteurs tiennent à remercier la région du CENTRE et le FSE pour leurs soutiens financiers.

References

[1] M. Borsotti, P. Campadelli, and R. Schettini. Quantitative evaluation of color image segmentation results. *Pattern Recognition Letters*, 19:741–747, 1998.

[2] B. Mc Cane. On the evaluation of image segmentation algorithms. *DICTA'97 and IVCN'Z97*, pages 455–459, october 1997.

[3] J.-P. Cocquerez and S. Philipp. *Analyse d'Images: filtrage et segmentation*. Masson, 1995.

[4] J. Freixenet, X. Muñoz, D. Raba, J. Marti, and X. Cufi. Yet another survey on image segmentation: Region and boundary information integration. *ECCV*, pages 408–422, 2002.

[5] R. H. Haralick and L. G. Shapiro. Image segmentation techniques. *Computer Vision Graphics Image Processing*, 29:100–132, 1985.

[6] C. R. Houck, J. A. Joines, and M. G. Kay. A genetic algorithm for function optimisation: a matlab implementation. Technical report, NCSU-IE, sept 1995.

[7] C.-W. Hsu and C.-J. Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, (13):415–425, 2002.

[8] H. Laurent, S. Chabrier, C. Rosenberger, B. Emile, and P. Marché. Etude comparative de critères d'évaluation de la segmentation. *Actes du 19ème Colloque GRETSI*, 3:150–153, 2003.

[9] S. Philipp-Foliguet. Evaluation de la segmentation. Technical report, ETIS, mars 2001.

[10] A. Rakotomamonjy. Variable selection using svm-based criteria. *Journal of Machine Learning Research*, 3:1357–1370, 2003.

[11] C. Rosenberger. *Mise en Oeuvre d'un Système Adaptatif de Segmentation d'Images*. PhD thesis, Université de Rennes 1, décembre 1999.

[12] L. Vinet. *Segmentation et mise en correspondance de régions de paires d'images stéréoscopiques*. PhD thesis, Université de Paris IX Dauphine, Juillet 1991.

[13] R. Zeboudj. *Filtrage, Seuillage Automatique, Contraste et Contours: du Pré-Traitement à l'Analyse d'image*. PhD thesis, Université de Saint Etienne, 1988.

[14] Y. J. Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.

[15] Y.J. Zhang. Evaluation and comparison of different segmentation algorithms. *Pattern Recognition Letters*, 18:963–974, 1997.

Chaînes de Markov cachées multivariées à bruit corrélé non gaussien, avec applications à la segmentation du signal radar

N. BRUNEL[1,*], W. PIECZYNSKI[1], F. BARBARESCO[2]

[1]CITI, GET/INT, 9, rue Charles Fourier, 91011 Evry cedex, France.

[2]Thales Air Defence, 9, rue des Mathurins, 92223 Bagneux, France

Email: nicolas.brunel@int-evry.fr, wojciech.pieczynski@int-evry.fr, frederic.barbaresco@fr.thalesgroup.com

Résumé – Nous proposons un modèle de signal multivarié, corrélé temporellement et spectralement, adapté à la modélisation du signal Radar. Le modèle proposé est une chaîne de Markov cachée telle que la loi des observations appartienne à la famille des SIRV et dont la corrélation temporelle de la texture est décrite au moyen d'une copule. Il demeure alors possible d'estimer les paramètres de la loi des observations, et nous étudions la robustesse de cette estimation en faisant varier le type et la force de la dépendance temporelle en utilisant différentes familles de copules. Finalement, nous explorons l'influence de la non prise en compte de cette dépendance dans les procédures de segmentation statistique des signaux Radar fondées sur des chaînes de Markov cachées.

Abstract *– We propose a multivariate signal model with temporal and spectral dependence, well-fitted for the modeling of Radar signals. The proposed model is a Hidden Markov Chain for which observations are Spherically Invariant Random Vectors (SIRV), and temporal correlation is described by a copula. It is still possible to estimate the parameters of the SIRV, and we study the robustness of the estimation under different kinds of copulas and various strength of dependence. Finally, we explore the influence of an omitted dependence for statistical segmentation of Radar signals based on Hidden Markov Chains.*

1. Introduction

Nous traitons dans cet article de la modélisation et du traitement des signaux multivariés, non gaussiens, corrélés temporellement et spectralement. Le caractère non gaussien des signaux en radar a motivé l'introduction de la loi K et de ses variantes pour la modélisation des intensités [1, 7], ou encore des «Spherically Invariant Random Vector» (SIRV) qui en sont l'extension pour les signaux multidimensionnels [16, 8]. Cette approche permet une représentation statistiquement correcte des signaux Radar effectivement mesurés [6], et reste suffisamment proche du modèle gaussien pour en permettre des interprétations simples tout en palliant les problèmes de robustesse de l'estimation et une décroissance trop rapide des queues de distribution. Les signaux reçus en Radar sont généralement dépendants selon l'axe distance, et plusieurs approches ont déjà été proposées pour décrire cette dépendance [12]. L'objectif est ici de proposer un famille de processus stochastiques qui permette de décrire simultanément ces deux caractéristiques : des marges appartenant à la famille des SIRV et dont les matrices de covariance représentent la corrélation spectrale, et une dépendance temporelle (ou spatiale). Nous nous intéressons principalement à cette dernière propriété que nous introduisons par le biais des copules [10, 14], outil statistique encore peu utilisé en traitement du signal (voir cependant [3]). Nous abordons alors l'influence de cette dépendance dans le contexte de la segmentation, laquelle est souvent fondée sur des modèles de Markov cachés dans lesquelles l'hypothèse d'indépendance des observations (conditionnellement aux classes recherchées) est faite. Nous rappelons dans la section suivante les définitions et propriétés des SIRV et des copules. Nous montrons ensuite comment celles-ci permettent de construire le modèle proposé. Nous utilisons alors ce dernier pour proposer un modèle triplet [15] afin de simuler des observations ne respectant pas les hypothèses d'indépendance des modèles « Hidden Markov Chain » (HMC). En variant le type de la copule et la force de la dépendance, nous expérimentons l'influence de cette dépendance dans la procédure d'estimation et de segmentation. Nous en tirons alors des conclusions quant à la robustesse et à la fiabilité des procédures de segmentation non-supervisée dans le traitement du signal radar, et de manière plus générale pour des processus stationnaires.

2. SIRV dépendants

2.1 Modèle SIRV et notations

Un vecteur aléatoire $Z \in \mathbb{R}^M$ est appelé SIRV si s'écrit :

$$Z = U^{-1/2}\varepsilon \tag{1}$$

avec ε vecteur gaussien de loi $N(0,\Sigma)$ appelé « speckle » et U variable aléatoire réelle positive, appelée « texture ». La loi de Z est donc un mélange continu de lois gaussiennes centrées. Nous étendons la définition des SIRV telle qu'elle est usuellement donnée en radar au cas où la moyenne de Z peut être non nulle. De telles lois sont dites à contour elliptique ou simplement elliptiques en raison de la forme de leur densité [9]. Les membres de cette famille les plus utilisés

dans les applications sont déduits de la loi gamma. Lorsque la variable U suit une loi gamma $\gamma(\nu/2, 2/\nu)$, la loi de Z est une loi de Student (loi T) paramétrée par le triplet (m,Σ,ν), et si la variable U^{-1} suit une loi gamma $\gamma(a,\frac{1}{a})$, c'est une loi K de paramètre (m,Σ,a).

La matrice Σ est proportionnelle à la matrice de variance de Z. Les paramètres a, ν règlent la vitesse de décroissance des queues, et la loi normale apparaît comme un cas limite des lois K et T lorsqu'ils tendent vers l'infini. Les densités s'écrivent

$$\text{Loi T} \quad f(z|\mathbf{m},\Sigma,\nu) = \frac{\Gamma(\frac{\nu+M}{2})|\Sigma|^{-1/2}}{(\pi\nu)^{M/2}\Gamma(\frac{\nu}{2})}\left(1+\frac{2p(z)}{\nu}\right)^{-\frac{\nu+M}{2}} \tag{2}$$

$$\text{Loi K} \quad f(z|\mathbf{m},\Sigma,a) = \frac{2a^a|\Sigma|^{-1/2}}{(2\pi)^{M/2}\Gamma(a)}\left(\sqrt{\frac{p(z)}{a}}\right)^{a-\frac{M}{2}} K_{a-\frac{M}{2}}\left(2\sqrt{ap(z)}\right) \tag{3}$$

en écrivant $p(z) = \frac{1}{2}(z-m)'\Sigma^{-1}(z-m)$, et $'$ est l'opérateur de transposition.

2.2 Copules

Une copule de dimension 2 est la fonction de répartition de 2 variables aléatoires uniformes sur le carré $[0,1]^2$, [5, 10, 14]. L'intérêt des copules est qu'elle permettent de faire le lien entre lois jointes et lois marginales (ou marges). Si $\tilde{F}$ est la fonction de répartition (f.d.r) jointe de deux vecteurs aléatoires (V_1,V_2), de f.d.r F_1, F_2, nous pouvons affirmer grâce au théorème de Sklar [14] que

$$\tilde{F}(v_1,v_2) = C(F_1(v_1),F_2(v_2)) \tag{4}$$

C est aussi la fonction de répartition du vecteur $(F_1(V_1),F_2(V_2))$. Pour cette raison, la fonction dérivée $c(x,y) = \frac{\partial^2}{\partial x \partial y}C(x,y)$ est appelée densité de la copule C. Les copules se généralisent en dimension M quelconque comme les f.d.r de variables à marges uniformes sur l'hypercube $[0,1]^M$. Nous proposons 3 familles différentes de copules paramétriques que nous utiliserons dans les simulations dans la section 4.

Les copules elliptiques sont déduites en inversant la relation de l'Eq. (4) ce qui permet d'obtenir l'expression de la copule (ou de sa densité) pour des familles multivariées connues.

La copule (bivariée) de la loi normale est

$$c(u,v;\rho) = (1-\rho^2)^{-1/2}\exp\left(-\frac{\varsigma_1^2+\varsigma_2^2-2\rho\varsigma_1\varsigma_2}{2(1-\rho^2)}+\frac{\varsigma_1^2+\varsigma_2^2}{2}\right) \tag{5}$$

avec $\varsigma_1 = \Phi^{-1}(u), \varsigma_2 = \Phi^{-1}(v)$, et Φ^{-1} l'inverse de la f.d.r de la loi normale (univariée) centrée et réduite.

La copule de Student a une expression plus complexe

$$c(u_1,..,u_M) = \frac{\Gamma(\frac{\nu+2}{2})\Gamma(\frac{\nu}{2})}{\sqrt{1-\rho^2}\Gamma(\frac{\nu+1}{2})^M}\frac{\left(1+\frac{\varsigma_1^2+\varsigma_2^2-2\rho\varsigma_1\varsigma_2}{\nu(1-\rho^2)}\right)^{-\frac{\nu+2}{2}}}{\left((1+\frac{\varsigma_1^2}{\nu})(1+\frac{\varsigma_2^2}{\nu})\right)^{-\frac{\nu+1}{2}}} \tag{6}$$

avec $\varsigma_1 = T_\nu^{-1}(u), \varsigma_2 = T_\nu^{-1}(v)$, et T_ν^{-1} l'inverse de la f.d.r de la loi T univariée à ν degrés de libertés. Le paramètre $\rho \in [-1,1]$ correspond au coefficient de corrélation de la matrice de covariance intervenant dans la définition d'une loi elliptique. Cependant, si ces copules sont utilisées pour définir la loi jointe d'un vecteur V_1, V_2, ρ ne correspond plus à la corrélation (usuelle) de Pearson entre V_1, V_2, mais au tau de Kendall τ entre deux variables aléatoires. Les deux mesures de dépendance sont liées par la relation $\tau = 2\arcsin(\rho)/\pi$.

Les copules archimédiennes constituent une autre famille générique de copules, définie sous la forme fonctionnelle suivante $C(u,v) = \varphi^{-1}(\varphi(u)+\varphi(v))$, avec certaines conditions sur la fonction φ (entre autre la positivité et la décroissance sur le segment $[0,1]$). La copule de Clayton est construite avec la fonction $\varphi_\alpha(t) = \frac{1}{\alpha}(t^{-\alpha}-1)$, et vaut

$$C_\alpha(u,v) = \max\left((u^\alpha+v^\alpha-1)^{-1/\alpha}, 0\right) \tag{7}$$

Le paramètre α peut varier dans $[-1,+\infty)\setminus\{0\}$. Il est relié lui aussi au tau de Kendall par l'expression $\tau = \alpha/(\alpha+2)$.

La copule gaussienne a la propriété de rendre indépendante les valeurs extrêmes (minimale ou maximale), à la différence de la copule de Student [10]. Pour cette dernière, la présence d'une valeur extrême sur une des composantes « entraîne » une valeur extrême (dans le même sens) pour l'autre variable. Lorsque ces copules sont utilisées pour la modélisation de dépendance de processus spatiaux, elles permettent de reproduire de phénomènes d'agrégation de valeurs extrêmes. La copule de Clayton ne corrèle que les valeurs minimales.

2.3 Modélisation de la dépendance par copule

Soit un processus stationnaire (au sens strict) $\mathbf{Y} = (Y_n)_{n\geq 1}$ à valeurs dans R^M dont les marges sont elliptiques. Nous proposons de modéliser la dépendance à l'aide de la théorie des copules. Si $\mathbf{Y}$ était à valeurs réelles, il serait possible de décrire tous les processus stationnaires à l'aide d'une copule et de la f.d.r de la loi Y_n, grâce au théorème de Sklar. Cependant une telle construction pose deux difficultés : l'Eq. (4) ne s'étend pas aux vecteurs aléatoires (théorème d'impossibilité, [14]) et la manipulation d'une copule à n arguments pour modéliser la loi de $(Y_1,\ldots,Y_n)$ lorsque n est grand pose des problèmes pratiques. Afin de proposer une classe assez large de processus stationnaires à marge SIRV nous introduisons la dépendance par le biais d'un processus scalaire markovien latent. L'hypothèse SIRV permet d'introduire grâce à l'Eq. (1) deux processus $\mathbf{U} = (U_n)_{n\geq 0}, \boldsymbol{\varepsilon} = (\varepsilon_n)_{n\geq 0}$,. Nous supposons alors que $\mathbf{U}$ et $\boldsymbol{\varepsilon}$ sont des processus indépendants et que le processus speckle $\boldsymbol{\varepsilon}$ est i.i.d, alors que le processus de texture $\mathbf{U}$ est un processus markovien stationnaire (hypothèse similaire à celle faite dans [12]). Pour connaître entièrement sa loi, il suffit de se donner la loi de (U_1,U_2) que nous exprimons en fonction de sa copule C, ses marges étant connues et dépendantes du type de SIRV recherché. L'intérêt de l'utilisation d'une copule pour modéliser le processus $\mathbf{U}$ est double :

(i) La facilité de construction d'une chaîne de Markov à loi stationnaire fixée de densité g (et fonction de répartition G). Le noyau de transition s'écrit alors $g(u_{n+1})c(G(u_n),G(u_{n+1}))$;

(ii) La richesse des types de dépendance envisageables grâce aux nombreuses familles de copules existantes (copules elliptiques, archimédiennes).

Nous avons $p(y_n|y_{n-1},\ldots,y_1) \neq p(y_n|y_{n-1})$ parce que le processus $(\mathbf{U},\mathbf{Y})$ est une chaîne de Markov Cachée. L'estimation du processus caché $\mathbf{U}$ à partir d'observations $y_1,\ldots,y_n$ peut se révéler intéressante dans les applications.

Remarque : Des modélisations similaires ont déjà été introduites en économétrie pour la modélisation des séries temporelles financières dans lesquelles la variance (ou volatilité) est elle-même stochastique [11]. Cependant le plus souvent, il est supposé que le logarithme de la volatilité a une évolution linéaire. Les copules permettent de proposer des dynamiques non-linéaires pour le processus de variance.

3. Estimation des SIRV dépendants

Nous cherchons à estimer les paramètres $(m, \Sigma, \theta = \nu \text{ ou } a)$ d'un SIRV à partir de la réalisation $y_1, \ldots, y_N$ du processus précédent, qui est un échantillon identiquement distribué mais non indépendant. Nous estimons cependant (m, Σ, θ) à partir des équations déterminant l'estimateur du maximum de vraisemblance (EMV) dans le cas indépendant. Ceci évite d'avoir à traiter le problème de la maximisation de la log-vraisemblance d'une HMC, nécessitant des procédures complexes souvent coûteuses en temps de calcul [4].

3.1 Estimation

Pour calculer l'EMV d'un SIRV, nous proposons un algorithme EM exploitant le processus de texture $\mathbf{U}$ [13]. La log-vraisemblance complète du processus joint $(\mathbf{U}, \mathbf{Y})$ dans le cas indépendant se décompose en deux termes, ce qui permet de proposer une procédure de recherche du maximum de la vraisemblance s'opérant par une succession de maximisation sur de petits espaces, au lieu d'une recherche sur de grands espaces ($\theta = a$ ou ν) :

$$\log p\left(y_1^N, u_1^N \middle| m, \Sigma, \theta\right) = \sum_{i=1}^{N} \log\left(f(y_i | u_i, m, \Sigma)\right) + \sum_{i=1}^{N} \log\left(g_\theta(u_i)\right) \quad (8)$$

La log-vraisemblance complète de la HMC $(\mathbf{U}, \mathbf{Y})$ ne diffère que par la présence du terme $\sum_{i=1}^{N-1} \log\left(c\left(G_\theta(u_i), G_\theta(u_{i+1})\right)\right)$ représentant la dépendance entre les observations. Ainsi l'estimation utilisée sous l'hypothèse (fausse) d'indépendance revient à modifier l'algorithme EM correspondant au vrai modèle lors de la mise à jour du paramètre θ. Les formules de ré-estimations dans le cas indépendant sont :

$$m^{(n+1)} = \frac{\sum_{i=1}^{N} w_i^{(n)} z_i}{\sum_{i=1}^{N} w_i^{(n)}} \text{ et } \Sigma^{(n+1)} = \frac{1}{N} \sum_{i=1}^{N} w_i^{(n)} (z_i - \mathbf{m}^{(n+1)})(z_i - \mathbf{m}^{(n+1)})' \quad (9)$$

avec $w_i^{(n)} = E[U_i | z_i, m^{(n)}, \Sigma^{(n)}, \theta^{(n)}]$. Le paramètre de la texture s'obtient par la résolution d'une équation non-linéaire unidimensionnelle dépendant de la densité g_θ que l'on peut écrire sous la forme :

$$\theta^{(n+1)} = \psi(\theta^{(n)}, (w_i^{(n)})_{1 \le i \le N}) \quad (10)$$

3.2 Influence de la dépendance

La maximisation de la vraisemblance (8) donne toujours des estimateurs consistants mais dont la variance est plus grande que celle de l'EMV. L'algorithme EM, pour des échantillons de faible taille donne souvent un estimateur biaisé, en raison de maxima locaux de la vraisemblance le piégeant. Dans le cas d'échantillon assez grand ($N = 500$), ce problème est atténué et les tableaux 1 et 2 permettent d'évaluer l'effet de la dépendance sur l'estimation du paramètre de queue (estimés par Monte-Carlo sur 200 tirages). Nous utilisons la racine carrée de l'écart quadratique moyen (REQM) de l'estimateur afin de donner une indication de la fluctuation de l'estimateur, mais aussi de son biais.

Les paramètres (m, Σ) sont correctement estimés, et la variance des estimateurs reste stable et proche du cas indépendant. La différence devient notable pour $\tau = 0.8$. Le paramètre de queue est le plus affecté par la dépendance, ce qui revient à mal estimer le facteur multiplicatif de la covariance d'un SIRV.

TAB. 1 : Moyenne et REQM de l'estimateur du paramètre de queue de la loi T ($\upsilon = 10$)

τ	0	0.13	0.33	0.59	0.8
Gauss	10.7 (3.2)	11.1 (4.7)	11 (4)	11.5 (4.9)	15.4 (12.8)
Student	10.5 (2.5)	10.6 (2.8)	11.6 (4.8)	11.5 (6.4)	14.7 (12.6)
Clayton	-	11.4 (4.4)	12.5 (8.4)	13.7 (11.1)	23.5 (25)

TAB. 2 : Moyenne et REQM de l'estimateur du paramètre de queue de la loi K ($a = 5$)

τ	0	0.13	0.33	0.59	0.8
Gauss	4.6 (1.1)	4.7 (1.2)	4.6 (1.1)	4.9 (1.5)	5.2 (1.7)
Student	4.7 (1.2)	4.7 (1.1)	4.7 (1.4)	4.8 (1.2)	5.4 (1.9)
Clayton	-	4.7 (1.1)	4.6 (1.1)	4.9 (1.5)	5.7 (2.2)

Lorsque les queues de distribution sont plus épaisses (cas de la loi T par rapport à la loi K considérée) l'influence de la dépendance de la texture est plus importante. La corrélation des extrêmes par la copule de Student ne détériore pas la qualité de l'estimation par rapport à la copule Gaussienne. C'est par contre le type de la structure de la dépendance qui modifie les performances de la segmentation : à même tau de Kendall, la copule Clayton dégrade plus fortement la variance des estimateurs.

4. Segmentation

4.1 Dépendance conditionnelle dans les HMC

Contrairement au cas homogène traité dans la section 2, nous supposons que nous avons plusieurs zones aux caractéristiques différentes, et que celles-ci sont représentables par un processus markovien caché $\mathbf{X} = (X_n)_{n \ge 1}$ à K classes, tel que la loi de Y_n conditionnellement à $X_n = k$ soit une loi elliptique de paramètres $(m_k, \Sigma_k, \theta_k)$. Nous cherchons à estimer de façon non-supervisée le processus $\mathbf{X}$ par l'estimateur bayésien Maximum des Marges a Posteriori (MPM), en supposant que nous avons une chaîne de Markov cachée.

Nous introduisons une dépendance entre les observations, conditionnellement aux états par le biais du processus de texture $\mathbf{U}$ ce qui permet d'évaluer la robustesse (par rapport aux écarts aux hypothèses du modèle) des procédures de segmentation non-supervisée dans le cas où l'hypothèse d'indépendance conditionnelle des observations est mise en défaut. Cette corrélation spatiale supplémentaire est introduite par une copule c modélisant la dépendance de (U_n, U_{n+1}), tel que le processus $(\mathbf{X}, \mathbf{U}, \mathbf{Y})$ soit une chaîne de Markov homogène et stationnaire (c'est un cas particulier de chaînes de Markov Triplet). Ce nouveau modèle généralise aussi bien le modèle univarié [3], que le modèle multivarié [2].

4.2 Segmentation non-supervisée avec HMC-IN

Les segmentations sont obtenues sous hypothèse d'indépendance conditionnelle des observations, pour un modèle à 3 classes en dimension 2. Dans l'exemple traité, les moyennes sont $m_1 = (0\,0)'$, $m_2 = (1.5\ 1.5)'$, $m_3 = (3\ 3)'$, les variances sont toutes les normalisées, avec des coefficients de corrélation distincts $\rho_1 = 0.4$, $\rho_2 = 0.2$, $\rho_3 = 0.5$. Enfin, les paramètres de queue sont $\nu_1 = 5$, $\nu_2 = 10$, $\nu_3 = 15$ pour les lois T, et $a_1 = 2.5$, $a_2 = 5$, $a_3 = 7.5$ pour les lois K. Les taux d'erreur se dégradent lentement par rapport au cas indépendant (cas Gauss avec $\tau = 0$), et ne montre un fort écart que dans le cas de tau de Kendall à 0.8. Les écart-types des taux d'erreur restent stables mais deviennent supérieurs au cas indépendant à partir de $\tau \geq 0.59$, indiquant une plus grande variabilité dans la qualité des segmentations non-supervisées, ce qui est particulièrement net dans le cas de la copule de Clayton (pour la loi T et dans une moindre mesure pour la loi K).

TAB. 3 : Taux d'erreur et écart-type en % (loi T)

τ	0	0.13	0.33	0.59	0.8
Gauss	17	17	17.2	17.3	17.7
	(3.4)	(3)	(3)	(3.9)	(4.5)
Student	17.5	17.5	17.5	17.6	18
	(3.5)	(3.4)	(3.4)	(4)	(4.9)
Clayton	-	17	17	17.9	19.1
		(3.2)	(3.4)	(4.3)	(7.7)

TAB. 4 : Taux d'erreur et écart-type en % (loi K)

τ	0	0.13	0.33	0.59	0.8
Gauss	13.9	13.9	13.3	14.1	13.6
	(2.6)	(2.9)	(2.5)	(2.8)	(3.1)
Student	13.7	13.7	13.6	14.3	14
	(2.8)	(2.7)	(2.7)	(3.2)	(3.3)
Clayton	-	13.7	13.7	13.7	14
		(2.5)	(2.6)	(3.2)	(3.5)

5. Conclusion

Le modèle statistique proposé permet de décrire différentes structures de dépendance dans les processus SIRV, souvent utilisés en radar. Le processus $(\mathbf{U}, \mathbf{Y})$ est une HMC dont l'originalité est l'utilisation des copules pour la modélisation de la dépendance du processus caché, et nous avons présenté plusieurs copules induisant des types de dépendance différents. Une courte étude expérimentale montre que la dépendance dégrade l'estimation des queues de distribution, d'autant plus que celle-ci augmente et que la taille des échantillons est faible. Dans le cadre de la segmentation non-supervisée, nous avons étendu le modèle de chaîne couple $(\mathbf{X}, \mathbf{U})$ introduit dans [3] au processus triplet $(\mathbf{X}, \mathbf{U}, \mathbf{Y})$ afin d'introduire une dépendance conditionnelle dans les observations. L'influence de cette dernière reste faible lorsque les classes sont bien séparées (notamment par des moyennes distinctes), mais est notable lorsque la dépendance augmente fortement. Une étude plus approfondie nécessite d'être menée lorsque les classes ont des caractéristiques proches (entre autre les moyennes nulles), et pour identifier des familles de copules susceptibles de représenter la dépendance de données réelles radar. Le développement des méthodes d'estimation basées sur la vraie vraisemblance du modèle constitue aussi un axe de développement, afin d'appliquer des méthodes de segmentation non-supervisée dans le contexte des modèles de Markov Triplet [15].

Références

[1] R. Barakat. Direct derivation of intensity and phase statistics of speckle produced by a weak scatterer from the random sinusoid model. *J. Opt. Soc. Amer.*, 71(1) :86–90, 1981.

[2] N. Brunel, W. Pieczynski and S. Derrode, Copulas in vectorial hidden Markov chains for multicomponent image segmentation, ICASSP'05, Philadelphia, USA, March 2005.

[3] N. Brunel and W. Pieczynski, Unsupervised signal restoration using hidden Markov chains with copulas, *Signal Processing*, 2005 (to be published).

[4] O. Cappé, E. Moulines and T. Ryden, *Inference in Hidden Markov Models*, Springer-Verlag, 2005.

[5] X. Chen and Y. Fan. Estimation of copula-based semi-parametric time series models. Technical report, Department of Economics, Vanderbilt University, 2004.

[6] E. Conte, A. De Maio, and C. Galdi. Statistical validation of the compound-Gaussian model on clutter data from IPIX radar, *Proc. of Toulouse Radar 2004 Conference*, October 2004.

[7] Y. Delignon and W. Pieczynski. Modeling non-rayleigh speckle distribution in sar images. *IEEE Trans. on Geoscience and Remote Sensing*, 40(6) :1430–1435, 2002.

[8] F. Gini and M. Greco. Covariance matrix estimation for CFAR detection in correlated heavy tailed clutter. *Signal Processing*, 82(12) :1847–1859, December 2002.

[9] A.K. Gupta and T. Varga, *Elliptically contoured models in statistics*, Mathematics and its applications, Kluwer Academic Publ., 1993.

[10] H. Joe. *Multivariate Models and Dependence Concepts*, volume 73 of Monographs on Statistics and Applied Probability. Chapman and Hall, 1997.

[11] S. Kim, N. Shepard and S. Chib, Stochastic Volatility: Likelihood Inference and Comparisons with ARCH models, *Review of Economic Studies*, 65, pp. 361-393, 1998.

[12] P. Lombardo, A. Farina, Coherent radar detection against K-distributed clutter with partially correlated texture, *Signal Processing*, 48, pp. 1-15, 1996.

[13] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, 2000.

[14] R. B. Nelsen. *An introduction to Copulas*. Number 139 in *Lecture notes in Statistics*. Springer-Verlag, 1998.

[15] W. Pieczynski, C. Hulard, and T. Veit, Triplet Markov Chains in hidden signal restoration, *SPIE's International Symposium on Remote Sensing*, September 22-27, Crete, Greece, 2002.

[16] M. Rangaswamy, D. Weiner, and A. Ozturk. Non-Gaussian random vector identification using spherically invariant random process. *IEEE Trans. on Aerospace and Electronic Systems*, 29(1) :111–124, January 1993.

* La thèse de N. Brunel est financée par la Délégation Générale pour l'Armement (DGA/DSP).

Représentation d'images par chaînes de symboles : Application à la recherche par le contenu *

Isabelle SIMAND, Jean-Michel JOLION

LIRIS, FRE CNRS 2672 INSA
Bât. J. Verne
INSA Lyon, 69621 Villeurbanne cedex - France
{Isabelle.Simand, Jean-Michel.Jolion}@liris.cnrs.fr

Résumé – Cet article introduit une nouvelle représentation d'images basée sur des chaînes de symboles. Mêlant les notions de points d'intérêt, de contraste, et d'ordre, cette signature offre une description à la fois concise et précise de l'image. La comparaison de deux images consiste en la comparaison des symboles qui composent les chaînes signature. D'où l'élaboration d'une distance entre symboles dérivée de la distance d'édition.

Abstract – This paper introduces a new representation of images based on strings of symbols. This signature, both precise and compact, is based on notions such as interest points, contrast and order. The comparison between two images consists in the comparison of the symbols which compose the strings. It leads to the development of a distance between symbols such as the edit distance.

1 Introduction

Le domaine de l'indexation a connu un essor très rapide ces dernières années, conduisant à un ensemble de méthodologies et de techniques de plus en plus stables. Les travaux exposés ici renouent avec l'approche exploratoire initiale, en s'appuyant, entre autres, sur les outils de la reconnaissance des formes. En effet si les caractéristiques extraites dans les images sont souvent numériques, celles qui constituent notre code d'images sont symboliques. Par conséquent nous conservons toute la puissance de l'approche structurelle de la reconnaissance des formes. En particulier, la phase de comparaison pourra faire intervenir une distance entre symboles, telle que la distance de Levenstein. L'apport d'une technique de ce type sera illustré dans la suite.
Cette étude trouve sa source dans divers travaux sur le parcours visuel de l'oeil humain, qui est particulièrement attiré par les zones de fort contraste [1], ce qu'illustre d'ailleurs la loi de Naka-Rushton. En outre, dans [2], Thorpe montre l'importance de la notion d'ordre dans l'analyse des images, ce qui justifie l'usage d'une liste triée de symboles. Nos travaux ont donc pour but de traduire ces deux éléments (ordre et contraste) dans une nouvelle définition d'un code d'images pour la classification, la reconnaissance et l'indexation d'images.

2 Architecture générale

Lors de l'analyse d'une image, l'oeil effectue une succession de focus d'attention aussi appelée saccade. L'extraction de points d'intérêt dans une image, si elle se base sur un critère pertinent, constitue une bonne traduction de ce comportement, sans en être une modélisation. En outre la représentation, peu coûteuse, d'une image par un tel ensemble de points a aujourd'hui fait la preuve de ses performances (ce depuis les travaux de Schmid et Mohr [3]). C'est pourquoi notre code reprend cette approche, combinée avec les notions d'ordre et de contraste. La figure 1 illustre les différentes étapes de notre méthode, de l'image à sa signature.

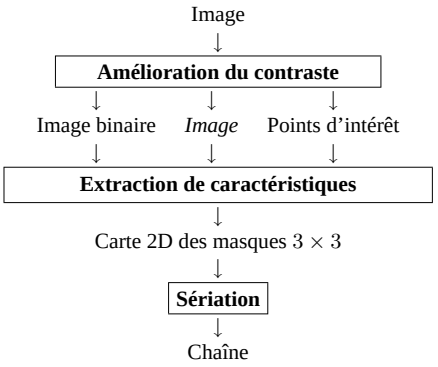


FIG. 1: *Schéma général de l'extraction du code à partir d'une image.*

2.1 Binarisation

La première étape du codage consiste en la binarisation de l'image, qui est préalablement convertie en niveaux de gris si nécessaire. Nous utilisons à ce stade une méthode développée il y a quelques années [4], et basée sur le rehaussement de contraste. Elle vérifie les étapes suivantes :

- représentation multirésolution d'une image sous la forme d'une pyramide,
- calcul d'une pyramide de contraste par ratio de deux niveaux successifs,

*Cette étude est financée par France Télécom R&D dans le cadre du contrat ECAV3 42568725.

FIG. 2: *Plusieurs niveaux de la pyramide initiale (à gauche) et sa binarisation obtenue en 7 itérations.*

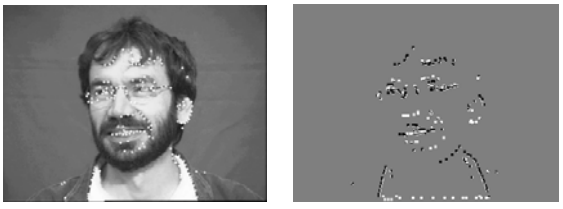
- réhaussement itératif du contraste de l'image à tous les niveaux de résolution,
- reconstitution de l'image réhaussée par projection top-down.

Le point de convergence de ce processus est une image binaire dont l'information de contraste a été préservée, tout en réduisant l'entropie, ce qui permet de conserver un maximum de détail. La figure 2 illustre cette convergence.

2.2 Points d'intérêt

La première étape de binarisation permet aussi l'extraction des points d'intérêt, point de départ de notre signature. En effet, ces points sont les maxima locaux de contraste, calculés lors de la première itération du mécanisme de binarisation. Ces points peuvent être extraits à plusieurs niveaux de résolution, dans le but de récupérer davantage d'information. La figure 3 est un exemple d'extraction de ces points.
Jouant le rôle d'un filtre sur l'image binaire, ces points d'intérêt et leur voisinage immédiat seront les seules informations retenues pour constituer le code d'une image. Ce code comprend deux composantes :

- un symbole binaire, centré sur un point d'intérêt, extrait de l'image binaire
- une double composante couleur associée à ce symbole, extrait à partir de l'image originale.



(a) (b)

FIG. 3: *(a) Image originale avec superposition des points d'intérêt, (b) masques binaires ,extraits de la version binaire de (a), dans le voisinage des points d'intérêt.*

2.3 Sériation

L'étape précédente fournit donc un ensemble de points, répartis dans l'image. A l'heure actuelle, comparer deux distributions 2D (deux graphes) reste encore très complexe, surtout pour plusieurs milliers de points (sommets). Le passage en 1D, suppose l'utilisation d'un procédé de sériation, qui peut être de différentes natures : parcours du graphe, parcours de l'image, etc. Ici, la sériation s'effectue simplement à partir d'un critère scalaire. En effet la liste de points est ordonnée par ordre décroissant de la valeur de contraste, obtenue lors de l'extraction des points. Cela ne permet malheureusement pas de conserver l'information spatiale contenue dans la distribution 2D de points, mais les autres méthodes de sériation développées lors de cette étude n'ont pas encore permis de mettre en évidence l'apport d'une telle information [5].

3 Un nouveau code

3.1 Définition

Le code associé à chaque image est donc une chaîne ordonnée d'entités, chacune constituée d'un symbole et d'une double composante couleur. Ces symboles, ou masques, composés de pixels noirs et blancs, sont de taille 3 par 3 pixels. Il en existe 512 différents, ce qui fournit une diversité conséquente, tout en restant raisonnable à étudier en détail.
Considérée comme un complément, l'information couleur devait être grossière. Pour chaque masque sont calculés deux couleurs : l'une correspondant aux pixels noirs N et l'autre aux pixels blancs B du masque. Elles résultent de la moyenne seuillée des composantes couleurs de ces deux zones N et B. On se ramène alors à seulement 8 valeurs pour chacune des deux couleurs extraites, ce qui permet déjà d'améliorer les résultats (cf. section 4), en gardant un code compact. En effet, sa taille est de 9 bits pour un masque et 2x3 bits pour la couleur, soit 15.n bits pour une chaîne de longueur n points. Pour une image I, on a :

$$\mathrm{I} \rightarrow (\ \begin{smallmatrix} c_{0.2} \\ c_{0.1} \end{smallmatrix}, \begin{smallmatrix} c_{1.2} \\ c_{1.1} \end{smallmatrix}, \begin{smallmatrix} c_{2.2} \\ c_{2.1} \end{smallmatrix}, \begin{smallmatrix} c_{3.2} \\ c_{3.1} \end{smallmatrix}, \begin{smallmatrix} c_{4.2} \\ c_{4.1} \end{smallmatrix}, \begin{smallmatrix} c_{5.2} \\ c_{5.1} \end{smallmatrix} \ldots)$$

3.2 Comparaison de deux codes d'images

Le développement d'un code symbolique et non numérique permet d'avoir accès à l'ensemble des outils de la reconnaissance de formes, qui sont à la fois performants et originaux dans le domaine de l'indexation d'images. En particulier, une distance de type Levenstein, qui fait intervenir des coûts élémentaires de substitution, de suppression et d'insertion entre symboles, est dans notre cas tout à fait appropriée. En effet, pour savoir si deux images se ressemblent, il s'agit de déterminer la similarité de leur signature. Autrement dit, calculer quel est le coût pour passer d'une chaîne à l'autre, ce coût devant être d'autant plus élevé que les images comparées sont différentes. C'est exactement ce qu'offre ce genre de distances. Qui plus est, la notion d'ordre est conservée si les coûts élémentaires sont calculés entre deux entités de même rang dans la chaîne.
La distance de Levenstein étant trop coûteuse pour de longues chaînes, nous avons mis au point d'autres distances qui s'en inspirent. A commencer par une distance de distributions. Dans

ce cas on construit pour chaque chaîne un histogramme à 512 entrées qui contient les occurences des masques qui la composent. La différence entre les deux histogrammes ainsi obtenus représente alors la valeur de similarité entre les deux images comparées. A noter qu'on peut faire de même pour la couleur avec un histogramme à 8 entrées. Ici malheureusement, la notion d'ordre disparaît totalement.
Nous avons donc élaboré une troisième distance, à mi-chemin entre les deux précédentes : ici, l'incrément lors de la construction des histogrammes H_1 et H_2 correspond au coût de substitution entre deux masques de même rang. En outre, ce coût est pondéré par la position des masques dans les chaînes, afin de favoriser les masques placés en tête. Plus formellement, cela se traduit par :

$$H_1 \leftarrow 0\,; H_2 \leftarrow 0\,; W \leftarrow 0$$
$$\text{for } i : 1 \rightarrow \min(l(s_1), l(s_2))$$

$$\text{do} \begin{cases} H_1(m_1^i) = H_1(m_1^i) + \omega_i \times c(m_1^i, m_2^i) \\ H_2(m_2^i) = H_2(m_2^i) + \omega_i \times c(m_1^i, m_2^i) \\ W = W + \omega_i \times c(m_1^i, m_2^i) \end{cases}$$

$$d(s_1, s_2) = \textstyle\sum_i |H_1(i) - H_2(i)|/W$$

où $l(s_1)$ et $l(s_2)$ sont les longueurs des chaînes s_1 et s_2 ; $c(m_1^i, m_2^i)$ le coût de substitution entre deux masques de rang i ;
et ω_i la pondération du type $K - i$ (K constante ou égale à $\min(l(s_1), l(s_2))$+1). W permet de normaliser la valeur obtenue afin qu'elle soit indépendante de la longueur des chaînes comparées.
Une nouvelle fois, on peut procéder de même avec la couleur.

3.3 Distance entre deux symboles

Ainsi, pour connaître la similarité de deux chaînes, il est nécessaire d'établir le coût de substitution d'un masque par un autre, ces coûts étant alors stockés dans une matrice symétrique 512x512.
Le coût de base est la distance de Hamming, qui cumule le nombre de bits différents entre deux mots de 9 bits qui codent deux masques quelconques. Cependant, pour prendre en compte l'information symbolique contenue dans un masque, nous avons établit une distance dite structurelle. Pour se faire, les masques ayant les mêmes caractéritiques de forme ont été regroupés, et la distance de Hamming a alors été utilisée pour définir des distances moyennes inter- et intra-classes.
Par ailleurs, les coûts de substitution entre deux couleurs élémentaires ont été stockées dans une matrice 8x8, compte tenu du nombre de leurs composantes R, G et B qui diffèrent.
D'autre part, ces coûts peuvent aussi faire l'objet d'un apprentissage par l'exemple [6].

4 Expérimentations

Afin d'aborder au mieux les différents aspects offerts par la vaste thématique de l'indexation, nous avons mis au point différentes bases d'images permettant de répondre à un ensemble de questions (classification, reconnaissance d'une classe particulière, similarité à une image requête, robustesse à l'ajout de bruit, etc.). Par manque de place nous ne présenterons qu'une partie de nos résultats, dont le détail est présenté dans [6].

FIG. 4: *Echantillon d'images de la base 1. Sept catégories : bateaux, voitures, avions, outils, trains, légumes, objets vikings*

La figure 4 représente un échantillon d'images, issues de sept classes de 30 à 50 images, regroupées selon un critère sémantique tel que "voiture". Cette première base a pour objectif d'expérimenter la classification d'un ensemble d'images parmi un nombre prédéfini de classes.

La figure 5 montre l'avantage de notre méthode par rapport à une classification de référence basée sur les histogrammes couleurs. Avec tous les pixels de l'image, cette méthode permet d'obtenir 68.7% de bonne classification, quand la nôtre peut atteindre près de 74% avec seulement 2500 symboles. Cette figure présente aussi le résultat que l'on obtient si on optimise globalement la matrice de coûts élémentaires sur cette base d'image. Cette optimisation consiste en un "apprentissage", sur une base ayant les mêmes classes, des coûts de substitution entre deux masques quelconques : si les chaînes comparées appartiennent à la même classe, ces coûts sont réduits, sinon, ils sont augmentés. Par ailleurs, différents tests ont été menés afin de montrer l'apport de la couleur dans ce type de classification. En effet, la couleur peut être combinée aux masques suite à une étape de fusion de données que nous ne détaillerons pas ici. En outre, comme nous l'avons souligné plus haut, il est possible

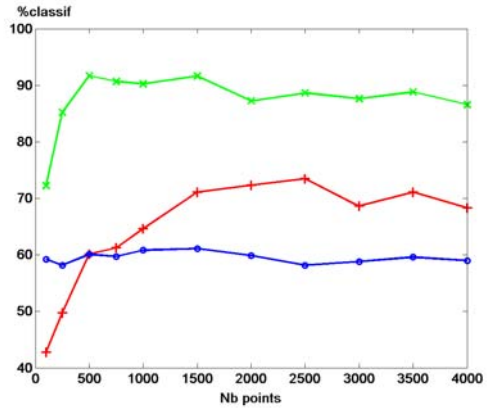


FIG. 5: *Pourcentage de bonne classification en fonction du nombre de points extraits. Valeurs obtenues avec les histogrammes couleurs (•), avec notre méthode prenant en compte l'ordre(+), avec l'optimisation (×).*

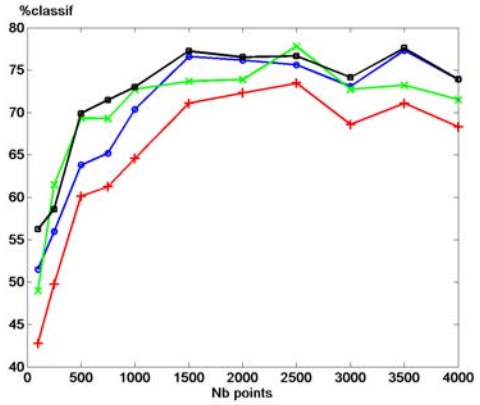


FIG. 6: *Pourcentage de bonne classification en fonction du nombre de points extraits. Valeurs obtenues avec les masques binaires seuls (+), avec les masques binaires et la couleur (•), avec les masques binaires extraits à deux niveaux de résolutions (×), avec les masques binaires et la couleur extraits à deux niveaux de résolution (□).*

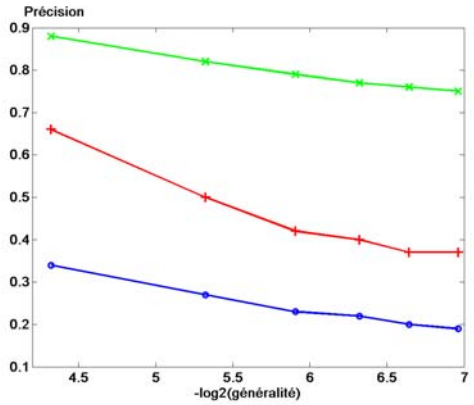


FIG. 7: *Précision moyenne pour 300 requètes en fonction de la généralité. Valeurs obtenues par Laurent* et al. *(+), avec notre méthode (masques et couleur, un seul niveau de résolution)(•), avec optimisation (×).*

d'extraire des points d'intérêt à divers niveaux de résolution, qui peuvent ensuite se combiner pour améliorer les résultats. La figure 6 illustre l'apport de ces deux informations supplémentaires.

Une autre base, composée initialement de 300 images réparties en 20 classes, à laquelle on ajoute progressivement 300 à 1500 images quelconques, permet de mettre en évidence la robustesse de la signature. On parle aussi de généralité:

$$\text{Généralité} = \frac{\text{Nombre d'images pertinentes}}{\text{Nombre d'images dans la base}} \qquad (1)$$

Cette base est initialement composée de 20 classes de 15 images chacune. La généralité pour une image donnée est donc $\frac{15}{300}$. Chacune de ces 300 images est successivement utilisée comme requète au sein de la base, afin de retrouver les 14 images les plus similaires. On calcule alors la précision moyenne obtenue. Ensuite, des paquets d'images quelconques sont progressivement ajoutés, réduisant ainsi la généralité. Le but est de tester la robustesse de la signature face à l'addition de "bruit" dans la base. Pour se faire, on calcule la nouvelle précision moyenne après chaque ajout de paquet, et ce jusqu'à l'addition de 1570 images (à comparer avec les 14 images similaires à une requète donnée).
Cette base ayant déjà été utilisée pour ce genre de tests nous avons comparé notre approche avec celle développée par Laurent *et al.* [7]. La figure 7 illustre les résultats obtenus.
Les résultats que l'on obtient sans optimisation sont inférieurs à ceux de Laurent *et al.*, dont la méthode est, rappelons le, beaucoup plus complexe que la nôtre. En revanche, lorsqu'on optimise la matrice de coûts sur la base des 300 images, la précision moyenne est toujours supérieure à 0.7.

5 Conclusion

La méthode présentée ici propose une signature d'image originale, basée sur une approche structurelle. Cette démarche, qui se distingue des approches numériques traditionnelles, offre de ce fait des outils innovants en indexation d'image tels que ceux de la reconnaissance de formes. En outre, ce codage symbolique semble plus à même de refléter le contenu sémantique d'une image, ce que confirment les résultats encourageants obtenus.
Différents points restent cependant à améliorer (la couleur, pour l'instant très grossière) et à prospecter : une étude plus précise des chaînes permettra de déterminer des longueurs optimales ; la mise en place d'une méthode de comparaison à la volée offrira un résultat plus rapide ; etc.

Références

[1] P.J. Parkhurst et E. Niebur. Scene content selected by active vision. *Spatial Vision*, 16(2):125-154, 2003.

[2] S.J. Thorpe, A. Delorme, et R. VanRullen. Spike-based strategies for rapid processing. *Neural Networks*, 14:715-725, 2001.

[3] C. Schmid et R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Trans. on PAMI*, 19(5):530-535, 1997.

[4] J.M. Jolion. Analyse multirésolution du contraste dans les images numériques. *Traitement du Signal*, 11(3):245-255, 1994.

[5] J. Ros, C. Laurent, J.M. Jolion et I. Simand. Comparing String Representations and Distances in a Natural Images Classification Task, L. Brun and M. Vento (Eds), *GbR-PR'05*, 2005, LNCS 3434, 71-83.

[6] J.M. Jolion, I. Simand et P. Prabhat. Développement d'une nouvelle représentation de formes contenues dans des images par analyse du parcours visuel et points saillants. Rapport technique ECAV-3, LIRIS, INSA Lyon, 2004.

[7] C. Laurent, N. Laurent, M. Maurizot et T. Dorval. In depth analysis and evaluation of saliency-based color image indexing methods using wavelet salient features. *Multimedia Tools and Application*, 2004.

Histogrammes spatiaux couleur optimisés pour l'indexation d'images par le contenu

S. KHARBOUCHE, P. VANNOORENBERGHE, C. LECOMTE, P. MICHE

Laboratoire Perception Systèmes Information, FRE 2645 CNRS

Université de Rouen, UFR des Sciences

Place Emile Blondel, 76821 Mont Saint Aignan cedex, France

{Said.Kharbouche, Patrick.Vannoorenberghe, Christele.Lecomte}@univ-rouen.fr, Pierre.Miche@insa-rouen.fr

Résumé – Dans cet article, nous présentons une méthode qui augmente l'efficacité et la précision des systèmes d'indexation et de recherche d'images couleur par le contenu basés sur les histogrammes spatiaux. Cette technique consiste essentiellement à extraire les informations spatiales de l'image en la divisant en plusieurs sous-images rectangulaires interférées (grille des histogrammes). La méthodologie proposée optimise, via une combinaison linéaire, le découpage horizontal et vertical de l'image pour le calcul des histogrammes spatiaux. Nous augmentons ainsi la performance de la recherche en réduisant en même temps le coût de stockage et le temps de calcul par rapport aux méthodes classiques dont le choix concernant le découpage reste arbitraire.

Mots clef : Indexation d'images, Recherche d'image par le contenu, Couleur, Histogramme spatial.

Abstract *– In this paper, a methodology that improves efficiency and accuracy of content-based image retrieval (CBIR) systems is presented. It is based on spatial histograms and used for colour images databases. This technique primarily consists in extracting spatial information from the image by dividing it horizontally and vertically (colour histograms grid). With a linear combination, the number of horizontal and vertical splits is optimised. It allows us to increase the performance of the CBIR system and reduce the storage cost and computing time generated by other methods based on splitting the image into equal sub-images.*

Keywords*: Image indexing, Content based image retrieval, Colour, Spatial histogram;*

1. Introduction

L'indexation d'une base d'images a été envisagée dans un premier temps en associant à chaque image des commentaires textuels généralement sous forme de mots-clefs. Ces dernières années, une autre voie consiste à indexer et rechercher directement les images à partir de leurs contenus. Cette deuxième technique est plus intéressante parce qu'elle est indépendante des thésaurus associé aux commentaires, plus riche que le texte et permet une description moins subjective que le texte et souvent plus efficace. Cette technique semble bien adaptée à plusieurs domaines d'application (albums personnels, images biomédicales, imagerie satellitaire, télé-surveillance, média, …).

L'indexation de la base étant basée sur le contenu des images, on étudie la similarité perceptuelle entre les images, l'image pouvant alors être caractérisée par trois attributs : la forme, la texture et la couleur [10]. L'extraction de ces attributs pour chacune des images de la base permet d'obtenir une signature de chaque image. Avant de calculer cette signature, il faut souvent résoudre les problèmes du choix de l'espace couleur, de la quantification de l'image et des traitements à effectuer au préalable sur les images :

Choix de l'espace couleur : Il n'y a pas d'espace couleur idéal pour l'indexation. Les espaces couleur les plus utilisés sont RGB (Red, Green, Blue), HSV (Hue, Saturation, Value), Lab ou Luv (L : luminance, (a,b) et (u,v) représentent la chrominance). Quelques auteurs proposent d'utiliser des espaces couleur hybrides [3].

Quantification de l'image : Elle consiste à réduire le nombre de couleurs présentes dans l'image afin de ne conserver uniquement que les couleurs dominantes. Cette étape facilite la segmentation et le calcul des signatures. On distingue deux catégories de quantification, la quantification statique (uniforme ou non-uniforme) et la quantification dynamique (classification).

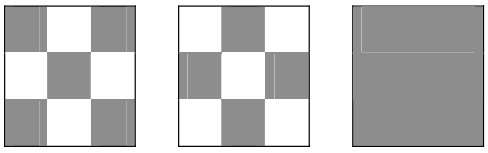
Pré-traitements : Cette étape est abordée surtout si on veut extraire des informations de forme. Elle peut permettre également d'éliminer le bruit dans l'image, de localiser et déterminer des frontières de zones ou des points d'intérêt dans l'image.

Dans cet article, nous nous intéressons uniquement à l'attribut couleur du fait de son haut pouvoir discriminant par rapport aux autres attributs mais l'approche reste concevable pour tous autres attributs extraits de l'image. Nous travaillons dans l'espace HSV qui s'avère plus dédié à l'évaluation de la similarité perceptuelle entre deux couleurs. On réalise sur les images de la base une quantification statique et uniforme sur les trois axes de l'espace HSV. La section 2 décrit les histogrammes spatiaux couleur ainsi que leur utilisation pour l'indexation d'images par le contenu. Des simulations sur une base d'images généraliste sont ensuite présentées dans la section 3 de cet article.

2. Histogrammes spatiaux couleur

Une des particularités de l'utilisation de l'histogramme en tant que signature réside dans le fait que l'image reste invariante par translation ou rotation. Ce grand avantage est obtenu par rapport aux autres approches au détriment des informations spatiales qui ne sont pas prises en considération par l'histogramme. Pour réduire les effets de cet inconvénient majeur, plusieurs travaux s'attachent à calculer les proportions de couleurs et leurs positions relatives par exemple la méthode CRT [4], spatial-chromatic histograms [5], color co-occurance matrix [6]. Il y a également d'autres méthodes qui consistent à diviser l'image en sous-images par exemple en 9 rectangles égaux [7], en 25 rectangles égaux [3]. Après décomposition, le système associe à chaque sous-image son histogramme couleur correspondant. Avec cette technique, la dimension du vecteur des indices sera multipliée par le nombre de rectangles, ce qui augmente la taille des signatures (espace de stockage) dans la base indexée et donc allonge le temps de calcul. Un autre inconvénient généré par ces méthodes concerne la perte d'invariance de l'image. Un exemple de ce type d'inconvénient est illustré à la figure 1 où un système de découpage en 3x3 rectangles donne une image (c) plus proche de l'image (a) que de l'image (b) !

.



(a) (b) (c)

FIG 1 : Le système basé sur la division de l'image en 9 rectangles estime que l'image (c) est plus proche de l'image (a) que de l'image (b).

La méthode que nous proposons cherche à réduire la taille de la signature (par conséquent le temps de recherche dans la base) et à augmenter l'invariance de l'image. Notre méthode consiste à subdiviser l'image quantifiée (en **c** couleurs différentes) en **m** parties horizontales et en **n** parties verticales. A partir de ce découpage, les signatures associées à chaque sous-image sont calculées. L'effet obtenu en divisant l'image verticalement, est l'invariance de l'image par translation verticale. L'effet obtenu en divisant l'image horizontalement, est l'invariance de l'image par translation horizontale. Cet avantage est illustré sur la figure 2. Cette technique de division d'image donne une certaine souplesse au système pour mesurer la similarité perceptuelle par rapport aux méthodes classiques de découpage, parce que plus le nombre de rectangles est grand, plus l'image reste invariante. De plus, les intersections entre les **n** zones verticales et les **m** zones horizontales fournissent une idée sur la proportion des couleurs dans les **nxm** rectangles qui forment l'image. Alors que la division classique de l'image en $\mathbf{n^2}$ rectangles augmente $\mathbf{n^2}$ fois la taille de la signature, la subdivision de l'image en **n** parties horizontales et en **m** parties verticales augmente (**m+n**) fois la taille de la signature.

FIG 2 : Illustration de l'invariance du découpage horizontal pour le calcul de la signature d'une image .

Pour la mesure de similarité, le système fait une pondération entre les parties horizontales et les parties verticales, les découpages verticaux restant les plus intéressants [8][4], surtout si on travaille avec des bases généralistes. Ainsi, la similarité entre deux images $\boldsymbol{I_i}$ et $\boldsymbol{I_j}$ est évaluée par le calcul de la distance suivante :

$$d(I_i,I_j)=\frac{\alpha}{m}\sum_{k=1}^{m} f(V_{i,h(k)},V_{j,h(k)})+\frac{\beta}{n}\sum_{q=1}^{n} f(V_{i,v(q)},V_{j,v(q)}) \qquad (1)$$

où $(\alpha,\beta) \in [0,1]^2$ tels que $\alpha+\beta=1$ représentent les coefficients de pondérations entre les parties verticales et les parties horizontales des deux images. Dans cette équation, $V_{i,h(k)}$, respectivement $V_{j,v(q)}$, représentent les vecteurs des indices calculées sur la *k*-ème partie horizontale $I_{i,h(k)}$ de l'image $\boldsymbol{I_i}$ et respectivement la *q*-ème partie verticale $I_{j,v(q)}$ de l'image $\boldsymbol{I_j}$. Une caractéristique importante d'un système de recherche concerne la fonction de similarité entre les images (les parties d'images dans cette approche) c'est-à-dire la fonction $\boldsymbol{f}$ définie par f: (espace d'indice)$^2 \rightarrow [0,1]$. Le choix de cette fonction sera discutée dans la partie suivante.

3. Simulations

Dans cette section, nous proposons d'évaluer les performances de la méthode proposée en la comparant avec une méthode classique de division des images en plusieurs rectangles. La base de test à indexer contient 1115 images classées en 20 catégories sélectionnées à partir de la base COLUMBIA :
www.cs.washington.edu/research/imagedatabase/groundtruth/
Cette base généraliste comporte essentiellement des images de paysages naturels ou des lieux habités. Pour cette raison, nous avons choisi de diviser les images horizontalement et verticalement en **m=n=3** parties horizontales et verticales (ciel, objets, terre). En ce qui concerne la signature des images, nous avons travaillé uniquement sur l'attribut couleur. La quantification est faite dans l'espace colorimétrique HSV et est statique et uniforme sur les trois axes avec : 18 niveaux pour l'axe Hue (teinte), 3 niveaux pour l'axe Saturation et 3 niveaux pour l'axe Value (intensité). Ceci permet de quantifier l'image sous 162 couleurs différentes. Ce choix qui reste arbitraire peut être discutable mais a montré son efficacité dans plusieurs travaux [9],[4],[3]. Ainsi, chaque image $\boldsymbol{I_i}$ est représentée

par un vecteur de dimension 972 : soit le nombre des couleurs (=162) multiplié par le nombre d'histogrammes (=6). La similarité entre deux sous-images horizontales $V_{i,h(k)}$ et $V_{j,h(k)}$ extraites respectivement des images $\boldsymbol{I_i}$ et $\boldsymbol{I_j}$ est calculée par la fonction [8] :

$$f(V_{i,h(k)}, V_{j,h(k)}) = \frac{\sum_{l=1}^{c} \min(p^l_{i,h(k)}, p^l_{j,h(k)})}{\min(\sum_{l=1}^{c} p^l_{i,h(k)}, \sum_{l=1}^{c} p^l_{j,h(k)})} \quad (2)$$

Cette même équation peut être utilisée pour les parties verticales de l'image permettant ainsi de quantifier la similarité entre $V_{j,v(q)}$ et $V_{j,v(q)}$:

$$f(V_{i,v(q)}, V_{j,v(q)}) = \frac{\sum_{l=1}^{c} \min(p^l_{i,v(q)}, p^l_{j,v(q)})}{\min(\sum_{l=1}^{c} p^l_{i,v(q)}, \sum_{l=1}^{c} p^l_{j,v(q)})} \quad (3)$$

Dans les équations (2) et (3), $p^l_{i,h(k)}$ (resp $p^l_{i,v(q)}$.) représente la proportion de la couleur l dans la sous-image $I_{i,h(k)}$ (resp. $I_{i,v(q)}$). Ainsi le vecteur $V_{i,h(k)}$ (resp. $V_{i,v(q)}$) représente l'histogramme de la sous-image $I_{i,h(k)}$ (resp. $I_{i,v(q)}$). Cette fonction de similarité varie entre 0 et 1. Plus les deux sous-images sont similaires, plus f est faible ($f(V_{i,h(k)}, V_{i,h(k)})=f(V_{i,v(q)}, V_{i,v(q)})=0$).

Nous avons testé cette approche sur la base généraliste COMLUBIA en utilisant, pour mesurer la similarité entre les images, le système d'équations (1), (2) et (3) et en ajustant les paramètres de la façon suivante :

m=n=3, c=162, α=0.5 et β=0.5.

Les figures 3 et 4 montrent quelques résultats des tests, où l'on peut constater facilement que la division de l'image en sous-images horizontalement et verticalement améliore la performance de la recherche. La figure 3 illustre les résultats obtenus pour deux requêtes respectives (image en haut à gauche de chaque planche d'images. On compare respectivement de manière qualitative les images retrouvées par le système basé sur la division en 9 rectangles (images de la planche a) et la grille des histogrammes (images de la planche b). De manière visuelle, on constate que les images retrouvées par le système basé sur la grille des histogrammes est beaucoup plus robuste et permet de retrouver un nombre plus important d'images de la même catégorie. Ceci est confirmé par les courbes ^précision / rappel présentées à la figure 4 pour trois catégories respectives de la base d'images.

Ces résultats montrent clairement que la grille des histogrammes est plus discriminante que le module de sous-images non-interférées, malgré la non-pondération (α=0.5, β=0.5) afin de ne pas passer outre le choix de la division qui est généralement évincée dans les autres méthodes.

4. Conclusion

Dans cet article, nous avons présenté une méthode qui augmente l'efficacité et la précision des systèmes d'indexation et de recherche d'images couleur par le contenu basés sur les histogrammes spatiaux. Cette technique consiste à diviser l'image en plusieurs blocs rectangulaires interférés et à en calculer leurs histogrammes couleur respectifs. Ceci permet d'introduire des informations spatiales avec peu de contraintes dans les processus d'indexation et de recherche. Nous augmentons ainsi la performance de la recherche en réduisant en même temps le coût de stockage et le temps de calcul par rapport aux méthodes classiques dont le choix concernant le découpage reste arbitraire

Optimiser les coefficients de pondération et le nombre de division en fonction du type d'image à indexer constituent les perspectives de ce travail.

Références

[1] Arnold W.M and al. Content-Based Image Retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 22, N° 12. December 2000.

[2] R.Schettini and al,"A survey of methods for colour image indexing and retrieval in image databases", (R. Luo, L. MacDonald eds.), J. Wiley, 2001. (invited paper).

[3] Hum-Woo Yoo and al, "Visual information retrieval via content-based approach" Pattern Recoggnitio 35 (2002) 749-769

[4] J.R Smith and S.F.Chang Li,"Decoding image semantics using composite region templates", IEEE Workshop, CBAIVAL 1998.

[5] L.Cinque and all. Color-based retrieval using spatial-chromatic histograms. Image and Vision Computing 19(2001) pp979-989.

[6] Seong-Oshim and al,"Image Indexing by modified Color co-occurrence matrix", ICIP 2003

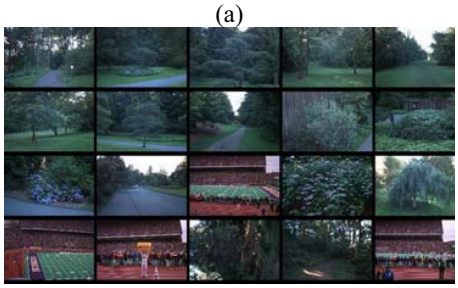
[7] Gong Y,"Image Indexing and retrieval using color histogram", Multimedia Tools and Applications, Vol 2, pp 133-156, 1996

[8] Jérôme FOURNIER, Indexation d'images par le contenu et recherché interactive dans les bases généralistes. Thèse : Université de Cergy -Pontoise 2002.

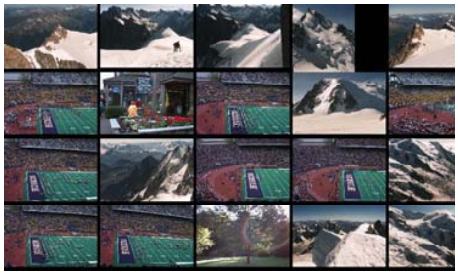
[9] J.R Smith and S.F.Chang, "A fully automated content-based image query system", ACMMM 1996, pp 87-98

[10]A.Trémeau, C.Fernandez-Maloigne, P.Bonton, "Image couleur :de l'acquisition au traitement";Editions DUNOD 2004.

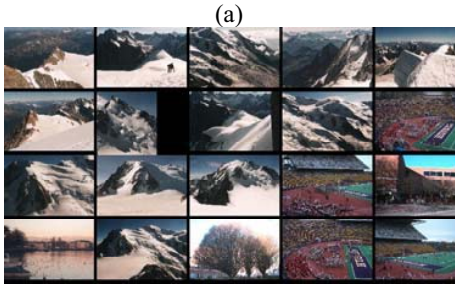
(a)



(b)

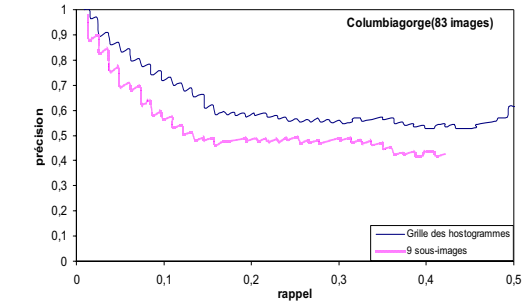


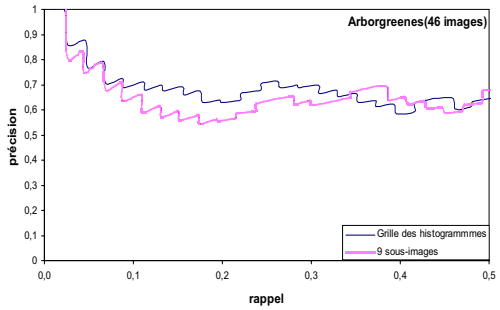
(a)



(b)

FIG 3 : Images résultats d'une requête (image en haut à gauche) donnée par les méthodes division en 9 rectangles (a) et la grille des histogrammes (b).





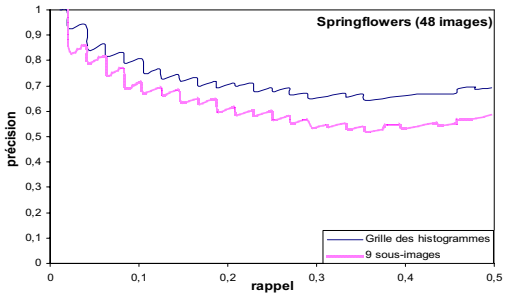


FIG 4 : Courbes précision / rappel pour trois catégories d'images dans la base COLUMBIA

Une approche de type filtrage pour la sélection de variables. Application à la reconnaissance automatique de poissons

Dahbia Semani, Carl Frélicot, Pierre Courtellemont
Laboratoire d'Informatique – Image – Interaction (UPRES EA 2118),
Université de La Rochelle, avenue M. Crépeau, 17042 La Rochelle Cedex 1, France
{dahbia.semani,carl.frelicot,pierre.courtellemont}@univ-lr.fr

Résumé – Cet article aborde le problème de la sélection de variables dans le cadre de la classification supervisée. Nous présentons un nouveau critère permettant de mesurer la pertinence d'un sous-ensemble de variables. Ce critère repose sur une mesure d'ambiguïté fondée sur la combinaison d'étiquettes représentant le degré d'appartenance aux classes en présence. Des tests sont menés sur des jeux de données réels issus de la littérature. L'application traitée concerne un problème réel de reconnaissance des formes, à savoir la reconnaissance d'objets dans les séquences vidéo.

Abstract – This paper addresses the feature selection problem for supervised classification. Feature selection methods are based on a selection algorithm and a criterion function assessing how effective feature subsets are. We propose an ambiguity measure that allows to define a new evaluation criterion. It is based on a combination of labels representing the degree of typicality to the classes. The new criterion is compared to others found in the literature on various real data sets. It is validated on a real object recognition problem.

1 Introduction

La sélection de variables joue un rôle très important en classification lorsqu'un grand nombre p de variables sont disponibles, certaines pouvant être peu significatives, corrélées ou non pertinentes au regard de l'application considérée [4]. Elle consiste à sélectionner un sous-ensemble de q variables ($q < p$) sans que les performances de la règle de classement diminuent trop voire même augmentent. La sélection permet également de faciliter l'étape d'apprentissage et de réduire la complexité des algorithmes ainsi que les temps de calcul.

Une méthode de sélection repose principalement sur un algorithme de recherche et un critère d'évaluation pour mesurer la pertinence des sous-ensembles potentiels de variables. Nous nous intéressons aux critères d'évaluation. Ainsi, nous proposons un nouveau critère fondé sur une mesure d'ambiguïté. Cette mesure repose sur la combinaison d'étiquettes représentant le degré de spécificité ou d'appartenance des données aux classes en présence. Des opérateurs d'agrégation sont utilisés pour la combinaison de ces étiquettes.

L'article est organisé comme suit : Un bref état de l'art sur les algorithmes de sélection et les critères d'évaluation est dressé à la section 2. Le critère proposé ainsi que sa validation sur des jeux de données réels issus de la littérature sont présentés à la section 3. La section 4 est consacrée à la description du système de reconnaissance automatique de poissons. Ce système constitue un cadre d'application de notre approche de sélection de variables. Les résultats de sélection et les taux de reconnaissance sont présentés à la section 4.1.

2 État de l'art

Algorithmes de recherche

Il existe divers algorithmes de recherche et études comparant leurs avantages et inconvénients [3]. Les méthodes les plus populaires procèdent par ajouts et/ou suppressions séquentiels de variables. Elles évitent la recherche exhaustive du meilleur sous-ensemble et conduisent à une solution sous-optimale. Nous utilisons l'algorithme SFFS (*Sequential Forward Floating Search*) [7] permettant à chaque pas, d'ajouter une variable et d'en supprimer plusieurs tant que le sous-ensemble résultant améliore le critère d'évaluation. Il est considéré comme la méthode sous-optimale la plus efficace [3]. Les deux étapes (ajout / suppression) de l'algorithme sont alternées jusqu'à ce qu'une condition d'arrêt soit vérifiée. Parmi celles-ci, citons une borne sur q ou un seuil sur la valeur du critère d'évaluation.

Critères d'évaluation

Deux approches sont couramment utilisées pour évaluer la pertinence d'un sous-ensemble de variables sélectionnées [5] : l'approche de type *filtrage* (*filter approach*) et celle de type « *enveloppante* » (*wrapper approach*). Dans la première, les critères sont fondés uniquement sur les données et sont donc totalement indépendants du discriminateur utilisé. Les variables sont alors filtrées avant le processus d'apprentissage et de classification. Parmi les différentes fonctions utilisées, citons celles fondées sur des mesures de distance probabilistes (ex : Mahalanobis, Battacharyya), d'information (ex : entropie) ou de dépendance (ex : corrélation, information mutuelle). La seconde approche tient compte de la règle de classement dans le calcul du critère d'évaluation. Celui-ci est simplement la probabilité d'erreur estimée sur l'ensemble des données.

3 Un nouveau critère d'évaluation

Étiquetage

Considérons un point $\mathbf{x} = (x_1, ..., x_p)^t$ dans un espace de représentation de dimension p et un ensemble de c classes $\omega = \{\omega_1, ..., \omega_c\}$. On peut lui associer un vecteur d'étiquettes à l'aide d'une fonction : $\Re^p \rightarrow [0,1]^c$, $\mathbf{x} \mapsto \mu(\mathbf{x}) = (\mu_1(\mathbf{x}), ..., \mu_c(\mathbf{x}))^t$ où $\mu_i(\mathbf{x})$ représente le degré d'appartenance de $\mathbf{x}$ à la classe ω_i, par exemple :

$$\mu_i(\mathbf{x}) = (1 + d(\mathbf{x}, \omega_i))^{-1} \quad (1)$$

où $d(\mathbf{x}, \mathbf{p}_i)$ est une distance (ex : distance de Mahalanobis) entre $\mathbf{x}$ et le vecteur moyenne de la classe ω_i.

La mesure d'ambiguïté

En classification, une variable est d'autant moins discriminante que les projections des classes sur cette variable se chevauchent. Le chevauchement est révélateur d'une certaine ambiguïté. Nous voulons quantifier l'ambiguïté entre les classes en combinant les étiquettes μ_i. Pour cela, nous avons choisi d'utiliser les normes (et conormes) triangulaires ou *t-normes* (et *t-conormes*) qui sont des opérateurs de combinaison tout à fait adaptés à notre problème ; le lecteur intéressé peut en trouver une synthèse dans (Klir et Yuan 1995). Dans ce qui suit, $\top$ désigne une t-norme arbitraire et $\perp$ sa t-conorme duale.

Dans [8] nous avons introduit un nouvel opérateur d'agrégation dans le cadre de la classification supervisée avec double option de rejet et en particulier option de rejet d'ambiguïté. Nous l'avons baptisé *OU-2 flou*, noté $\overset{2}{\perp}$, défini par :

$$\overset{2}{\underset{i=1,c}{\perp}} \mu_i = \underset{i=1,c}{\top} \left(\underset{j=1,c;j\neq i}{\perp} \mu_j \right) \quad (2)$$

Nous avons également montré que dans le cas où la t-norme utilisée est le *min*, $\overset{2}{\perp}_{i=1,c} \mu_i$ est égal au deuxième plus grand des μ_i. Cet opérateur possède plusieurs propriétés mathématiques dont certaines découlent naturellement de celles des t-normes et t-conormes (ex : bornes, monotonie, continuité, symétrie, etc.) [8].

Un moyen assez naturel de mesurer l'ambiguïté entre les classes est d'effectuer le rapport entre le *deuxième plus grand* et le *plus grand* des μ_i. La mesure d'ambiguïté que nous proposons généralise ce rapport :

$$A(\mathbf{x}) = \overset{2}{\underset{i=1,c}{\perp}} \mu_i(\mathbf{x}) \Big/ \underset{i=1,c}{\perp} \mu_i(\mathbf{x}) \quad (3)$$

Le critère d'évaluation proposé

Nous proposons d'utiliser la mesure d'ambiguïté (3) pour définir le nouveau critère d'évaluation d'un sous-ensemble S_q de q variables, suivant :

$$J_A(S_q) = \sum_{\mathbf{x}} A^{[q]}(\mathbf{x}) \quad (4)$$

où l'exposant $^{[q]}$ indique que la mesure d'ambiguïté est définie à partir d'étiquettes $\mu_i^{[q]}$ représentant le degré d'appartenance, donné par exemple par (1), du point $\mathbf{x}$ à la classe ω_i dans l'espace $\mathbb{R}^q$. Il s'agit alors de sélectionner l'ensemble des q variables parmi les p d'origine qui minimisent le critère $J_A(S_q)$.

Validation du critère d'évaluation

Nous avons associé le critère J_A à l'algorithme SFFS pour définir une méthode de sélection, baptisée *SFFS–fOU2*. Les normes standard et d'Hamacher avec $\gamma = 0$ et $\gamma = 1$ ont été utilisées dans la définition de la mesure J_A. Bien que les résultats soient comparables, ceux reportés correspondent à chaque fois aux normes ayant donné les meilleures performances.

Nous avons testé la méthode *SFFS–fOU2* sur des jeux de données standard issus de la base de données UCI : *Iris*, *Pima Indian* et *Breast Cancer Wisconsin*. La méthode de sélection a permit de réduire, pour tous les jeux, le nombre de variables d'au moins la moitié. La comparaison des taux de bon classement obtenus avec les p variables d'origine et les q variables sélectionnées est faite pour deux discriminateurs classiques : la règle des *k-Plus Proches Voisins* (k-PPV) et la règle de *Bayes Quadratique*, sous hypothèse gaussienne (BQ). Rappelons que la règle BQ correspond à la règle du *Maximum A Posteriori* (MAP) en considérant une matrice de covariance Σ_i propre à chaque classe ω_i. L'estimation des taux a été réalisée selon une procédure 10-CV (*Validation Croisée*). Pour réduire le biais dû au caractère aléatoire de la construction des ensembles test et apprentissage par la procédure 10-CV, cette dernière est répétée en réalisant 10 essais indépendants. Nous disposons donc de dix taux estimés, nous avons reporté les intervalles de confiance à 95% sur la moyenne de ces taux dans le tableau 1. Les résultats reportés montrent que la méthode proposée est capable de sélectionner les variables pertinentes puisque aucune baisse des performances n'a été enregistrée. Les performances sont même augmentées dans certains cas.

Notre critère a été comparé avec d'autres critères issus de la littérature et appartenant à l'approche *filtrage* : la distance de Mahalanobis, l'entropie [6] et la distance floue [1]. Les résultats obtenus par notre critère et la distance de Mahalanobis sont comparables, quel que soit le discriminateur. Les résultats obtenus par notre critère sont cependant meilleurs que ceux obtenus par la mesure d'entropie et la distance floue. Nous avons reporté, à titre d'exemple, dans le tableau 2 les taux de bon classement obtenus en utilisant la règle BQ.

TAB. 1 – Moyennes et intervalles de confiance des taux de bon classement obtenus avec la procédure *10-VC*.

Jeux	Critère	BQ (%)	k-PPV (%)
Iris	$p = 4$	97.20±0.20	96.26±0.63
	$J_A(q = 2)$	97.13±0.32	96.27±0.56
Pima	$p = 8$	74.06±0.51	74.60±0.57
	$J_A(q = 3)$	75.57±0.30 +	75.75±0.51 +
Breast	$p = 9$	95.17±0.16	97.11±0.24
	$J_A(q = 3)$	96.27±0.11 +	96.82±0.20

Amélioration (+) ou dégradation (−) statistiquement significative (95%)

TAB. 2 – Taux de bon classement obtenus dans le cas de la règle BQ.

Jeux	p **variables**	J_A	$Maha$	DF	E
$Iris$	97.20±0.20	97.13±0.32	97.40±0.15	94.67±0.42−	97.07±0.33
$Pima$	74.06±0.51	75.57±0.30+	75.40±0.17+	74.39±0.25	74.73±0.22+
$Breast$	95.17±0.16	96.27±0.11+	95.99±0.18+	95.23±0.07	94.73±0.17−

Amélioration (+) ou dégradation (−) statistiquement significative (95%)

4 Application à la reconnaissance automatique de poissons

Nous décrivons, dans cette section, un système de reconnaissance automatique de poissons évoluant dans un aquarium. Ce travail s'inscrit dans le cadre du projet Aqu@thèque[1] dont l'objectif est de permettre au visiteur d'un aquarium de désigner un poisson sur un écran tactile diffusant les images acquises par une caméra vidéo, distante et fixe, orientée vers un bassin de l'aquarium. Le système de reconnaissance doit identifier automatiquement, et en temps réel, l'espèce correspondant au poisson désigné. Des informations multimedia (textuelles, graphiques, audio ou vidéo) à caractère pédagogique sur le poisson sont alors mises à disposition du visiteur. Le système de reconnaissance comporte les étapes suivantes :

1. **Segmentation** : Correspond à effectuer une partition de l'image en régions susceptibles d'être des poissons ou parties de poissons. L'objectif principal est donc de séparer les objets en mouvement du fond de la scène supposé statique. Pour cela, nous utilisons la méthode de *détection de changement temporels* d'Elgammal et al. [2].
2. **Extraction d'attributs** : Les régions extraites des images sont caractérisées par un ensemble d'attributs (ou variables) pour permettre leur identification. Différents attributs peuvent être calculés sur une région [9]. Nous avons extrait 85 variables réparties sur différents groupes [8] : attributs géométriques, photométriques, texture, moments couleur, moments de Hu, paramètres de déplacement. Le tableau 3 résume le nombre de variables par groupe.
3. **Sélection d'attributs** : Cette étape permet de ne retenir, parmi tous les attributs extraits, que ceux qui sont discriminants au regard de notre application.
4. **Reconnaissance** : L'identification de l'espèce d'un poisson se fait grâce à une procédure de classification supervisée, c'est-à-dire que seules les classes apprises seront reconnues.

4.1 Résultats expérimentaux

La base d'apprentissage
Un bassin de l'Aquarium de la Rochelle comprenant 12 espèces $\Omega = \{\omega_1, \omega_2, ..., \omega_{12}\}$ a été filmé pour obtenir les séquences d'apprentissage. La base d'apprentissage a été construite en étiquetant, suivant les 12 classes, 1900 régions extraites des images segmentées. Les régions sont décrites par les 85 attributs.

Sélection d'attributs
L'étape de la sélection d'attributs est appliquée à l'ensemble des données centrées réduites de la base d'apprentissage. Nous avons sélectionné les attributs les plus pertinents en utilisant deux approches différentes :

1. *Sélection globale* (*Sélection # 1*) : L'utilisation de la méthode *SFFS–fOU2*, permet de sélectionner 24 attributs, ce qui représente une réduction significative du nombre de variables d'origine (71.77% de réduction, voir le tableau 3).
2. *Sélection hiérarchique* (*Sélection # 2*) : En analysant les résultats de la classification obtenus avec les $p = 85$ variables d'origine, nous avons remarqué que certaines classes sont très confondues induisant ainsi une baisse des performances. La confusion provient des espèces appartenant en réalité à une même espèce de poissons. La ressemblance (visuelle) entre des espèces complètement différentes est également source de confusion. Il est alors raisonnable de penser qu'un seul ensemble de variables est insuffisant pour discriminer toutes les classes. Un ensemble de variables capable de discriminer une classe des autres n'est pas forcement capable de discriminer une autre classe des autres.
   Nous avons décider de fusionner certaines classes selon le principe suivant : 1) une première classification des objets, par rapport aux différentes espèces présentent dans le bassin, est effectuée dans l'espace des variables d'origine ($p = 85$ variables), 2) les espèces qui se confondent avec un taux d'erreur de classement supérieur à 10% sont fusionnées. Les classes ayant été discriminées avec un taux de reconnaissance supérieur à 70% sont gardées comme des classes singleton. Finalement, nous obtenons 19 nouvelles classes contenant des classes *fusionnées* et des classes *singleton*. Par exemple, si la classe ω_i est confondue avec la classe ω_j qui elle même se confond avec la classe ω_k, nous obtenons les deux classes *fusionnées* $\omega'_l = \{\omega_i, \omega_j\}$ et $\omega'_m = \{\omega_j, \omega_k\}$.

   La *sélection hiérarchique* s'effectue en deux étapes :
   – *Étape 1 :* nous considérons l'ensemble des 19 nouvelles classes et nous appliquons d'une manière globale la méthode *SFFS–fOU2* comme dans *Selection # 1*. Cette étape permet de réduire les 85 attributs en 27 attributs (voir tableau 3).

[1] Le projet Aqu@thèque est mené par le laboratoire L3i de l'Université de La Rochelle.

TAB. 3 – Résumé des attributs.

Nombre da variables	Avant sélection	*Sélection # 1*	*Sélection # 2*	
			Étape 1	Étape 2
Géométriques	10	2	3	5
Photométriques	37	11	10	30
Texture	18	7	9	15
Moments	13	4	4	7
Moments de Hu	4	0	1	3
Déplacement	3	0	0	0
Total	85	24	27	60
Taux de réduction		71.77%	68.23%	30%

– *Étape 2 :* les objets sont classés dans l'une des 19 nouvelles classes en utilisant le sous-ensemble des 27 variables sélectionnées lors de l'Étape 1. Les objets classés dans une classe fusionnée doivent être séparés en utilisant le sous-ensemble de variables approprié. L'objectif de cette étape est de sélectionner pour chaque classe fusionnée $\omega'_l = \{\omega_i, \omega_j\}$, le sous-ensemble qui discrimine au mieux les deux classes ω_i et ω_j. Une nouvelle sélection est donc effectuer par la méthode *SFFS–fOU2* au niveau de chaque classe fusionnée. Ceci permet d'obtenir différents sous-ensembles de variables. Au final, 60 variables ont été retenues dans cette étape, représentant un taux de réduction d'environ 30% (voir tableau 3).

Les attributs sélectionnés par les deux méthodes *Sélection # 1* et *Selection # 2* (voir tableau 3) sont essentiellement des moments de couleur (moments d'ordre 1 et 2, moments de chromaticité, moments généralisés couleur) et des attributs de texture (couleur et niveaux de gris).

Résultats de classification

Les taux de bon classement obtenus avec les p variables d'origine et les variables sélectionnées par les deux méthodes *Sélection # 1* et *Sélection # 2*, sont présentés dans le tableau 4. Les taux de bon classement sont estimés par une procédure *Holdout* [4] pour deux discriminateurs : BQ et MDAG. La règle de classement MDAG intègre deux options de *rejet* suivant la stratégie « *mélange d'abord généralisée* » que nous avons proposé dans [8] : le *rejet de distance* et le *rejet d'ambiguïté*. Le premier type de rejet permet de n'associer un point $\mathbf{x}$ à aucune des classes ; il concerne les $\mathbf{x}$ situés en général loin de toute classe. Le second consiste à associer $\mathbf{x}$ à plusieurs ou à toutes les classes ; il concerne en général les $\mathbf{x}$ se projetant dans des régions situées entre deux classes ou plus. L'introduction du rejet permet de réduire le risque de mauvais classement.

Les performances obtenues avec les attributs sélectionnés par la méthode de *sélection globale* sont meilleures que celles obtenues avec les variables d'origine, quel que soit le discriminateur. Des meilleures performances sont obtenues en utilisant la méthode de *sélection hiérarchique*. Le taux de bon classement est, dans ce cas, supérieur à 90% pour les deux discriminateurs.

TAB. 4 – Taux de bon classement obtenus avec les variables d'origine et les variables sélectionnées.

Méthode	BQ (%)	MDAG (%)
Avant sélection ($p = 85$)	63.73	73.79
Sélection # 1 ($q = 24$)	76.51	77.96
Sélection # 2 ($q = 60$)	**90.87**	**92.52**

5 Conclusion

Nous avons présenté, dans cet article, un nouveau critère permettant d'évaluer la pertinence d'un (sous-)ensemble de variables. Le critère proposé peut être associé à n'importe quel discriminateur. Ses performances ont été comparées avec d'autres critères issus de la littérature. Les tests menés sur des jeux de données réels ont montré que le critère proposé est capable de sélectionner les variables pertinentes et d'augmenter dans la plupart des cas les taux de bon classement. Nous avons validé la méthode de sélection de variables proposée sur un problème réel de reconnaissance des formes. L'application concerne un système de reconnaissance automatique de poissons évoluant dans un aquarium. La sélection de variables a permit une amélioration de plus de 14% du taux de reconnaissance.

Références

[1] T. E. Campos, I. Bloch, and R. M. Cesar Jr. *Feature Selection Based on Fuzzy Distances between Clusters : First Results on Simulated Data*. In : Lecture Notes in Comp. Sc. 2013 : Advances in Pattern Recognition, Springer-Verlag, 2001.

[2] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis. Background and foreground modeling using nonparametric kernel density for visual surveillance. *Proceedings of the IEEE*, 90(7) :1151–1163, July 2002.

[3] A. Jain and D. Zongker. Feature selection : Evaluation, application and small sample performance. *IEEE Trans. on PAMI*, 19(2) :153–158, 1997.

[4] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition : A review. *IEEE Trans. on PAMI*, 22(1) :4–37, 2000.

[5] P. Langley. Selection of relevant features in machine learning. *In AAAI Fall Symposium on Relevance*, pages 140–144, 1994.

[6] P. Mitra, C. A. Murthy, and S. K. Pal. Unsupervised feature selection using feature similarity. *IEEE Trans. on PAMI*, 24(3) :301–312, March 2002.

[7] P. Pudil, J. Novovicová, and J. Kittler. Floating search methods in feature selection. *Pattern Recognition Letters*, 15 :1119–1125, 1994.

[8] D. Semani. *Une méthode supervisée de sélection et de discrimination avec rejet. Application au projet Aqu@thèque*. Thèse de doctorat, Université de La Rochelle, Mai 2004.

[9] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press Inc., 1999.

Amélioration de la reconnaissance de partitions musicales par modélisation floue et indication des erreurs possibles

F. ROSSANT[1], I. BLOCH[2]

[1]ISEP, 28 rue Notre-Dame des Champs, 75006 Paris, [2]ENST CNRS UMR 5141, 46 rue Barrault, 75634 PARIS Cedex 13

florence.rossant@isep.fr , isabelle.bloch@enst.fr

Résumé – Nous proposons une méthode de reconnaissance de partitions musicales fondée sur la modélisation floue des informations extraites de la partition scannée et des règles de musique. L'objectif est de prendre en compte les ambiguïtés qui subsistent à l'issue de l'étape d'analyse individuelle des symboles, et d'obtenir une interprétation globale cohérente. Le formalisme proposé permet de modéliser l'imprécision sur la détection des symboles, la variabilité des polices, la souplesse des règles musicales, et ainsi de fiabiliser les résultats de reconnaissance. Nous montrerons qu'il permet également d'indiquer les erreurs potentielles, de manière à faciliter la correction manuelle.

Abstract *– We propose an OMR method based on fuzzy modeling of the information extracted from the scanned music score and of musical rules. The aim is to disambiguate the recognition hypotheses output by the individual symbol analysis process. Fuzzy modeling allows to account for imprecision in symbol detection, for typewriting variations, and for flexibility of rules. The reliability of the recognition is increased, and the results are also used in order to indicate the possible errors, and thus to facilitate the manual correction.*

1. Introduction

La recherche dans le domaine de la reconnaissance optique de la musique (OMR) a commencé dès les années 1970. Cependant, la difficulté d'extraction des symboles musicaux, due au haut degré de connectivité entre primitives aussi bien qu'aux défauts d'impression, ainsi que la complexité et la souplesse des règles d'écriture musicales [1], ne permettent pas encore d'obtenir des résultats de reconnaissance suffisamment fiables. Les efforts se sont concentrés depuis les années 1990 sur l'introduction d'informations structurelles. Cependant, les méthodes présentées se limitent souvent à la description des symboles [2], à la formalisation de règles graphiques strictes et locales [3,4], ou bien sont fondées sur des modèles probabilistes [5], nécessitant de nombreuses données d'apprentissage.

L'objet de notre recherche est de prendre en compte les sources d'incertitude et d'imprécision (variabilité des polices, défauts de segmentation), de modéliser et d'intégrer l'ensemble des règles d'interprétation de haut niveau, strictes ou souples, dans un processus global de décision, afin d'aboutir à une reconnaissance totalement cohérente de la partition. La théorie des ensembles flous et des possibilités offre un formalisme bien adapté à la modélisation et l'intégration de contraintes souples [7], ainsi qu'à la prise en compte de l'imprécision sur la forme et la localisation [8].

Une autre contribution de cet article est l'aide à la correction. Certains auteurs tentent de repérer des incohérences afin de tenter de corriger certains types d'erreurs (sur les durées des noires le plus souvent) par des post-traitements spécifiques [3,9]. Ici, nous utilisons directement certains résultats de la modélisation floue, exprimés sous forme de degrés de possibilité, et ajoutons d'autres critères portant sur la décomposition rythmique des mesures, afin d'indiquer les symboles ou groupes de symboles qui pourraient être erronés, sans restriction sur la nature de l'erreur, de manière à faciliter la vérification et la correction manuelle.

Nous débuterons par une brève présentation du système global [10]. Puis nous présenterons les éléments de la modélisation floue qui servent d'indicateurs pour la correction. Nous indiquerons ensuite la règle de décision globale, la méthode d'indication des erreurs, avant de conclure sur les performances obtenues en reconnaissance et en détection des erreurs.

2. Système global

En entrée du programme, nous avons l'image binaire (1 pour un pixel noir, 0 pour un pixel blanc) de la partition scannée à 300 dpi, ainsi que des informations globales, la clef, la métrique et la tonalité. Pour l'instant, le système ne traite que le cas monodique, et reconnaît les symboles nécessaires à la reconstitution de la mélodie.

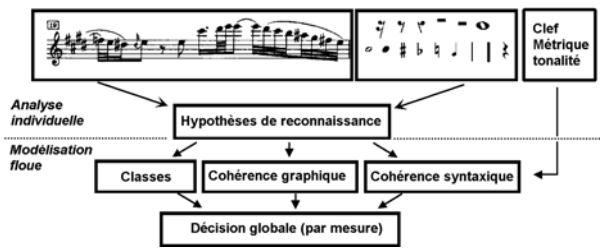


FIG. 1 : Système global

L'algorithme procède en deux étapes. Une première étape d'analyse [6], réalise la détection des objets, et, par corrélation avec des modèles de référence, propose pour chacun plusieurs hypothèses de reconnaissance. La théorie des ensembles flous et des possibilités permet dans une

seconde étape de combiner les informations de position et de corrélation fournies par la première, de modéliser les règles graphiques et syntaxiques de la musique, afin d'aboutir à une décision finale par optimisation de tous les critères.

3. Modélisation floue

3.1 Résultats initiaux exploités

L'étape d'analyse [6] fournit les positions (x_k,y_k) et les scores de corrélation $C^k(s)$ des objets s avec les modèles de référence M^k. Trois hypothèses de reconnaissance sont retenues, au maximum, si les scores de corrélation correspondants sont supérieurs à un seuil t_m (fixé à 0.3 dans nos expérimentations). On ajoute également la possibilité qu'il n'y ait pas de symbole ('-' dans les tableaux), lorsque le plus haut score de corrélation est inférieur au seuil de décision $t_d(k)$. Ces seuils ont été définis expérimentalement. Le seuil $t_d(k)$ est élevé lorsqu'on observe une grande probabilité de fausse détection pour la classe k, et plus faible lorsque le score de corrélation est très sensible aux variations de police de symboles. La figure 2 montre deux barres de mesure (a) et (b) extraites d'une même partition, et, en superposition, les hypothèses de reconnaissance. Le tableau 1 indique certains scores de corrélation obtenus.

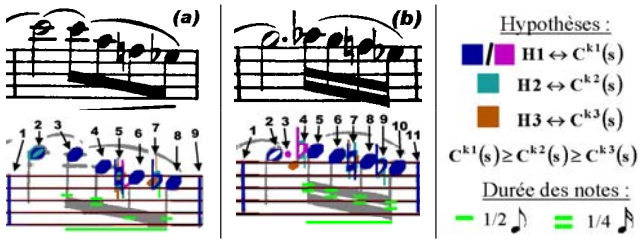


FIG. 2 : Image originale et hypothèses de reconnaissance

TAB. 1 : Scores de corrélation

(a)	2	5	7	(b)	3	4	7
H1	0.59	ou 0.56	0.72	H1	ou 0.74	ou 0.67	0.71
H2	0.38	0.49	0.54	H2		0.52	0.56
H3		0.42	0.43	H3		0.40	0.47

On voit sur cet exemple les limites de l'analyse individuelle des symboles : les scores de corrélation peuvent être très ambigus, et le plus haut ne correspond pas toujours à l'hypothèse correcte. Ces ambiguïtés sont dues aux défauts de segmentation et aux variations dans les polices de symboles inter et intra partition (voir par exemple les bécarres 5(a) et 7(b) ou les bémols 7(a) et 4(b)). Pour les résoudre, nous modélisons les classes de symboles par des distributions de possibilité apprises à partir des scores de corrélation, et nous modélisons les règles graphiques et syntaxiques de l'écriture musicale par des coefficients de compatibilité entre symboles.

3.2 Modélisation des classes de symboles

Les scores de corrélation obtenus indiquent la similarité entre les symboles de la partition et les modèles de classe. Nous définissons donc le degré de possibilité $\pi_k(s)$ d'appartenance de l'objet s à la classe k comme une fonction croissante du score de corrélation $C^k(s)$ (Fig. 3) :

$$\pi_k(s) = f_k\left(C^k(s)\right) \qquad (1)$$

Le paramètre D de la fonction f_k représente la zone d'incertitude et il est constant (toujours 0.3). En revanche, S_k est fonction de la ressemblance entre le modèle de référence M^k et les symboles de la partition, et il est déduit des scores de corrélation résultant de l'analyse individuelle. Soient n_k le nombre d'objets obtenant le plus haut score de corrélation pour la classe k, supérieur à $t_d(k)$, et $m(k)$ la moyenne de ces scores. On définit S_k par :

$$S_k = \frac{t_d(k) + n(k)m(k)}{n(k)+1} \qquad (2)$$

Lorsque le modèle de référence M^k ne correspond pas bien aux symboles de cette classe dans la partition, à cause des variations de police, alors $m(k)$ est proche de $t_d(k)$. Dans le cas contraire, il prend des valeurs élevées, translatant la distribution de possibilité vers la droite.

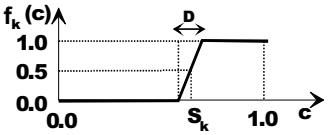


FIG. 3 : Distribution de possibilité de la classe k

Notons que la forme des distributions de possibilité n'a pas besoin d'être estimée de manière précise, et on constate expérimentalement une bonne robustesse par rapport à cette forme. L'essentiel est que ce n'est pas une fonction binaire, et qu'elle est croissante, c'est-à-dire que le degré de possibilité d'appartenance à la classe k est d'autant plus élevé que le score de corrélation l'est.

Le tableau 2 montre les résultats obtenus pour les deux barres de mesures de la figure 2. Les degrés de possibilité présentent généralement moins d'ambiguïté que les scores de corrélation du tableau 1. Notons par ailleurs que l'ordre peut changer, par exemple, l'objet 5 mesure (a) obtient un degré de possibilité maximal pour la classe bécarre, alors que le degré de possibilité pour la classe bémol est maintenant nul.

TAB. 2 : Degrés de possibilité d'appartenance aux classes

(a)	2	5	7	(b)	3	4	7
H1	0.35	ou 0.00	0.40	H1	ou 0.43	ou 0.28	0.68
H2	0.00	0.13	0.25	H2		0.20	0.35
H3		0.00	0.00	H3		0.00	0.00

3.3 Modélisation des règles graphiques

Elles permettent d'exprimer la cohérence sur les positions relatives des symboles. Aux règles déjà définies [10], concernant les relations entre les altérations et les notes, les points et les notes, a été ajoutée la définition d'un coefficient de compatibilité entre deux symboles voisins quelconques. Cette règle exprime que deux symboles ne devraient pas se superposer. On estime les positions des boîtes englobantes de deux symboles successifs s_n et s_m ($m>n$), à partir des dimensions typiques des deux classes k_n et k_m considérées (Fig. 4a). Soient Δl et Δh les décalages entre les bords horizontaux et verticaux. Les valeurs admissibles dans les deux directions sont définies par la fonction f (Fig. 4b), et le coefficient de compatibilité entre s_n et s_m par :

$$C_p\left(s_n^{k_n}, s_m^{k_m}\right) = Max\left[f(\Delta l), f(\Delta h)\right] \qquad (3)$$

La fonction f autorise des décalages négatifs, afin de prendre en compte l'imprécision de l'estimation, et afin de traiter les partitions de forte densité.

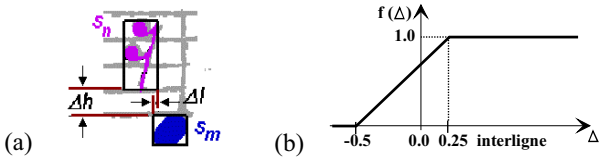


FIG. 4 : Compatibilité graphique entre deux symboles

Le coefficient de compatibilité étant calculé pour toutes les paires de symboles voisins, on peut maintenant en déduire pour chaque symbole son coefficient de compatibilité avec tous ses voisins, antérieurs et postérieurs :

$$C_p\left(s_n^{k_n}\right)=\left[\min_{j<n} C_p\left(s_n^{k_n}, s_j^{k_j}\right)\right]\cdot\left[\min_{l>n} C_p\left(s_n^{k_n}, s_l^{k_l}\right)\right] \quad (4)$$

Cette règle s'est révélée très efficace en cas de partitions de forte densité, car elle permet de ne pas éliminer des configurations qui théoriquement seraient inacceptables. Le bémol 9 de la mesure (b) par exemple, qui se superpose légèrement à la note précédente, obtient un coefficient de compatibilité graphique de 0.5.

3.4 Modélisation des règles syntaxiques

Nous appelons règles syntaxiques les règles musicales qui concernent les altérations et la tonalité [10], ou la métrique. Concernant la métrique, une règle stricte est celle du nombre de temps par mesure. Les méthodes de groupement des notes en temps, multiple ou fraction de temps, permettant de faciliter la lecture rythmique, sont au contraire des règles d'usage souples. En effet, il n'y a pas une décomposition unique pour une mesure donnée, et ces règles peuvent être relâchées, pour indiquer un phrasé par exemple.

L'étape première d'analyse individuelle des symboles a fourni pour chaque note noire sa durée [6], qui va maintenant être confirmée ou révisée en considérant les groupes de notes. Ceux-ci sont extraits par un algorithme de croissance de région. Lorsque la durée du groupe n'est pas usuelle, deux nouvelles hypothèses au maximum sont générées, augmentant et/ou diminuant la durée du groupe g, en changeant un minimum de durées dans le groupe. Le degré de possibilité affecté à chaque hypothèse H_l est fonction du nombre de corrections de durée de note $l(g)$ et du nombre total de notes $L(g)$ dans le groupe g :

$$C_l^{H_l}(g)=1.0-l(g)/L(g) \quad (5)$$

Dans la mesure (a) de la figure 2, le groupe de 4 notes a une durée inhabituelle de 1,75 temps (1/2, 1/4, 1/2, 1/2). Les deux hypothèses générées sont donc (1/2, 1/2, 1/2, 1/2), ou (1/2, 1/4, 1/8, 1/8), qui mènent le groupe respectivement à 2 temps ou 1 temps. Les degrés de possibilité sont 1.0 pour la configuration initiale, 0.75 et 0.5 pour les propositions de correction. Ces degrés n'évaluent pas les hypothèses par rapport aux règles d'usage de groupement, mais par rapport à l'interprétation initiale des durées, qui est supposée fiable.

4. Fusion et décision

La décision globale est réalisée mesure par mesure, en évaluant toutes les combinaisons possibles d'hypothèses. Pour chaque combinaison de symboles, indicée par j, il peut y avoir plusieurs combinaisons de durée H_l. Toutes les combinaisons de symboles qui contiennent au moins un coefficient de compatibilité graphique nul sont éliminées. Pour les autres, on moyenne les degrés de possibilité d'appartenance aux classes (1), les coefficients de compatibilité graphique (4) et les coefficients de compatibilité syntaxiques portant sur des altérations [10]. Le coefficient résultant $Conf_s(j)$ indique la cohérence mutuelle des symboles. Les différentes combinaisons de durées de notes sont ensuite évaluées en fusionnant les coefficients (5) en un coefficient $Conf_l(j,H_l)$ [10], exprimant le degré de possibilité des durées des notes pour la configuration H_l. On obtient le degré de possibilité final par multiplication de $Conf_s(j)$ et $Conf_l(j,H_l)$, de sorte que les critères sur les symboles et sur les durées de notes doivent être simultanément respectés. L'algorithme retient la configuration maximisant ce degré de possibilité, avec priorité aux configurations qui satisfont à la contrainte stricte du nombre de temps par mesure.

5. Indication des erreurs

Les erreurs sont de quatre sortes : symbole ajouté, confusion, symbole manquant, erreur de durée de note. Nous proposons d'analyser la solution retenue par l'algorithme de décision, afin d'indiquer à l'utilisateur les symboles potentiellement erronés. Les critères utilisés sont le degré de possibilité d'appartenance aux classes, la compatibilité graphique, la décomposition rythmique de la mesure.

5.1 Symboles et relations graphiques

Considérons de nouveau chaque symbole s_n^k, classé en classe k par l'algorithme de décision, avec un degré de possibilité $\pi_k(s_n^k)$ d'appartenance à la classe k et une compatibilité de position $C_p\left(s_n^k\right)$ avec les autres symboles de la mesure. Des valeurs faibles peuvent être révélatrices d'une erreur de classification du symbole s_n. La règle suivante est donc appliquée : si $\pi_k(s_n^k)<t_s^k$ et $C_p\left(s_n^k\right)<t_g^k$, alors le symbole s_n est indiqué comme potentiellement faux. Les seuils t_s^k et t_g^k ont été déterminés par apprentissage, de façon à minimiser sur chaque classe k la somme des occurrences « non détection » et « fausses alarmes ». La base d'apprentissage comprend la moitié des images de notre base de données et est représentative des différents types d'édition.

Il n'y a pas de règle permettant, sur la base des hypothèses non retenues par l'algorithme de décision, d'indiquer les symboles manquants. En effet une règle du type $\pi_k(s_n)>t'^k_s$ ou $C_p\left(s_n^k\right)>t'^k_g$, conduit essentiellement à des fausses alarmes, les symboles manquants étant en fait généralement absents de l'ensemble des hypothèses. Cependant, les autres indicateurs permettent également de localiser grossièrement des symboles manquants, car les erreurs dans une mesure sont souvent corrélées.

5.2 Métrique et décomposition rythmique

Nous indiquons tout d'abord toutes les mesures qui ne satisfont pas à la contrainte stricte de métrique. Nous analysons ensuite la décomposition rythmique de la mesure. La durée minimale des groupes de notes dans la mesure nous permet de fixer un pas de découpage, égal à 1.0 ou 0.5 dans

une métrique binaire, 1.5 ou 0.5 dans une métrique ternaire. Les groupes de notes sont ensuite associés aux silences voisins, de manière à ce que la durée totale de chacune des associations soit égale à un multiple du pas. Celles qui ne satisfont pas à ce découpage idéal sont affectées d'un degré de possibilité nul. Pour toutes les autres, nous affectons un degré de possibilité égal à 0.5 pour les groupes peu courants mais possibles, et 1.0 pour les groupes tout à fait possibles. Les critères portent sur la répartition des durées et sur le type de métrique. Par exemple le degré de possibilité d'un groupe croche pointée/double/croche est de 1.0 dans une métrique ternaire, 0.5 dans une métrique binaire avec un pas de 0.5.

6. Résultats et conclusion

Le taux de reconnaissance des symboles, évalué sur une centaine de partitions monodiques (environ 42500 symboles) est maintenant de 98.6%. Ce résultat a été amélioré par rapport à [10] (+0.2%), malgré l'introduction dans la base de partitions plus difficiles, et cela en particulier grâce au compléments sur les règles graphiques. 98.4% des notes (noires, croches, etc.) ont une durée correcte. De plus, notre méthode permet de résoudre des cas pour lesquels SmartScore [11] échoue (Fig. 5). Les ambiguïtés sur la classe des symboles, les symboles connectés, et les durées des croches sont bien résolues. Les taux de reconnaissance sur les deux pages complètes sont de 98.7% pour notre programme, 92.0% pour SmartScore.

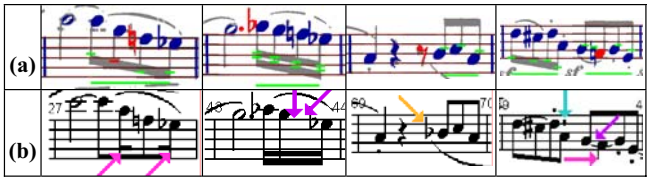


FIG. 5 : Résultats de reconnaissance
(a) notre méthode : 0 erreur, (b) Smartscore : 8 erreurs

Sur la base de test, 76% des confusions et ajouts sont bien indiqués, avec 30% des indications qui sont de fausses alarmes (0.4% de l'ensemble des symboles). 67% des erreurs de durée de note sont bien localisées : 40% directement par l'indication sur le groupe de notes, 27% par leur voisinage (mesure indiquée fausse ou symbole erroné dans la même mesure bien indiqué). Enfin, 64% des symboles manquants sont approximativement localisés par leur voisinage.

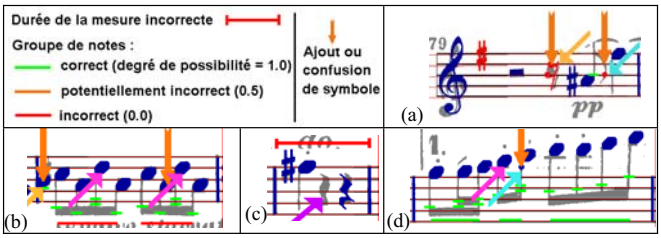


FIG. 6 : Résultats en détection

Sur la figure 6, on peut constater que les confusions et les ajouts de symboles sont bien indiqués (a)(b)(d) à l'exception d'une fausse alarme (croche en (b)); en (b), les deux erreurs de durée sont bien indiquées par l'indication de groupe erroné, et en (c), le soupir manquant est localisé par l'indication de mesure fausse; enfin, en (d), l'erreur sur la durée de la croche ne peut être détectée via le groupe car celui-ci est parfaitement cohérent (croche pointée + double croche) ; mais la détection du point ajouté permet de repérer rapidement cette erreur.

Ces résultats montrent l'intérêt de la méthode proposée, qui permet d'aboutir à une bonne fiabilité tant au niveau de la reconnaissance que de l'indication des erreurs possibles. On peut penser que cette méthodologie permet également d'adapter l'algorithme de reconnaissance à la partition traitée. En effet, l'analyse des degrés de possibilité d'appartenance aux classes et des coefficients de compatibilité graphique obtenus sur les premières pages permet certainement d'affiner la règle de décision, et ainsi d'obtenir une reconnaissance plus fiable sur le reste de la partition. D'autre part, les critères utilisés pour la détection des erreurs de durée pourraient être également intégrés dans l'algorithme de décision, de manière à obtenir un degré de possibilité exprimant simultanément le degré de possibilité des groupes de notes par rapport à l'interprétation initiale et la cohérence de la répartition des durées dans la mesure. Des cas d'erreurs qui se compensent (figure 6(b) par exemple) pourraient ainsi être éliminés au profit de corrections de durées conduisant à une décomposition rythmique correcte et à la bonne solution.

Références

[1] D. Blostein, H. Baird, *A Critical Survey of Music Image Analysis.* Structured Document Image Analysis, Eds H.Baird et al., Springer, Verlag, pp 405-434, 1992.

[2] G. Watkins, *The Use of Fuzzy Graph Grammar for Recognising Noisy Two-Dimensional Images*. NAFIPS Conference , pp 415-419, 1996.

[3] B. Coüasnon, B. Rétif, *Using a Grammar for a Reliable Full Score Recognition System*. International Computer Music Conference, Banff, Canada, pp 187-194, 1995.

[4] H. Fahmy, D. Blostein, *A graph-rewriting paradigm for discrete relaxation : application to sheet-music recognition*. Int. Journal of Pattern Recognition and Artificial Intelligence, Vol. 12, No. 6, pp 763-799, 1998.

[5] M.V. Stückelberg, C. Pellegrini, M. Hilaro, *An Architecture for Musical Score Recognition using High-Level Domain Knowledge*. 4th ICDAR Conference, vol. 2, pp 813-818, 1997.

[6] F. Rossant, *A Global method for music symbol recognition in typeset music sheets*. Pattern Recognition Letters 23 (10), pp. 1129-1141, 2002.

[7] D.Dubois, H. Prade, *Fuzzy Sets and Systems : Theory and Applications*. Academic Press, New-York, 1980.

[8] I. Bloch, H. Maître, *Fusion of Image Information under Imprecision*, In B. Bouchon-Meunier, Ed. , Aggregation and Fusion of Imperfect Information, Series Studies in Fuzziness, Physica Verlag, Springer, pp 189-213,1997.

[9] D. Blostein, L. Haken, *Using diagram generation software to improve diagram recognition : a case study of music notation,* IEEE Trans. on PAMI, Vol.21, No.11, pp 1121-1135, 1999.

[10] F. Rossant, I. Bloch, *A fuzzy model for optical recognition of musical scores*. Fuzzy Sets & Systems 141, pp165-201, 2004.

[11]SmartScore 3.2 Pro Demo, http://www.musitek.com/

Combinaison de classifieurs pour la localisation de visage

R. BELAROUSSI L. PREVOST M. MILGRAM

[1] Université Pierre & Marie Curie, Laboratoire des Instruments & Systèmes d'Ile de France Groupe PARC

4 Place Jussieu, 75252 Paris Cedex 5, BC252

Résumé – Dans cette communication, nous présentons une méthode de localisation de visages dans des images couleurs combinant trois détecteurs respectivement anthropomorphique (détecteur basé sur un modèle d'apparence neuronal), géométrique (détecteur d'ellipse utilisant une Transformée de Hough Généralisée sur l'image des orientations de gradient) et colorimétrique (détecteur de teinte chair utilisant un seuillage dans l'espace CbCr). La combinaison linéaire de ces trois détecteurs produit une carte de probabilités. La localisation du visage dans l'image correspond au maximum absolu de cette carte. Nous montrons l'apport de la combinaison sur le taux de localisation des détecteurs pris isolément.

Abstract *– In this communication, we present a method to localize faces in color images that combines three detectors: an anthropomorphic detector (using a neural appearance-based model), an ellipse detector (using the generalized Hough transform on gradient orientation image) and a coarse skin color model in CbCr space. Given an input image, we compute a kind of probability map by combining linearly the three detectors output. The face position is then determined as the location of the absolute maximum over this map. Improvement of localization rates of individual detectors is clearly shown and results are very encouraging.*

1. Introduction

La localisation de visages est un domaine de recherche en continuelle expansion en raison de ses nombreuses applications : interaction homme-machine, indexation, biométrie … De nombreuses méthodes ont été développées ces dernières années et sont décrites dans un état de l'art très complet [12]. On peut regrouper ces méthodes selon les deux grandes classes communes à la Reconnaissance des Formes : structurelle et globale. Les approches structurelles [11] cherchent à détecter des éléments caractéristiques du visage (yeux, bouche, nez, contour de la tête) puis à combiner les résultats de ces détections grâce à des modèles géométriques et radiométriques, notamment via des modèles déformables [13], ou par l'analyse de "constellations" [1]. Les approches globales traitent une vignette de l'image totale en la codant sous la forme d'un vecteur (codage rétinien, moments, projection) et estiment les paramètres de classifieurs neuronaux [4] ou statistiques [10]. Dans cette communication, nous proposons une nouvelle approche basée sur la combinaison de trois détecteurs respectivement anthropomorphique, géométrique et colorimétrique ; combinaison qui améliore clairement les performances de chacun des détecteurs considéré séparément.

2. Les détecteurs de base

Une partie de l'information contenue dans l'image d'un visage se trouve dans l'orientation du gradient des contours. Le premier avantage des orientations des contours est leur relative invariance à la couleur de la peau. Le second est lié à la présence de parties concaves (yeux, bouche) et convexes (nez) dans un visage qui créent des contours. Deux des trois détecteurs présentés dans cette section utilisent cette information : le modèle d'apparence (réseau de neurones Diabolo) et le modèle d'ellipse (Transformée de Hough Généralisée).

Notre technique, parfaitement classique pour déterminer les orientations, est la suivante :

- convolution de l'image par un masque circulaire dont le diamètre est de 10% de la largeur de l'image,
- estimation du champ de gradient (I_x, I_y) avec un masque de Roberts,
- seuillage des modules du gradient afin de déterminer les pixels de contours,
- calcul en ces points de l'orientation discrétisée sur N=36 valeurs.

Pour le Diabolo le seuil est défini pour chaque vignette 13x17, à 20% des modules les plus grands (on conserve donc 20% de pixels comme étant de contour). Pour la transformation de Hough, le seuil est appliqué sur l'image entière. Il est de 12 et a été optimisé sur 168 images d'apprentissage.

2.1 Le réseau diabolo : un modèle d'apparence du visage

Chaque pixel de contour de la vignette se voit attribuer deux caractéristiques (I_{cos},I_{sin}) :

$$I_{\cos}(i,j)=\cos\left(\frac{2\pi}{N}\cdot orienx(i,j)\right) \text{ et } I_{\sin}(i,j)=\sin\left(\frac{2\pi}{N}\cdot orienx(i,j)\right)$$

avec $orienx = round\left(\frac{N}{2\pi}\arctan\frac{I_y}{I_x}\right)$ mod(N). Les pixels

hors contours reçoivent les valeurs (0,0). Cette étape produit deux tableaux I_{cos} et I_{sin} que l'on réduit à la taille 13x17 par interpolation bi-cubique. On extrait les valeurs résultantes se trouvant à l'intérieur d'un masque elliptique afin d'éliminer les effets des bords de l'imagette. Un vecteur de 290 éléments est ainsi obtenu (on notera la réduction de dimension comparée à 13x17x2=442).

Le traitement de ces exemples est réalisé par un réseau de neurones auto-associateur (réseau Diabolo [3]). Ce réseau est entraîné à fournir en sortie une image identique à celle mise en entrée en réalisant une compression spécialisée car la couche cachée comporte un nombre de cellules nettement inférieur à celui de l'entrée ou de la sortie. Une image de non-visage sera en principe mal compressée et donnera une erreur de reconstruction plus importante (figure 1). 1 602 images de visage sont utilisées en apprentissage et 178 en validation croisée. Les poids et biais du réseau de neurones sont estimés par descente de gradient de l'erreur quadratique avec adaptation du pas d'apprentissage. Une recherche exhaustive a montré que pour 290 entrées (rétine de taille 13x17), une couche cachée de 18 neurones est optimale. L'image en niveau de gris est parcourue, à l'échelle correspondant à la taille du visage, par une rétine 13x17, et à chaque position de l'image une erreur de reconstruction est calculée. Une carte des erreurs de reconstruction est ainsi obtenue, nous l'appelerons *DiaboloMap* dans la suite.

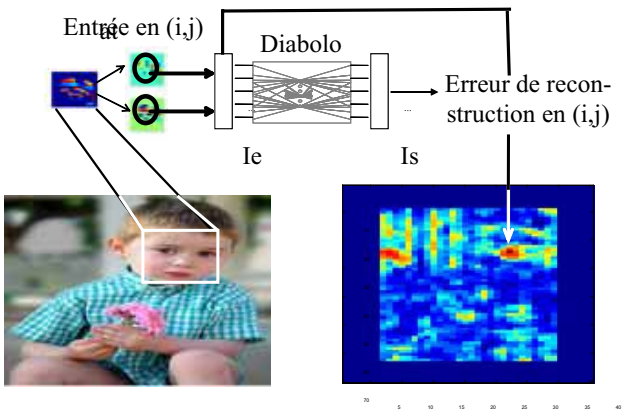


FIG. 1 : Erreur de reconstruction du diabolo (les couleurs chaudes correspondent aux erreurs de reconstruction faibles).

2.2 Détecteur d'ellipse basé sur la Transformation de Hough Généralisée

Une Transformation de Hough Généralisée (THG) est réalisée sur la carte des orientations. Concernant la détection d'ellipses, il existe une structure simplifiée de la THG basée sur les propriétés géométriques des ellipses. La méthode consiste à accumuler les votes d'une demi-droite ayant pour point de départ un pixel de contour et dont la direction et le sens sont donnés par l'orientation du contour en ce point. L'orientation de l'ellipse est supposée connue : les visages sont modélisés par une ellipse d'axe vertical d'excentricité donnée. Alors, pour chaque point de contour *M*, une simple table précalculée spécifie l'angle entre la tangente au contour *Mt* et le rayon *MO* (*O* étant le centre de l'ellipse passant par *M*).

Il en résulte un tableau de vote dont le maximum correspond à la position dans l'image du point le plus susceptible d'être le centre de l'ellipse. Comme le fond est souvent structuré et complexe, ce maximum ne localise qu'approximativement la position du visage. Afin de diminuer l'effet de ce fond sur le tableau de vote, ce dernier est d'abord lissé puis parcouru par un "chapeau mexicain" de taille 13x17. La somme des éléments de chaque sous-tableau, pondérés par les coefficients du masque, donne un nouveau score. La carte de l'image qui en résulte (figure 2) sera nommée *HoughMap* dans la suite de cet exposé.

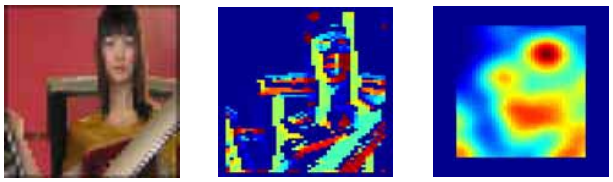


FIG. 2 : image originale, orientation du gradient des contours et tableau de vote.

2.3 Modèle de teinte chair

Indépendamment de l'information d'orientation des contours, une détection de peau basique est implémentée dans l'espace *YCbCr* [2]. L'avantage de l'espace colorimétrique *YCbCr* est qu'il sépare l'information de luminance (*Y*) de celle de la chrominance (*Cb-Cr*). Notre filtre de teinte chair est défini par les intervalles suivants : $Cb \in [105 \quad 130]$ et $Cr \in [135 \quad 160]$. La peau étant caractérisée par des informations de chrominance spécifiques, ce filtre peut s'appliquer aux différents types de couleur de peau rencontrés en fonction des origines ethniques. Ces seuils ont été déterminés expérimentalement en utilisant des images (hors test) contenant des visages. On présente ci-dessous la matrice de confusion de cette opération de seuillage sur un échantillon de 145 736 580 pixels de couleur peau (set 3 de la base ECU [8]) et 667 359 498 pixels non-peau :

TAB. 1 : Matrice de confusion du modèle de couleur peau.

Classification / Etiquette	Peau	Non-peau
Peau	76%	24%
Non-peau	19%	81%

Ce modèle de couleur peau est grossier, et dans certains cas tous les pixels de peau sont masqués, mais la fusion des trois détecteurs nous permet l'utilisation d'un modèle simple. Les seuils utilisés ne sont pas universels du fait que la chrominance dépend dans une certaine mesure de la valeur de la luminance Y [5]. Dans des conditions d'illumination faible ou saturée, les composantes filtrées laissent apparaître des trous. Le masque des pixels "non-peau" bruité ainsi obtenu est alors lissé, et des opérations de morphologie mathématique sont appliquées pour remplir les éventuels trous, et le résultat est binarisé. A chaque position de l'image, le pourcentage de pixel filtré dans une fenêtre 13x17 est calculé. Nous appellerons la carte qui en résulte *SkinMap*.

3. Combinaison des différents détecteurs

Nous avons défini trois détecteurs qui produisent, pour une image couleur donnée, en trois cartes : *DiaboloMap*, *HoughMap*, et *SkinMap*. La combinaison linéaire de ces trois réponses (figure 3) permet d'améliorer les taux de localisation de chaque détecteur pris isolément [9]. La carte résultante est appelée *FusionMap*.

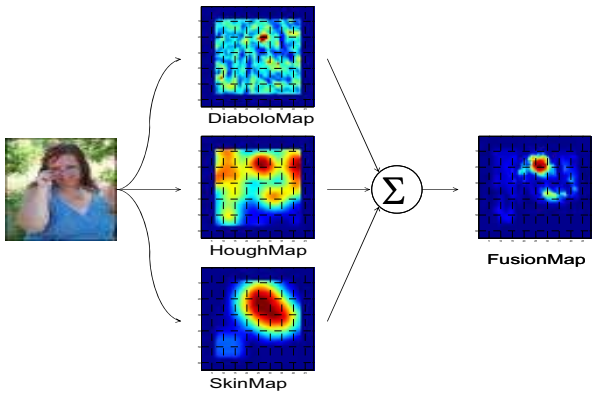


FIG. 3 : Système de localisation du visage

Dans cette optique, la réponse de chaque détecteur est ajustée linéairement dans l'intervalle $[-1\ \ 1]$. Appelant ces cartes normalisées *D*,*H* et *S*, l'imagette 13x17 à la position (i,j) dans l'image originale est alors caractérisée par $I_{i,j} = \left[H_{i,j}\ \ D_{i,j}\ \ S_{i,j} \right]$. Cent images de la base d'apprentissage ont été utilisées pour apprendre les coefficients de la combinaison linéaire par descente du gradient arrêtée par validation croisée :

$$FusionMap_{i,j} = a \cdot H_{i,j} - b \cdot D_{i,j} + c \cdot S_{i,j} + d$$

avec a=0. 2280, b=0.2620 , c=0. 1229 et d= -0.7198

On notera la faiblesse du poids du détecteur de couleur peau comparé à celui autres détecteurs.

4. Evaluation des performances

Pour déterminer la position du visage, l'image est parcourue par les trois détecteurs précédemment décrits, et le visage correspond à la position du maximum de la carte FusionMap résultante. Nos expérimentations sont réalisées sur la base ECU [8] qui se divise en images en couleur de personnes (set 1), la vérité terrain des visages (set 2) ainsi que la vérité terrain de la teinte chair (set 3). La tâche de localisation sur cette base est particulièrement complexe en raison de la grande variabilité apparaissant dans les images de visages (figure 4) au niveau des caractéristiques intrinsèques (origine ethnique, âge, orientation) et extrinsèques (taille dans l'image, résolution).

Un premier test est réalisé sur 1 353 images (distinctes des images des corpus d'apprentissage et de validation croisée) contenant un seul visage, permettant une évaluation du taux de localisation (nombre de visages correctement localisés divisé par le nombre total de visages). Pour chaque image, la taille du visage est supposée connue, ce qui nous permet d'appliquer une fenêtre glissante de taille fixe à l'image. Cette connaissance revient à celle de la distance du visage à la caméra. Les images en test sont donc réduites de manière à ce que le visage puisse être contenue dans une fenêtre 13x17. Cette taille permet de respecter le rapport hauteur sur largeur d'un visage, et est un bon compromis entre visibilité (pour un œil humain) des caractéristiques d'un visage et l'effort computationnel. Après combinaison, 1 166 visages sur les 1 353 que contient la base de test sont correctement localisés, soit un taux de localisation de 86%. La fusion des trois détecteurs permet ainsi de diminuer le taux d'erreur de 50% (tableau 2).

FIG. 4 : Exemples de localisation.

TAB. 2 Performances des détecteurs et combinaison.

détecteur	diabolo	hough	fusion
Taux de localisation	48.5 %	67 %	**86 %**

Un second test est réalisé sur 205 images (différentes de celle utilisée en apprentissage ou en validation croisée) contenant un total de 482 visages. Dans les images "mono-visage" la position du visage est définie par celle du maximum de FusionMap. Dans une image contenant N visages (N étant supposée connu pour un problème de localisation) les N plus grand maximum locaux (suffisamment distants pour éviter les recouvrements) définissent la position des visages. 396 visages sont correctement détectés (figure 5) sur 482 (82%).

FIG. 5 : Localisation multi-visage.

5. Conclusion et perspectives

Cette communication a pour but de présenter une contribution significative au problème de la localisation de visage. Nous avons présenté trois détecteurs représentant des sources d'information différentes : couleur de la peau, apparence et forme géométrique du visage. Nous avons montré qu'une fusion linéaire de ces trois sources améliore clairement les performances de chacun des détecteurs.

De nombreuses améliorations sont à l'étude. Pour la teinte chair d'abord, nous utilisons un modèle probabiliste avec mélange de gaussiennes [6]. Pour la combinaison de détecteurs, un expert de fusion neuronale devrait aussi améliorer les résultats comparé à la fusion linéaire. De plus un algorithme de détection robuste des yeux a été développé [7] et sera utilisé pour la validation des visages candidats.

Références

[1] Bileschi S.M. & Heisele B., Advances in Component Based Face Detection, *IEEE Int Workshop on Analysis and Modeling of Face and Gestures, 2003.*

[2] Chai D. & Nang K.N., Locating facial region of a head-and-shoulders color image, *Int Conf on Automatic Face and Gesture Recognition, pp.124-129, 1998.*

[3] Féraud R., Bernier O., Viallet J. & Collobert M., A Fast and Accurate Face Detector Based on Neural Networks, *IEEE Trans. PAMI, 23(1), pp. 42-53, 2002.*

[4] Garcia C. & Delakis M., Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection *IEEE Trans. PAMI, 26(11), pp. 1408-1422, 2004.*

[5] Hu M., Worrall S., Sadka A.H. & Kondoz A.M., Automatic scalable face model design for 2D model-based video coding, *Signal Processing: Image Communication, 19, pp. 421-436, 2004.*

[6] Milgram M., Prevost L. & Belaroussi R., Une nouvelle transformation pour la localisation des yeux dans une image de visage monochrome, à paraître, *GRETSI 2005.*

[7] McKenna S.J., Gong S. & Raja Y., Modelling facial colour and identity with gaussian mixtures, *Pattern Recognition 31(12), pp.1883-1892, 1998.*

[8] Phung S.L., Bouzerdoum A. & Chai D., Skin segmentation using color pixel classification: Analysis and comparison, *IEEE Trans. PAMI, 27(1), pp 148-154, 2005.*

[9] Prevost L. & Milgram M., Automatic Allograph Selection and Multiple Expert Classification for Totally Unconstrained Handwritten Character Recognition, *IEEE Int Conf on Pattern Recognition, (1), pp 381-383, 1998.*

[10] Sung K.K & Poggio T., Example-based learning for view-based human face detection, *IEEE Trans. PAMI, 20(1), pp 39-51, 1995.*

[11] Yang G. & Huang T. S., Human Face Detection in Complex Background, *Pattern Recognition, 27(1), pp. 53-63, 1994.*

[12] Yang M.H., Kriegman D. & Ahuja N., Detecting Faces in Images: A Survey, *IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 1, pp. 34-58, 2002.*

[13] Yuille A., Hallinan P. & Cohen D., Feature Extraction from Faces Using Deformable Templates, *Int Journal Computer Vision, 8(2), pp. 99-111, 1992.*

Sélection de variables pour la reconnaissance de formes

Sébastien GADAT[1], Laurent YOUNES[2]

[1]Centre de Mathématiques et de Leurs Applications
ENS de Cachan, 61 avenue du président Wilson, 94235 Cachan Cedex, France

[2]Center for Imaging Science
The Johns Hopkins University, 3400 N-Charles Street Baltimore MD 21218-2686, USA
sebastien.gadat@cmla.ens-cachan.fr
laurent.younes@cis.jhu.edu

Résumé – Nous étudions le problème d'extraction d'éléments informatifs d'un signal à partir d'un grand ensemble de variables descriptives $\mathcal{F}$, en préservant des capacités de classification optimales. Notre travail s'appuie sur une nouvelle approche sous-optimale du problème d'extraction de variables. Nous étudions une loi de probabilité sur $\mathcal{F}$ qui permet à la fois d'en extraire des éléments en mesurant leur efficacité, mais aussi d'optimiser un critère de performance. Nous donnons alors des algorithmes stochastiques de minimisation de ce critère et obtenons des améliorations notables de performance sur différents exemples.

Abstract – We introduce a new model formalizing the extraction of meaningfull features from a large set of variables $\mathcal{F}$ describing a signal or an image. Our work is based on a subobtimal study using a probability distribution $\mathbb{P}$ on $\mathcal{F}$. We can extract features with respect to this probability map and optimize the efficiency of the method of extraction. Meaningfull features are then detected by looking at the weights denoted by $\mathbb{P}$ on $\mathcal{F}$. We give stochastic algorithms to learn optimal distribution of weights and we apply our method to several problems like spam detection or face recognition.

1 Introduction

De nombreux problèmes de reconnaissance de formes dans des domaines aussi divers que l'interprétation de scène dans une image, de textes ou de *microarrays* génétiques engendrent la manipulation d'un très grand nombre de variables. Pour des raisons statistiques ou algorithmiques, la multiplication de ces variables descriptives nuit aux performances de classification. En effet, les variables peu informatives d'un signal agissent comme un bruit artificiel tandis que la variance statistique d'un échantillon de données augmente avec la dimension des signaux. Il s'agit donc de réduire la variance des signaux tout en préservant un biais faible sur les décisions prises par les algorithmes de classification [1]. Par ailleurs, certains algorithmes ne sont pas implémentables dans le cadre des grandes dimensions à la vue des temps de calcul qui leur sont alors nécessaires. De plus, le nombre d'échantillons nécessaire pour maintenir un niveau de précision donné croît exponentiellement avec le nombre de variables (piège des grandes dimensions [2]). Enfin, il peut être aussi important qu'optimiser des performances, en génétique par exemple, d'identifier les variables qui permettent l'obtention de ces performances.

Les méthodes telles que l'analyse en composantes principales ou indépendantes ne permettent pas d'identifier les variables importantes puisque celles-ci fournissent des espaces de petite dimension sans pour autant donner les variables les plus importantes pour le problème traité [3]. D'autres méthodes basées purement sur des tests statistiques (critère sur des p-valeurs par exemple) fournissent souvent des coordonnées explicatives sur les signaux sans pour autant assurer l'optimalité du jeu de variables sélectionnées.

Nous formalisons notre modèle de sélection de variables comme la recherche d'une distribution de poids optimale sur ces variables. Notre stratégie est sous-optimale, étant donnée une loi de probabilité sur l'ensemble des variables $\mathcal{F}$, nous munissons notre modèle d'une énergie que nous cherchons à minimiser. Nous donnons alors plusieurs algorithmes destinés à optimiser un tel critère avant d'appliquer notre étude à la détection de spam et à la reconnaissance de visages.

2 Modèle d'extraction de variables

2.1 Cadre statistique

Nous suivons le cadre statistique supervisé classique de la reconnaissance de formes, en utilisant un *Training Set* et un *Test Set*. Le signal I à classer est vu comme la réalisation d'une variable aléatoire et est décrit par des variables explicatives (questions, tests ...) notées $\mathcal{F} = \{\delta_1, \dots \delta_f\}$, on notera alors $\delta_i(I)$ ses coordonnées. L'objectif est alors de prédire parmi $\mathcal{C}_1, \dots \mathcal{C}_N$ la classe à laquelle appartient I. Il peut s'agir d'identifier par exemple si une image contient un visage, ou si un e-mail est en fait du spam. Nous supposons par ailleurs que nous disposons d'un algorithme $\mathbb{A}$ pour classer ces signaux, pouvant s'exécuter à la fois à partir de la totalité des variables $\delta(I)$ et sur un sous-ensemble de variables $\omega(I), \omega \subset \mathcal{F}$. Nous utiliserons ici les algorithmes de Support Vector Machine ou de

plus proches voisins. Ces algorithmes $\mathbb{A}$ permettent alors, pour chaque ω, de définir des indices de performance relatifs au problème de classification étudié. Par exemple, $q(\omega) = P[\mathbb{A}(\omega(I)) \neq \mathcal{C}(I)]$ désigne le taux d'erreur de classification renvoyé par $\mathbb{A}$ que nous utiliserons comme indice de pertinence de notre sous-ensemble ω. Nous estimons alors q par la probabilité empirique mesurée sur l'ensemble d'apprentissage de données :

$$\hat{q}(\omega) = \hat{P}[\mathbb{A}(\omega(I)) \neq \mathcal{C}(I)]$$

2.2 Modèle sous-optimal

Étant donné un entier k, nous cherchons à choisir le meilleur k-uplet de variables ω issues de $\mathcal{F}$ ayant une valeur optimale pour $\hat{q}$. Pour de simples raisons combinatoires, il est impossible de concevoir l'énumération de tous les k-uplets ω afin de déterminer pour chacun $\hat{q}(\omega)$ puisque si $\mathcal{F}$ contient 2000 variables et si $k = 50$, il est alors nécessaire d'exécuter C^{50}_{2000} algorithmes $\mathbb{A}$!

Afin de contourner ce problème combinatoire, nous adoptons une stratégie sous-optimale. Nous munissons l'ensemble des variables $\mathcal{F}$ d'une loi de probabilité $\mathbb{P}$, et pour un entier k quelconque fixé, notre modèle de sélection de variables consiste alors en k tirages de tests de $\mathcal{F}$ indépendants, avec remise, effectués selon $\mathbb{P}$. La loi sur ces k-uplets de $\mathcal{F}$ sera notée $\mathbb{P}^k$. Nous souhaitons alors minimiser l'erreur moyenne commise par $\mathbb{A}$ lorsque les sous-ensembles de k questions suivent la loi $\mathbb{P}^k$, cela nous amène naturellement à considérer l'énergie $\mathcal{E}$:

$$\mathcal{E}(\mathbb{P}) = \mathbb{E}_{\mathbb{P}^k}[q(\omega)] = \sum_{\omega \in \mathcal{F}^k} q(\omega)\mathbb{P}^k(\omega) \qquad (1)$$

2.3 Estimation de q

Afin de calculer $\mathcal{E}$, nous devons être capable d'évaluer q. Nous utilisons la procédure suivante pour approcher $q(\omega)$ par $\hat{q}(\omega)$:

- Tirer un échantillon de données T_1 du *Training Set*.
- Éxécuter l'algorithme $\mathbb{A}_{\omega,T_1}$ sur les coordonnées ω et l'échantillon T_1.
- Tirer un échantillon T_2 sur le *Training Set* et mesurer l'erreur empirique de A_{ω,T_1} sur T_2.

L'objectif recherché, étant donné une telle méthode pour mesurer l'erreur empirique sur l'ensemble d'apprentissage, est de minimiser $\mathcal{E}$. Une distribution de poids optimale aura alors tendance à favoriser les tests δ privilégiés pour le problème de classification en chargeant de façon importante ces variables. En revanche, les « mauvaises » variables ne permettant pas à $\mathbb{A}$ d'obtenir de bonnes performances seront faiblement pondérées par $\mathbb{P}$. Nous quantifierons donc l'utilité d'une variable δ de $\mathcal{F}$ par la valeur $\mathbb{P}_\infty(\delta)$: plus $\mathbb{P}_\infty(\delta)$ est grande, plus δ est utile à la tâche de classification où $\mathbb{P}_\infty$ sera alors une distribution optimale.

3 Algorithmes de recherche

Nous cherchons donc à minimiser $\mathcal{E}$. Tout comme il est impossible de calculer toutes les valeurs de $q(\omega)$, il est impossible de se ramener à l'optimisation d'un polynôme de variables $\mathbb{P}$. Nous avons donc choisi d'approcher une distribution $\mathbb{P}$ optimale en utilisant une descente de gradient. L'algorithme d'apprentissage proposé suivra dès lors le schéma récursif résumé par la figure 1.

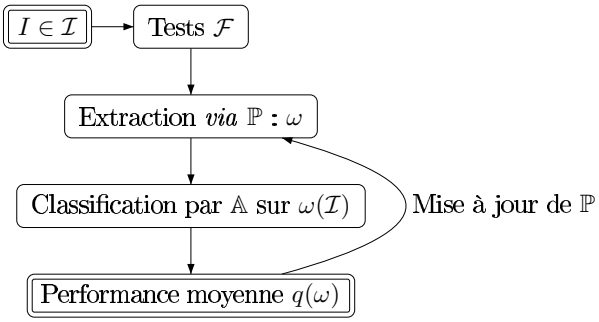


FIG. 1 – Schéma général de l'apprentissage de $\mathbb{P}$

3.1 Descente de gradient euclidienne et riemannienne

Métrique euclidienne L'optimisation de $\mathcal{E}$ se fait sur le simplexe de $\mathbb{R}^f$ noté $\mathcal{S}_\mathcal{F}$ et défini par

$$\sum_{\delta \in \mathcal{F}} \mathbb{P}(\delta) = 1 \qquad \text{et} \qquad \mathbb{P}(\delta) \geq 0$$

En métrique euclidienne, le gradient de $\mathcal{E}$ s'exprime en :

$$\forall \delta \in \mathcal{F} \qquad \nabla\mathcal{E}(\mathbb{P})(\delta) = \sum_{\omega \in \mathcal{F}^k} \frac{C(\omega,\delta)q(\omega)\mathbb{P}(\omega)}{\mathbb{P}(\delta)} \qquad (2)$$

où $C(\omega,\delta)$ désigne simplement le nombre d'occurrences du test δ dans le k-uplet ω. Le schéma de descente correspondant est alors

$$\mathbb{P}_{n+1} = \mathbb{P}_n - \alpha_n \pi(\nabla\mathcal{E}(\mathbb{P})) \qquad (3)$$

π désigne simplement la projection vectorielle d'un vecteur sur l'hyperplan de $\mathbb{R}^f$ portant $\mathcal{S}_\mathcal{F}$.

Métrique riemannienne La principale difficulté d'une approche basée sur (3) est que la contrainte de positivité des poids $\mathbb{P}_n$ est difficile à satisfaire. Il est donc naturel de choisir une paramétrisation de $\mathcal{S}_\mathcal{F}$ en variables exponentielles. En posant $y = \log(\mathbb{P})$, on obtient une nouvelle énergie $\tilde{\mathcal{E}}(y) = \mathcal{E}(\mathbb{P})$ et la descente de gradient sur y correspondante s'écrit :

$$y_{n+1} = y_n - \alpha_n \overline{\pi}\left(\nabla\tilde{\mathcal{E}}(y)\right)$$

Cette descente se traduit alors sur les poids par une équation de la forme

$$\mathbb{P}_{n+1} = \mathbb{P}_n \frac{e^{\alpha_n \nabla\tilde{\mathcal{E}}(y_n)}}{K_n} \qquad (4)$$

où K_n est une constante de renormalisation positive assurant la normalisation de la somme des poids à 1. Cette paramétrisation permet donc de satisfaire la contrainte de positivité sur $\mathbb{P}_n$ naturellement. Il est possible d'interpréter une telle propriété en remarquant qu'en munissant $\mathcal{S}_\mathcal{F}$ de la métrique riemannienne de Kullback sur les lois de probabilité (*i.e.* la métrique du χ_2), on a alors la proposition suivante.

Proposition 3.1 (Gradient riemannien) *Pour toute distribution de poids strictement positive de $\mathcal{S}_\mathcal{F}$, on a*

$$\nabla_{\chi_2}\mathcal{E}(\mathbb{P}) = \nabla\tilde{\mathcal{E}}(y)$$

Il est ainsi aisé d'interpréter la descente en variables exponentielles comme étant une descente sur les variables $\mathbb{P}$ avec le produit scalaire dérivé de la distance du χ_2. Cette distance a pour propriété de mettre à distance « infinie » un point sur la frontière du simplexe d'un point strictement à l'intérieur.

3.2 Diffusion réfléchie

La descente de gradient riemannienne précédente permet de définir structurellement la positivité des poids, mais son évolution n'est pas aussi rapide que l'évolution euclidienne (paramétrisation exponentielle de petites quantités entre 0 et 1). Afin de garder la vitesse d'évolution de la méthode euclidienne et la stabilité structurelle de $\mathcal{S}_\mathcal{F}$, il est donc judicieux d'utiliser un processus de diffusion réfléchie dans $\mathcal{S}_\mathcal{F}$ avec terme de dérive dirigé par le gradient euclidien de $\mathcal{E}$.

$$\mathrm{d}\mathbb{P}_t = -\pi\left(\nabla\mathcal{E}(\mathbb{P}_t)\right) + \sqrt{\Sigma}\mathrm{d}W_t + \mathrm{d}Z_t \qquad (5)$$

Σ est alors une matrice non-dégénérée dans $\mathcal{S}_\mathcal{F}$, W_t un mouvement brownien standard alors que $\mathrm{d}Z_t$ est un processus de rappel sur la face la plus proche du simplexe lorsque le terme $(\nabla\mathcal{E}(\mathbb{P}_t)) + \sqrt{\Sigma}\mathrm{d}W_t$ amène le processus hors de $\mathcal{S}_\mathcal{F}$.

L'existence et l'unicité de (5) sont assurées dans notre cas et dépendent en réalité de la construction de l'application de Skorokhod [4]. C'est ce processus que nous simulerons lors de nos expériences pour faire diminuer $\mathcal{E}$.

3.3 Approximation stochastique

En étudiant l'expression (2) du gradient de $\mathcal{E}$, on constate que pour implémenter numériquement une équation de la forme (5), il faut évaluer le terme $\nabla\mathcal{E}(\mathbb{P}_n)$ à chaque itération, terme qui nécessite une trop grande énumération de $\hat{q}(\omega)$ pour être réaliste en temps de calculs. Afin de pallier cette difficulté numérique insurmontable de manière exacte, il est possible de mettre en place un algorithme d'approximation stochastique dès qu'on remarque que

$$\nabla\mathcal{E}(\mathbb{P})(\delta) = \mathbb{E}\left[\frac{q(\omega)C(\omega,\delta)}{\mathbb{P}(\delta)} \,|\, \omega\sim\mathbb{P}^k\right] \qquad (6)$$

L'apprentissage stochastique suivant permet alors de simuler asymptotiquement (3) ou (5) [5, 6, 7].

- Initialiser $\mathbb{P}_0$ à la loi uniforme sur l'ensemble des variables $\mathcal{F}$.
- Pour calculer $\mathbb{P}_{n+1}$, extraire ω_n selon $\mathbb{P}_n^k$ et mettre à jour les poids *via* :

$$\mathbb{P}_{n+1} = \mathbb{P}_n - \alpha_n d_n + \sqrt{\alpha_n}\mathrm{d}\xi_n + \mathrm{d}z_n \qquad (7)$$

avec la définition suivante du vecteur d_n :

$$d_n(\delta) = \frac{q(\omega)C(\omega_n,\delta)}{\mathbb{P}_n(\delta)} \qquad (8)$$

3.4 Dynamique d'apprentissage

L'hypothèse fondamentale d'espérance (6) étant également satisfaite pour le terme d_n, on peut alors démontrer ([8]) le théorème fondamental suivant :

Théorème 3.2 (Asymptotique de (7)) *Le processus discret donné par (7) est une trajectoire pseudo-asymptotique de (5). De plus, le processus continu (5) est récurrent positif et les processus continu et discret convergent vers l'unique mesure stationnaire μ de (5) qui est le champ de Gibbs associé à l'énergie $\mathcal{E}$, donné par*

$$\mu(\mathbb{P}) = \frac{e^{-\mathcal{E}(\mathbb{P})}}{Z}$$

Ce théorème assure la convergence des simulations numériques vers des états d'énergie faible, et contourne la difficulté de la simulation exacte de (5).

4 Expériences

Nous implémentons notre algorithme de sélection de variables par apprentissage de poids optimaux sur deux exemples classiques de classification.

4.1 Reconnaissance de visages

Dans le cas de la détection de visages, nous utilisons la base de donnéees fournies par le MIT, des imagettes de taille 19×19, centrées codées sur 8 bits en niveaux de gris.

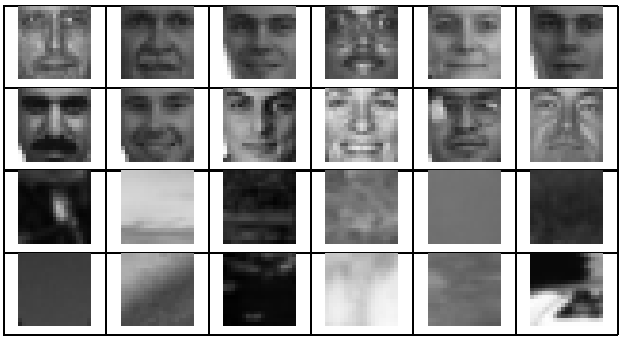


FIG. 2 – Échantillon issu de la base de données

Les variables utilisées sont des détecteurs de bords orientés, binaires décrits dans [9], très rapides à calculer. L'algorithme $\mathbb{A}$ est un algorithme de séparation par hyperplan de *Support Vector Machine* linéaire. Le nombre de variables initiales préselectionnées est de l'ordre de 2000 alors que nous cherchons à ne retenir que quelques dizaines de tests pour obtenir des performances de classification correctes.

La figure 3 suivante représente l'évolution, en fonction du nombre k, du gain apporté par une sélection des variables par une loi optimale plutôt que par la loi uniforme $\mathcal{U}_\mathcal{F}$.

On note une sensible amélioration des performances après apprentissage d'une loi d'extraction, surtout lorsque le nombre de variables à retenir est faible. La performance finale obtenue est 1.6% d'erreur sur le *Test Set* avec (5). Enfin, les détecteurs de bords privilégiés (*i.e.* avec un poids

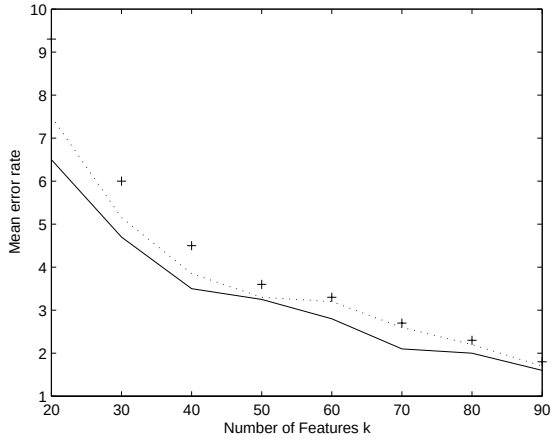


FIG. 3 – Erreur moyenne sur les données Tests avec des extractions par loi uniforme (+) et loi limite $\mathbb{P}_\infty$ apprises par descente riemannienne (pointillés) et diffusion réfléchie (ligne continue) en fonction de $k = |\omega|$.

$\mathbb{P}_\infty$ important) sont localisés sur les zones intuitivement informatives de l'image : bouche, nez, contour du visage comme le représente la figure 4.

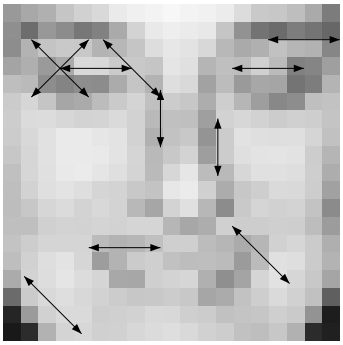


FIG. 4 – Représentation des détecteurs de bords privilégiés après apprentissage.

4.2 Détection de spam

Nous évaluons également nos algorithmes sur des données fournies par l'UCI de messagerie électronique. Il s'agit de messages étiquetés (spam/non spam) séparés en ensembles d'apprentissages et de tests, issus d'une même boîte mail.

Les variables utilisées mesurent la fréquence d'apparition de 54 mots pré-définis dans un e-mail pour former $\mathcal{F}$. $\mathbb{A}$ est cette fois un algorithme de 4-plus proches voisins. La figure 5 représente alors l'évolution de l'erreur moyenne de l'algorithme $\mathbb{A}$ avec le temps.

Là encore, l'algorithme utilisant la diffusion réfléchie semble plus performant que la simple descente de gradient en variables exponentielles. Les mots privilégiés lors de nos expériences sont par ordre décroissant *remove, business, receive, internet, ...* pour le spam et *cs, 857, 415, project, ...* pour les « vrais e-mails ». Si les mots associés au spam nous sont familiers, les mots relatifs aux « vrais e-mails » sont en réalité relatifs au caractère personnel de l'auteur de cette base de données (métier, coordonnées ...). Enfin, avec un vote de 10 détecteurs utilisant 15 mots extraits selon $\mathbb{P}_\infty$, nous obtenons 7.5% d'erreur, la meilleure performance sur cette base étant de 7% d'erreur.

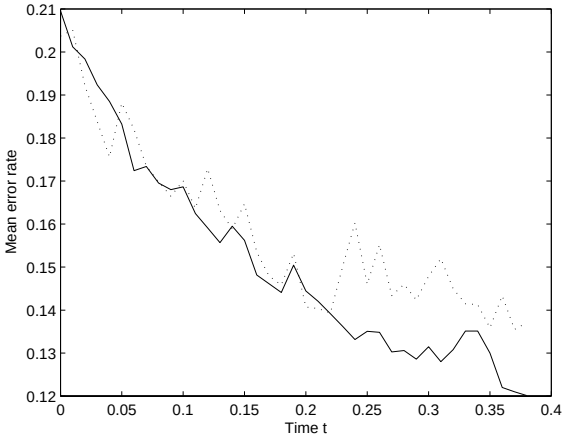


FIG. 5 – Erreur en fonction du temps obtenue par descente riemannienne et diffusion réfléchie (pointillés).

Des travaux futurs généraliseront cette approche en permettant à $\mathcal{F}$ d'évoluer en utilisant une dynamique inspirée des algorithmes génétiques (suppression et composition de variables).

Références

[1] S. Geman, E. Bienenstock and R. Doursat. *Neural networks and the bias/variance dilemma*, Neural Computation, vol. 4, pp. 1-58, 1992.

[2] R. Bellman. *Adaptative Control Processes : A guided Tour*, Princeton University Press, 1961.

[3] C. Jutten and J. Hérault. *Blind separation of sources, part 1 : An adaptative algorithm bases on neuromimetic architecture*, Signal Processing, vol. 24, pp. 1-10, 1991.

[4] P. Dupuis and H.Ishii. *On Lipschitz continuity of the solution mapping to the Skorokhod problem, with applications*, Stochastics Stochastics Rep., vol. 35, no. 1,pp. 31-62,1991.

[5] S. Gadat and L. Younes. *A stochastic algorithm of features extraction for pattern recognition*, Preprint CMLA, pp. 1-22, 2004.

[6] H. Kushner and G. Yin. *Stochastic approximation and recursive algorithms and applications*, Springer-Verlag, 2003.

[7] M. Benaim. *A dynamical system approach to stochastic approximations*, SIAM Control Optim., vol. 34, no. 2, pp. 437-472, 1996.

[8] S. Gadat. *Jump diffusion over Feature space for object recognition*, Preprint CMLA, pp. 1-27, 2005.

[9] Y. Amit and D. Geman. *A computational model for visual selection*, Neural Computation, vol. 11, pp. 1691-1715, 1999.

Classification de signaux invariante en translation

Vincent Guigue Alain Rakotomamonjy Stéphane Canu

Lab. Perception, Systèmes, Information - CNRS - FRE 2645
Avenue de l'Université, 76801 St Étienne du Rouvray
Tél : 02 32 95 97 03 Fax : 02 32 95 97 08
Vincent.Guigue@insa-rouen.fr

Résumé – Cet article présente une méthode de classification de signaux non-stationnaires dans le cas où les signaux contiennent des motifs caractéristiques dont la position temporelle est variable est inconnue. Nous avons montré que la construction d'une représentation graphique des signaux, basée sur une transformée en ondelettes continues permet de s'affranchir d'une référence temporelle absolue et de faire face efficacement à ce problème. Après avoir défini un produit scalaire entre graphes, nous avons comparé les résultats en classification entre les SVM et les kppv.

Abstract – This paper addresses the question of the classification of non-stationary signals. We make the hypothesis that each signal includes a pattern, the shape of which is discriminant but the time location of which is random and unknown. We propose to build a graphical representation, based on continuous wavelet transform in order to eliminate the absolute time reference of the description. Then, we define a dot product between graphs and compare the results of SVM and k-nn.

1 Introduction

La classification de signaux non-stationnaires est un problème complexe. Nous avons cherché à le résoudre dans le cas où chaque signal contient un motif discriminant dont la position est aléatoire et inconnue. Les descripteurs statistiques classiques comme la moyenne ou les coefficients de Fourier ne donnent pas une représentation satisfaisante pour la classification de signaux non stationnaires. Les deux approches usuelles consistent à traiter directement les signaux ou à utiliser les représentations temps-échelles et temps-fréquences [Dav00], [BJ93]. Dans les deux cas, la description est de grande dimension. Une solution pour traiter l'hypothèse d'un motif discriminant de position inconnue est d'optimiser la forme des ondelettes afin de rendre l'apparition du motif plus nette [LDHD02]. Il est également possible de filtrer les signaux afin de faire ressortir les parties discriminantes.

Cependant, nous souhaitons introduire le minimum de connaissance *a priori* et de paramètres dans le problème. Le but de cet article est de proposer une description originale des données pour caractériser des motifs indépendamment de leurs positions à l'intérieur du signal. Nous montrerons qu'il est possible de créer une telle représentation en décrivant les signaux sous la forme de graphes. Ces graphes sont issus d'une représentation en ondelettes continues, afin de conserver une description temporelle tout en s'affranchissant d'une référence temporelle absolue. Le calcul de produits scalaires entre les graphes [KTI03] nous permet ensuite de classer les signaux avec des *Support Vector Machines* (SVM) [Vap98]. Nous avons comparé les résultats SVM avec un k plus proches voisins basé sur la distance induite par ce produit scalaire.

2 Méthode

2.1 Représentation des données

Soit $\{S_1, \ldots, S_n\}$ un ensemble de signaux d'apprentissage de longueur N étiqueté en $\{-1, 1\}$. Nous cherchons une description discriminante, invariante en translation, pour la classification de ces signaux. Le caractère non-stationnaires des signaux nous a conduit à opter pour les représentations temps échelles. Nous noterons $S(t)$ et $r(S(t))$ un signal et sa représentation :

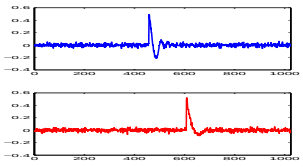
$$\begin{array}{rccc} r: & \mathcal{X} & \rightarrow & \mathcal{F} \subset (\mathbb{R}\times\mathbb{R}\times\mathbb{R})^k \\ & S & \rightarrow & r(S) = \{t_i, E_i, P_i\}_{i=1}^k \end{array} \quad (1)$$

$$\text{avec : } P_{E_i,t_i} = \langle S, \psi_{E_i,t_i}\rangle = \int_{-\infty}^{+\infty} S(t)\psi^\star\left(\frac{t-t_i}{E_i}\right)dt$$

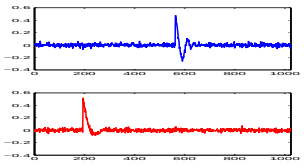
où k est le nombre de coefficients composant le plan, chaque coefficient étant défini comme un triplet temps-échelle-poids.

Pour faire face à l'hypothèse d'un motif dont la position est variable dans le signal, nous avons besoin d'une représentation invariante en translation, c'est à dire : $r(S(t-\tau)) = r_\tau(S(t))$, où r_τ est une translation de la représentation du signal. Nous nous sommes donc tournés vers les représentations en ondelettes continues. Cette famille analysante $\{\psi_{E_i,t_i}\}_{i=1,\ldots,k}$ ne forme pas une base, elle est redondante mais elle est invariante en translation [Mal97].

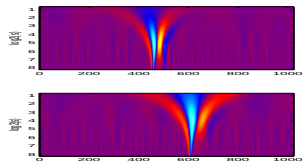
Les motifs que nous cherchons à identifier ont une signature dans le plan temps-échelles. Le but est de la faire apparaître en mettant en avant les caractéristiques discriminantes des motifs. Le temps est à la fois un facteur important pour décrire le motif et un facteur pénalisant puisque la position du motif est variable. La solution



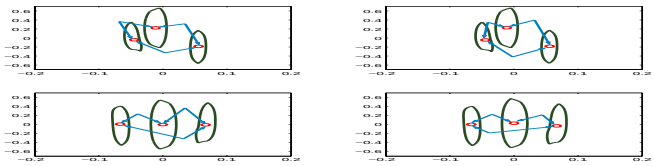
(a) Signaux peu bruités ($\sigma_b = 0.02$)



(b) Signaux fortement bruités ($\sigma_b = 0.2$)



(c) Transformées en ondelettes continues ($\sigma_b = 0.02$)



(d) Représentations graphiques ($\sigma_b = 0.02$)

Figure 1: Les 2 représentations graphiques du haut sont de la classe 1 ($u = 1 \cdot 10^{-3}, v = 2 \cdot 10^{-3}$), les 2 du bas sont de la classe -1.($u = 5 \cdot 10^{-4}, v = 6 \cdot 10^{-4}$)

consiste à relativiser les informations temporelles. Cette deuxième représentation r_2 est basée sur une matrice A des arcs entre les points du plan temps-échelles :

$$\begin{array}{rccc} r_2 : & \mathcal{F} & \rightarrow & \mathcal{G} \subset (\mathbb{R}\times\mathbb{R})^k \times \mathbb{R}^{k\times k} \\ & \{t_i, E_i, P_i\}_{i=1}^k & \rightarrow & r_2(S) = \{E_i, P_i\}_{i=1}^k, A \end{array} \quad (2)$$

où $\mathcal{G}$ est l'espace des graphes. Dans cet espace, les signaux prennent la forme de graphes : chaque coefficient de la transformée en ondelettes continues est un nœud lié à tous les autres coefficients du graphe. Les nœuds portent les informations d'échelle et de poids, les arcs portent les informations de temps Δ_t.

2.2 Réduction de la dimensionnalité

La représentation r_2 est particulièrement volumineuse, elle compte k nœuds et $k(k-1)$ arcs. De plus, la complexité du calcul du produit scalaire entre graphes dépend essentiellement du nombre de nœuds des graphes (section 2.4). Il est donc nécessaire de réduire la taille de la représentation pour pouvoir calculer efficacement des produits scalaires dans ce nouvel espace. Cette réduction de la dimensionnalité devra conserver une part significative de l'énergie du signal pour que la représentation ne perde pas son sens.

Nous faisons l'hypothèse que l'information discriminante se trouve dans les zones de fortes énergies du signal. La dimension du problème sera donc réduite tout en conservant une partie significative de la représentation du signal initial. Donoho [Don95] a montré que l'élimination des coefficients les plus faibles dans une transformée en ondelettes orthogonales permet de réduire le bruit dans un signal. Mallat [MZ92] a complété cette étude dans le cas des ondelettes continues, en montrant que les valeurs maximales de la transformée en ondelettes permettent de reconstruire le signal.

La solution consiste à diminuer le nombre de nœuds des graphes tout en augmentant le nombre d'informations présentes dans chacun des nœuds. Nous proposons de prendre un modèle gaussien pour les zones de fortes énergies du plan temps-échelles (cf Fig. 1(d)). Chacun des nœuds est alors une gaussienne, définie par : la matrice de covariance dans le plan, la somme des coefficients de la zone, le nombre de coefficients de la zone et la localisation de la gaussienne en échelle. Les arcs représentent les Δ_t entre les positions des gaussiennes dans le temps.

2.3 Produit scalaire entre graphes

Le produit scalaire entre ces représentations graphiques que nous utilisons est dérivé de [KTI03]. L'idée est de comparer des chemins de toutes longueurs, partant de chaque nœud dans les deux graphes, pondérés par la probabilité de ces chemins :

Soient :

- h_i^1 et h_j^2 les nœuds i et j des graphes G^1 et G^2,
- $p(h_i|h_{i-1})$ la probabilité de transition du nœud h_{i-1} au nœud h_i,
- $p_q(h_\ell)$ la probabilité de s'arreter au nœud ℓ,
- $K_n(h_k^1, h_k^2)$ le produit scalaire entre les nœuds h_i^1 et h_j^2 de G^1 et G^2,
- et $K_a(a^1_{h^1_{k-1}h^1_k}, a^2_{h^2_{k-1}h^2_k})$ le produit scalaire entre 2 arcs a^1 et a^2 de G^1 et G^2.

$$K(G^1, G^2) = \sum_{\ell=1}^{\infty} \sum_{h^1} \sum_{h^2} p(h_1^1) \prod_{i=2}^{\ell} p(h_i^1|h_{i-1}^1) p_q(h_\ell^1) \times p(h_1^2) \prod_{j=2}^{\ell} p(h_j^2|h_{j-1}^2) p_q(h_\ell^2) \times K_n(h_1^1, h_1^2) \prod_{k=2}^{\ell} K_a(a^1_{h^1_{k-1}h^1_k}, a^2_{h^2_{k-1}h^2_k}) K_n(h_k^1, h_k^2) \quad (3)$$

Les détails sont donnés dans [KTI03], les modifications concernent la gestion des nœuds qui contiennent plus d'informations que dans le papier original.

Le calcul de se produit scalaire conduit à la résolution d'un système linéaire de taille $(|G^1||G^2|)^2$, où $|G^i|$ est le nombre de nœuds du graphe i. Elle nous a forcés à réduire au maximum le nombre de nœuds dans chacun des graphes. Pour conserver une représentation significative des signaux, nous avons choisi d'augmenter le nombre d'informations présentes dans chacun des nœuds à travers une modélisation gaussienne.

2.4 Alternatives aux graphes

Afin de valider notre approche, nous avons implémenté quatre descriptions alternatives au noyau de graphes.

- **Les signaux bruts** : la méthode la plus simple pour décrire les données.
- **Les descripteurs statistiques classiques**, généralement utilisés pour faire face à des signaux stationnaires [Hea00, Has00]. Nous avons utilisé 16 descripteurs statistiques (moyenne, écart type, amplitude...) complétés par une analyse en bande de puissances sur 6 bandes de fréquence.
- **La combinaison des translations** possibles entre deux signaux, qui est très couteuse mais bien adaptée au problème d'invariance en translation. L'idée est de trouver la distance tangente [HHB04, SCDV98] entre les deux familles de signaux issues des translations des deux signaux originaux :
$$\langle S_1, S_2 \rangle = \max_{\tau \in \Omega} \left(k(r(S_1), r_\tau(S_2)) \right) \tag{4}$$
où Ω est l'ensemble des translations considérées.
- **Les sacs de vecteurs** : les gaussiennes issues de la modélisation du plan temps-échelles (section 2.2) sont comparées deux à deux, seules les informations temporelles des arcs du graphes sont laissées de coté [WCG03].
$$K_{match} = \frac{1}{2}\left(K + K^T\right)$$
$$\text{avec: } K(S_1, S_2) = \frac{1}{|G^1|} \sum_{i=1}^{|G^1|} \max_j \left(k(V_i^1, V_j^2) \right) \tag{5}$$
où V_i^k est le vecteur décrivant la i^e gaussienne du signal S_k.

Les noyaux (4) et (5) ne sont pas définis positifs, du fait de l'utilisation de la fonction max. Nous faisons l'approximation suivante :

$$\max_{y_j \in \mathcal{Y}} \left(k(x_i, y_j) \right) \approx \frac{1}{|\mathcal{Y}|} \sum_{j=1}^{|\mathcal{Y}|} \exp\left(-\frac{||x_i - y_j||^2}{2\sigma^2} \right)$$

Le noyau (5) est alors équivalent au *Multiple Instance Kernel* [GFKS02].

3 Expérience

3.1 Données

Dans beaucoup d'applications, les données sont composées de motifs et de bruits. La forme du motif est discriminante mais sa position est aléatoire et inconnue. Nous travaillons sur des données artificielles présentant de telles caractéristiques. Soit $m(t)$ un *chirp* à décroissance exponentielle :

$$m(t) = e^{-\alpha t} \cos((u + vt)t + \phi) \tag{6}$$

Pour compliquer ce problème, un bruit blanc gaussien $b(t)$ de variance σ_b^2 est ajouté sur les données. Soit $\Gamma(t)$ la fonction échelon, le signal $S(t)$ s'écrit :

$$S(t) = m_{u,v}(t - \tau)\Gamma(t - \tau) + b(t) \tag{7}$$

Le couple $\{u, v\}$ caractérise le motif et donc la classe du signal, l'instant de déclenchement τ n'est pas discriminant. Dans cet article, nous nous sommes limités à l'utilisation d'ondelettes simples de type *Sombrero* pour la transformée en ondelettes continues.

3.2 Résultats

Les résultats (Tab. 1) sont présentés suivant deux méthodes de classification : les *Support Vector Machine*, sur les différents produits scalaires, et les k plus proches voisins, sur les distances induites. Les résultats utilisent 1 ou 400 échantillons en apprentissage.

Les noyaux de base utilisés (pour les nœuds, les arcs, les signaux bruts etc...) sont des noyaux gaussiens de largeur de bande σ. Les paramètres σ, C (compromis biais-variance des SVM) et k (nombre de plus proches voisins) sont optimisés par validation croisée sur l'ensemble d'apprentissage.

Le tableau 1 présente les résultats obtenus en utilisant plusieurs produits scalaires avec différentes méthodes de classification et pour différents niveaux de bruit dans les signaux. La figure 2 montre l'évolution du taux d'erreur de classification en fonction du nombre de points disponibles dans la base d'apprentissage.

Le problème faiblement bruité devient trivial dès lors qu'une représentation adaptée à l'hypothèse du motif de position inconnue est utilisée. Le problème du bruitage des données et celui de la diminution de la taille de la base d'apprentissage mettent en avant les capacités discriminantes de la représentation graphique des données. La comparaison entre représentation graphique et sacs de vecteurs permet de quantifier l'apport lié à la description de la structure des données. En effet, les seules différences entre les deux descriptions sont les arcs qui portent les informations Δ_t.

4 Conclusion

La description non vectorielle des données offre de nouvelles perspectives dans de nombreux domaines. Dans le cas de la classification de signaux non stationnaires, cette description graphique des données permet de s'affranchir de la référence absolue au temps tout en réduisant la dimension du problème. Le but de l'étude est donc atteint puisque les motifs sont décrits de la même manière quelque soit leur position dans le signal et que cette représentation permet d'améliorer sensiblement les taux de reconnaissance par rapport aux autres descriptions.

Bruit	$\sigma_b = 0.2$			$\sigma_b = 0.02$		
|App.|/|Test.|	1/1000	400/1000		1/1000	400/1000	
Classifieurs	1-ppv and SVM	1-ppv	SVM	1-ppv and SVM	1-ppv	SVM
Sig. bruts	49.99% (±0.84)	48.10% (±2.01)	47.08% (±0.97)	48.19%(±3.52)	41.45% (±3.12)	38.54% (±1.96)
Descr. stat.	47.16% (±8.22)	35.54% (±2.12)	19.27% (±0.98)	6.12% (±4.30)	5.62% (±2.12)	5.79% (±2.22)
Translations	49.73% (±1.15)	43.17% (±1.99)	42.27% (±1.63)	28.75% (±2.28)	15.12% (±1.99)	14.91% (±1.93)
Sacs de vect.	29.40% (±15.24)	11.9% (±1.13)	8.29% (±0.80)	**0% (±0)**	**0% (±0)**	**0% (±0)**
Graphes	**13.98% (±13.26)**	6.66% (±0.64)	**5.25% (±0.35)**	**0% (±0)**	**0% (±0)**	**0% (±0)**

Table 1: Taux de mauvaise classification, moyennes et écart-types sur 30 itérations avec des données faiblement bruitées ($\sigma_b = 0.02$) et fortement bruitées ($\sigma_b = 0.2$). Les résultats sont donnés avec 1 (cas extrême) et 400 échantillons en apprentissage.

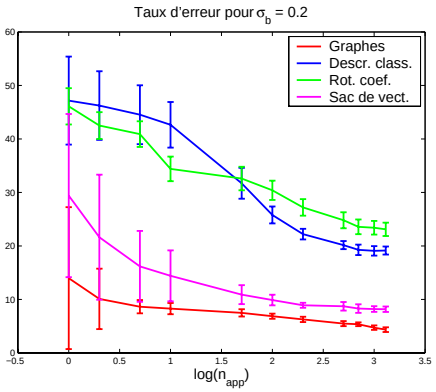


Figure 2: Évolution des taux d'erreur de classification en fonction de la taille de l'ensemble d'apprentissage.

Les perspectives de ce travail concernent l'amélioration de la modélisation du plan temps-fréquences, afin de prendre en compte les cas où l'information discriminante se trouve dans des zones de faibles énergies.

Remerciements

Ce travail est financé en partie par le programme IST de la communauté européenne, avec le réseau d'excellence PASCAL, IST-2002-506778. Cette publication reflète uniquement le point de vue des auteurs.

References

[BJ93] R.G. Baraniuk and D.L. Jones. A signal dependant time-frequency representation : Optimal kernel design. In *IEEE Transactions on Signal Processing*, volume 41, pages 1589–1601, 1993.

[Dav00] M. Davy. *Noyaux optimisés pour la classification dans le plan temps-fréquence - Proposition d'un algorithme constructif et d'une référence bayesienne basée sur les méthodes MCMC - Application au diagnostic d'enceintes acoustiques*. PhD thesis, Universite de Nantes, 2000.

[Don95] D.L. Donoho. De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41:613–627, 1995.

[GFKS02] T. Gärtner, P.A. Flach, A. Kowalczyk, and A.J. Smola. Multi-instance kernels. In *ICML*, pages 179–186, 2002.

[Has00] E. Haselsteiner. *Neural Based Methods for Time Series Classification*. PhD thesis, Technischen Universität Graz, 2000.

[Hea00] J. Healey. *Wearable and Automotive Systems for the Recognition of Affect from Physiology*. PhD thesis, MIT, 2000.

[HHB04] B. Haasdonk, A. Halawani, and H. Burkhardt. Adjustable invariant features by partial haar-integration. In *ICPR 2004*, volume 2, pages 769–774, 2004.

[KTI03] H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. In *20th International Conference on Machine Learning*, pages 321–328. AAAI Press, 2003.

[LDHD02] M.F. Lucas, C. Doncarli, E. Hitti, and N. Dechamps. Wavelet optimization for classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, Orlando, Florida, USA, May 2002.

[Mal97] S. Mallat. *A Wavelet Tour Of Signal Processing*. Academic Press, 1997.

[MZ92] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Anal. and Mach. Intell.*, 14(7):710–732, 1992.

[SCDV98] P. Simard, Y. Le Cun, J. Denker, and B. Victorri. Transformation invariance in pattern recognition, tangent distance and tangent propagation. *Lecture Notes in Computer Science*, 1524:239–274, 1998.

[Vap98] V. N. Vapnik. *The Statisitcal Learning Theory*. Springer, 1998.

[WCG03] C. Wallraven, B. Caputo, and A.B.A. Graf. Recognition with local features : the kernel recipe. In *ICCV 2003 Proceedings*, volume 2, pages 257–264. IEEE Press, 2003.

Treillis à complexité réduite pour le décodage de codes à longueur variable

Gholam-Reza MOHAMMAD-KHANI, Chang-Ming LEE, Michel KIEFFER, Pierre DUHAMEL

LSS — CNRS – Supélec – Université Paris-Sud
Plateau de Moulon - F-91192 Gif-sur-Yvette Cedex
Tél. : 00 33 1 69 85 17 32 – Fax : 00 33 1 69 85 17 65
{Gholam-Reza.Mohammad-Khani,lee,kieffer,Pierre.Duhamel}@lss.supelec.fr

Résumé – De nombreux algorithmes ont été proposés pour le décodage souple de données codées à l'aide de codes à longueur variable (CLV), la plupart travaillant avec des treillis. Pour un code réaliste, ces treillis sont très complexes à cause du nombre de mots de code à considérer. Cet article présente un algorithme de regroupement de mots de code à longueur variable en un nombre minimal de classes, ce qui permet de réduire significativement la complexité du décodage souple. Un exemple sur les CLV de la norme H263+ ainsi qu'une d'application à la localisation des frontières de blocs de texture H263+ sont présentés.

Abstract – Many trellis-based soft decoding techniques have been proposed for data encoded using variable-length codes (VLC). However, for actual VLC tables, these trellises are too complex to allow real-time soft decoding. This paper presents an algorithm for grouping VLC codewords into classes, which allows significant reductions of the size of the resulting trellises and of the complexity of soft decoding. Illustrations are provided on the VCL table used for texture encoding of H263+. The performances of decoding techniques for the localization of block frontiers in a bitstream generated by an H263+ coder are also presented.

1 Introduction

Dans de nombreux standards de compression, tel que JPEG, H263+ ou le mode de base de H264, une étape de codage à longueur variable (CLV) des données est effectuée. Les trains binaires générés par un CLV sont particulièrement sensibles à l'égard d'erreurs de transmission, principalement à cause de la perte de synchronisation qui peut résulter de certaines erreurs introduites par le canal.

Une solution classique consiste à insérer dans le train binaire des marques de synchronisation ou à faire appel à un codage canal puissant. Cependant, ceci résulte en une augmentation significative du débit binaire sur le canal. Des techniques plus récentes ont exploité la structure particulière des codes CLV pour développer des algorithmes de décodage souple de trains binaires issus d'un CLV [1–5]. L'ensemble des successions de mots de code possibles est décrite à l'aide de treillis exploitant certaines informations supposés disponibles *a priori* au niveau du décodeur (nombre de mots de code, longueur du train binaire à décoder, *etc.*). Les performances obtenues sont bien meilleures que celles atteintes par des décodeurs classiques. Des performances encore meilleures peuvent être obtenues en tenant compte de la syntaxe du train binaire généré par le codeur source considéré, voir par exemple [6, 7].

Toutes ces méthodes ont en commun la construction et l'élagage de treillis intégrant plus ou moins d'information *a priori*. La complexité de ces treillis peut devenir très importante lorsque des CLV comportant un nombre important de mots de code sont considérés. Cet article présente un algorithme de construction de tables simplifiées pour le décodage de données issues d'un CLV. Les mots de code sont regroupés en un nombre minimal de classes. Au lieu de travailler sur l'ensemble des mots de code, les algorithmes de décodage souple peuvent utiliser ces classes en nombre réduit, ce qui diminue la complexité des treillis et les temps de calcul.

Le paragraphe 2, présente brièvement un type de treillis utilisé pour le décodage de données issues d'un CLV. Le paragraphe 3 décrit le principe de l'algorithme de simplification des mots de code utilisés par un CLV donné. Un exemple de simplification des tables utilisées pour le codage de la texture par H263+ est présenté au paragraphe 5. Enfin, les performances d'un algorithme de décodage à entrées souples utilisant une table normale sont comparées à celle d'un algorithme utilisant une table compactée dans le paragraphe 4.

2 Décodage souple de CLV

On considère une source X générant des symboles appartenant à un alphabet $\mathcal{X} = \{X_1, \ldots, X_K\}$. A chaque symbole X_k est associé une probabilité d'occurence p_k. Un CLV associe à chaque symbole un mot de code $\mathbf{x}_k$ de $\ell_k = \ell(\mathbf{x}_k)$ bits, tel que $\ell_{\min} \leqslant \ell_k \leqslant \ell_{\max}$ pour tout $k \in \{1, \ldots, K\}$. A une suite de N symboles de la source peut donc être associé un vecteur de N mots de code résultant du CLV

$$\mathbf{x}_1^N = (\mathbf{x}(1), \ldots, \mathbf{x}(N)), \text{ avec } \sum_{i=1}^{N} \ell(\mathbf{x}(i)) = L.$$

Cette suite peut aussi être vu comme un vecteur de L bits $\mathbf{b}_1^L = (b_1, \ldots, b_L)$. Lorsque $\mathbf{b}_1^L$ est placé à l'entrée d'un canal de transmission, la sortie du canal est notée $\mathbf{y}_1^L = (y_1, \ldots, y_L)$.

Au niveau du décodeur, lorsque seuls L et la table des mots de code utilisée par le CLV sont connus, l'estimée $\hat{\mathbf{x}}$ au sens du maximum de vraisemblance de $\mathbf{x}_1^N$ est

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x} \in \mathcal{S}_L} p\left(\mathbf{y}_1^L \mid \mathbf{x}\right), \tag{1}$$

où $\mathcal{S}_L$ est l'ensemble de toutes les suites de mots de codes de L bits pouvant être générées par le CLV considéré. Lorsque le canal est gaussien sans mémoire, (1) se traduit par la minimisation sur $\mathcal{S}_L$ d'une norme euclidienne.

Pour résoudre ce problème de minimisation sous contrainte, [8] propose la construction d'un treillis représentant l'ensemble des suites de mots de code de longueur totale L pouvant être généré par un CLV. Une fois le treillis disponible, des algorithmes tels que SOVA [9] ou BCJR [10] peuvent être employés. Dans le treillis proposé, chaque nœud (a) represente la fin d'une suite de mots de code dont la longueur cumulée est a bits. Chaque branche correspond à un mot de code, ainsi, un groupe de branches parallèles reliant les nœuds (a) et $(a+\ell)$ representent tous les mots de code de ℓ-bits. La figure 1 montre qu'après un certain temps, la structure devient périodique. Chaque période du treillis contient au plus $\ell_{\max}$ nœuds. Le nombre total de nœuds est égal à L, le nombre de bits du vecteur $\mathbf{y}_1^L$, cependant, le nombre de branches arrivant et émergeant d'un nœud donné peut être très grand. Si l'on considère par exemple le CLV réalisé pour la texture dans la norme H263+, pour les mots de code de 10 bits, 42 branches parallèles relient $(a-10)$ et (a). Pour le décodage, le nombre de métriques de branche à évaluer est donc très important.

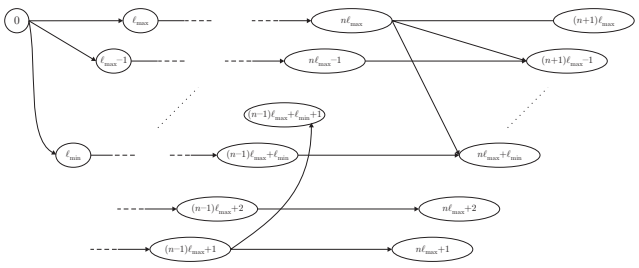


FIG. 1: *Structure de trellis pour des CLV de longueur comprise entre $\ell_{\min}$ et $\ell_{\max}$*

3 Simplification de tables de CLV

Une méthode permettant de réduire la complexité de décodage consiste à utiliser une table de mots de code de taille réduite. Cet article présente un algorithme de regroupement des mots de code de même longueur en classes. Ainsi, par exemple, les 16 mots de code de 8 bits utilisés pour le codage de la texture dans la norme H263+

$$\mathcal{A}_8 = \{00100000, 00100001 \ldots 00101110, 00101111\}$$

sont formés du préfixe 0010 et de tous les suffixes possible de 4 bits. $\mathcal{A}_8$ peut ainsi être décrit de manière compacte par $0010\$\$\$\$$, où $\$$ représente soit 0, soit 1. A l'aide de cette simplification, dans le treillis présenté au paragraphe 2, 16 branches parallèles entre les nœuds (a) et $(a+8)$ peuvent être remplacées par une branche unique. Le principe de l'algorithme proposé généralise la technique de réduction présenté précédemment.

3.1 Notations et définitions

On considère l'ensemble des N mots de code de L bits, $\mathcal{A}_L = \{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$, utilisé par un CLV donné. Le cardinal de $\mathcal{A}_L$ est noté $|\mathcal{A}_L|$. Il s'agit de trouver un regroupement des éléments de $\mathcal{A}_L$ de manière à obtenir un nombre minimal de représentants pour ces mots de code.

Soit $\mathbb{B} = \{0,1\}$, l'ensemble des mots binaires, qui peut être étendu en $\overline{\mathbb{B}} = \{0,1,\$\}$ avec l'élément *indéterminé*, correspondant à $\$ = \{0,1\}$. Une *classe* $\mathbf{c}$ est un élément de $\overline{\mathbb{B}}^L$. Soit $\mathcal{S}(\mathbf{c})$, l'ensemble des éléments de $\overline{\mathbb{B}}^L$ correspondant à $\mathbf{c}$. Ainsi, si $\mathbf{c} = (c_1, \ldots, c_L) \in \overline{\mathbb{B}}^L$, alors

$$\begin{aligned}\mathcal{S}(\mathbf{c}) \quad = \quad & \{(x_1, \ldots, x_L) \in \mathbb{B}^L \text{ tq } x_i = c_i \text{ si } c_i \neq \$ \\ & \text{et } x_i \in \{0,1\} \text{ si } c_i = \$\}.\end{aligned}$$

Une *classification* $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_M\}$ de $\mathcal{A}_L$ est un ensemble de classes tel que

$$\begin{cases} \forall \mathbf{x} \in \mathcal{A}_L, \exists i \text{ tel que } \mathbf{x} \in \mathcal{S}(\mathbf{c}_i), \\ \forall \mathbf{x} \in \mathcal{S}(\mathbf{c}_i),\ i = 1, \ldots, M, \mathbf{x} \in \mathcal{A}_L. \end{cases}$$

La méthode de simplification proposée consiste ainsi à représenter $\mathcal{A}_L$ à l'aide d'une *classification minimale* $\mathcal{C}$, c'est-à-dire une classification comportant un nombre minimal de classes.

Avant de présenter l'algorithme de construction de la classification minimale, quelques notions complémentaires doivent être introduites. La distance de Hamming $d(\cdot,\cdot)$ se généralise de $\mathbb{B}$ à $\overline{\mathbb{B}}$. Ainsi, si $(x,y) \in \overline{\mathbb{B}}$, $d(x,y) = 1$ si $x \neq y$ et $d(x,y) = 0$ sinon. La distance de Hamming entre deux classes est la somme des distances entre les composantes de ces classes. Deux classes sont *adjacentes* lorsque leur distance de Hamming est de un. L'*ordre* $o(\mathbf{c}) = \log_2(|\mathcal{S}(\mathbf{c})|)$ d'une classe $\mathcal{C}$ correspond au nombre d'éléments indéterminés de cette classe. Deux classes adjacentes $\mathbf{c}_1$ et $\mathbf{c}_2$ peuvent être *réunies* pour former une nouvelle classe $\mathbf{c} = \mathbf{c}_1 \sqcup \mathbf{c}_2$, on a alors $\mathcal{S}(\mathbf{c}) = \mathcal{S}(\mathbf{c}_1) \cup \mathcal{S}(\mathbf{c}_2)$ et $o(\mathbf{c}) = o(\mathbf{c}_1) + 1$.

3.2 Construction d'une classification minimale

Pour des raisons de place, seul le principe de l'algorithme de construction d'une classification minimale est présenté ici. Le détail de l'algorithme ainsi qu'une preuve d'optimalité est présenté dans [11].

L'algorithme a pour entrée un ensemble $\mathcal{A}_L$ de mots de codes de longueur L. Chaque élement de $\mathcal{A}_L$ est également une classe d'ordre 0. Une première classification $\mathcal{C}_0$ de $\mathcal{A}_L$ est formée à l'aide des éléments de $\mathcal{A}_L$. Un ensemble de classes $\mathcal{C}_1$ est formée de toutes les classes adjacentes de $\mathcal{C}_0$ qu'il est possible de réunir. $\mathcal{C}_1$ n'est plus nécessairement une classification de $\mathcal{A}_L$ (par exemple, s'il existe un élément $\mathbf{c} \in \mathcal{C}_0$ n'ayant pas de classe adjactente, alors $\mathbf{c} \notin \mathcal{C}_1$. La procédure est itérée jusqu'à l'obtension d'une classe $\mathcal{C}_{k_{\max}}$ ne comportant plus aucune classe adjacente.

L'union de tous les ensembles ainsi obtenus $\overline{\mathcal{C}} = \bigcup_{k=0}^{k_{\max}} \mathcal{C}_k$ est une classification *maximale* de $\mathcal{A}_L$, c'est-à-dire qu'elle contient toutes les classes qu'il est possible de former par réunion de classes adjacentes obtenues à partir des éléments de $\mathcal{A}_L$. Une classification minimale $\underline{\mathcal{C}}$ de $\mathcal{A}_L$ est alors un sous-ensemble de $\overline{\mathcal{C}}$ qui reste une classification de $\mathcal{A}_L$. L'obtension de $\underline{\mathcal{C}}$ se fait en deux étapes. La première consiste à éliminer de $\overline{\mathcal{C}}$ tous les éléments des ensembles intermédiares $\mathcal{C}_k$ qui ont pu être regroupés en classes appartenant à $\mathcal{C}_{k+1}$. La classification $\mathcal{C}$ est alors obtenue. Pour former la classification minimale $\underline{\mathcal{C}}$, une méthode combinatoire consistant à éliminer un maximum de

classes de $\mathcal{C}$ tout en vérifiant que le sous-ensemble obtenu reste une classification de $\mathcal{A}_L$ est ensuite considéré.

Cette technique peut fournir une classification minimale pour laquelle certains mots de code peuvent appartenir à plusieurs classes simultanément. Cette situation peut poser des problèmes lors du décodage à l'aide d'algorithmes de type MAP, la somme des probabilités *a priori* des classes n'étant plus égale à 1. Une solution à ce problème nécessite une adaptation de l'algorithme précédent, présentée dans [11].

Exemple 1 *Considérons l'ensemble $\mathcal{A}_7$ formé de 24 mots de code de 7 bits.*

$$\begin{aligned}\mathcal{A}_7 \;=\; \{&\mathbf{000}0000, \mathbf{000}0001, ..., \mathbf{000}1010, \mathbf{000}1011,\\ &\mathbf{0010}100, \mathbf{0010}101, \ldots, \mathbf{0010}111,\\ &\mathbf{0101}000, \mathbf{0101}001, \ldots, \mathbf{0101}011,\\ &\mathbf{1010}100, \mathbf{1010}101, \ldots, \mathbf{1010}111\}\,.\end{aligned}$$

Le calcul de la classification maximale et l'application de la procédure d'élimination des classes inutiles permet d'obtenir la classification suivante de $\mathcal{A}_7$

$$\mathcal{C}'_7 \;=\; \{0000\$\$\$, 000\$0\$\$, 00\$01\$\$, 0\$010\$\$, \$0101\$\$\}\,.$$

Une méthode combinatoire permet ensuite de fabriquer une classification minimale $\underline{\mathcal{C}}_7 = \{0000\$\$\$, 0\$010\$\$, \$0101\$\$\}$ à partir de $\mathcal{C}'_7$.

4 Propriétés

Les tables de CLV obtenue par la procédure de simplification décrite au paragraphe 3 contiennent légèrement moins d'information que les tables initiales. Ceci est du au regroupement des mots de code en classes. Cependant, l'impact de cette perte d'information peut être très limité, selon l'algorithme de décodage utilisé.

Si un décodage au sens du maximum de vraisemblance est utilisé, l'algorithme utilisant les tables simplifiées fournira les mêmes résultats que s'il avait utilisé les tables complètes, en effet, les probabilités *a priori* des symboles ne sont pas utilisées.

Lorsque l'algorithme de décodage cherche à fournir une estimée au sens du maximum de vraisemblance et fournit des sorties souples (en utilisant par exemple SOVA [9], [12]), on peut montrer qu'à nouveau, les résultats sont identiques à ceux obtenus avec une table complète.

Pour un décodeur optimal utilisant un critère de maximum *a posteriori*, tel que celui présenté dans [13], le résultat restera optimal si les probabilités *a priori* de chacun des mots de code regroupés au sein d'une même classe sont les mêmes. Comme les mots de code qui peuvent être regroupés doivent avoir la même longueur, il est raisonnable de penser que cette condition sera vérifiée. Cependant, dans le cas d'une utilisation des tables simplifiées dans le cadre d'une procédure itérative, les différences entre les probabilités *a priori* des mots de code formant une classe pourront être source de sous-optimalité.

5 Application

Dans [7], un algorithme de décodage des informations de texture générées par un codeur H263+ est considéré. Chaque bloc de texture de 8×8 pixels subit un CLV puis est placé dans des paquets qui sont ensuite envoyés sur le canal. Un paquet contient en général plusieurs blocs de texture. Le décodeur proposé par [7] comprend deux étapes. La première consiste à localiser les blocs de texture dans un paquet. La seconde consiste à décoder chaque bloc ainsi localisé en exploitant certaines contraintes liées à la syntaxe du train binaire que peut générer un codeur H263+, voir [6] pour plus de détails.

TAB. 2: *Temps de décodage moyen (en ms/bloc)*

	ML, table CLV		MAP, table CLV	
	initiale	compacte	initiale	compacte
INTRA	165	62 (−62%)	242	92 (−62%)
INTER	356	108 (−70%)	445	131 (−71%)

Pour la première étape, seul le nombre de blocs que comporte un paquet est connu *a priori*. Pour localiser la position des blocs, le treillis présenté au paragraphe 2 doit être modifié pour faire apparaître comme propriété de chaque nœud le nombre de blocs accumulés. Ceci revient à ajouter une dimension au treillis. Pour chaque mot de code, il convient également de distinguer ceux qui marquent la fin d'un bloc.

Ainsi, les 204 mots de code dédiés au codage de la texture dans le standard H263+ sont partitionnés en deux sous-ensembles $\mathcal{A}$ et $\mathcal{E}$ contenant respectivement les mots de code ne marquant pas et marquant la fin d'un bloc. La table 1 rassemble les résultats obtenus par la procédure de simplification des tables de CLV appliquée aux mots de code de même longueur de $\mathcal{A}$ et de $\mathcal{E}$.

Les 204 mots de code peuvent être regroupés en 34 classes. Un algorithme de Viterbi est ensuite employé pour la localisation des frontières de blocs dans des paquets de texture. Les performances sont données en terme de taux d'erreur bloc pour des images codées INTER et INTRA sur la figure 2 et en terme de temps de décodage moyen par bloc dans le tableau 2.

Avec des temps de calcul presque trois fois moindres, les performances sont la plupart du temps très semblables, excepté pour le décodage au sens du maximum *a posteriori* des blocs de type INTRA, où une perte de l'ordre de 0.2 dB peut être notée. La diminution de performances dans le cas des paquets de type INTRA est due au fait que les mots de code regroupés dans une même classe n'ont pas toujours des probabilités *a priori* identiques. Le regroupement se fait alors avec une perte d'information. Dans le cas de paquets de type INTER, les probabilités *a priori* des mots de code regroupés sont plus proches, ce qui explique d'une part des performances du décodeur MAP à peine meilleures que celles du ML et d'autre part, une dégradation moins importante lors de l'utilisation de la table CLV simplifiée.

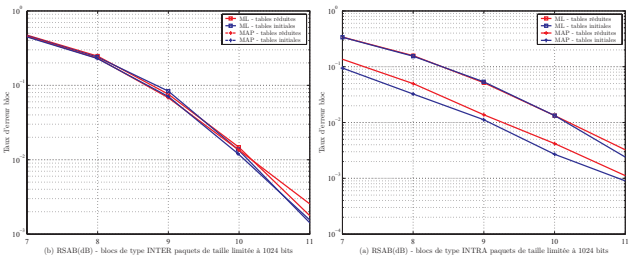


FIG. 2: *Décodage de paquets de blocs codés INTER et INTRA*

TAB. 1: *Classification minimale des mots de code utilisés pour le codage de la texture dans la norme H263+. Les mots marquant la fin d'un bloc (EOB) sont distingués des autres*

Longueur	EOB	Nb de mots de code	Classification minimale	Nb de classes
3	0	2	10$	1
4	0	2	110$	1
5	0	4	111$$	1
	1	2	0111$	1
⋮	⋮	⋮	⋮	⋮
10	0	24	00010010$$, 0001000$$$, 0000111$$$, 00001101$$	4
	1	18	00001100$$, 0000101$$$, 00001001$$, 000010001$	4
11	0	20	000010000$$, 0000001$$$$	2
	1	8	00000001$$$	1
12	0	12	0000000011$$, 000001000$$$	2
	1	12	0000000010$$, 000001001$$$	2
13	0	16	000001010$$$$	1
	1	16	000001011$$$$	1
Total		204		34

6 Conclusion

Cet article présente un algorithme de simplification des tables utilisés pour le décodage de CLV à l'aide d'algorithme de type maximum de vraisemblance ou maximum *a posteriori*. Certains mots de code peuvent être regroupés au sein de classes, ce qui permet de réduire la complexité des algorithmes de décodage. Une preuve d'optimalité de la procédure présentée sera fournie dans la version longue de l'article.

Dans l'exemple traité, la simplification a permi de représenter 204 mots de code par 34 classes, ce qui réduit largement la complexité des treillis représentatnt l'ensemble des succession de mots de code possibles pour une longueur de séquence codée reçue donnée. Les performances du décodeur au sens du maximum de vraisemblance restent inchangées lorsque des tables simplifiées lui sont fournies, mais le temps de calcul est presque trois fois moindre. Pour un décodeur utilisant un critère de maximum *a posteriori*, une légère perte d'optimalité peut être notée lorsque les probabilités *a priori* des mots de code regroupés au sein d'une même classe sont différents. Comme les mots de code qui peuvent être regroupés doivent avoir la même longueur, les différences restent faibles.

Références

[1] V. Buttigieg and P.G. Farrell. A MAP decoding algorithm for variable-length error-correcting codes. In *Codes and Cyphers: Cryptography and Coding IV*, pages 103–119, Essex, England, 1995. The Inst. of Mathematics and its Appl.

[2] V. B. Balakirsky. Joint source-channel coding with variable length codes. In *Proceedings of lEEE ISIT*, Ulm, Germany, 1997.

[3] N. Demir and K. Sayood. Joint source/channel coding for variable length codes. In *IEEE Data Compression Conf.*, pages 139–148, Snowbird, UT, 1998.

[4] K. Sayood, H. H. Otu, and N. Demir. Joint source/channel coding for variable length codes. *IEEE Trans. Commun.*, 48:787–794, 2000.

[5] R. Thobaben and J. Kliewer. On iterative source-channel decoding for variable-length encoded markov sources using a bit-level trellis. In *Proc. IV IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications (SPAWC'03)*, Rome, 2003.

[6] H. Nguyen and P. Duhamel. Compressed image and video redundancy for joint source-channel decoding. In *Proc. Globecom*, 2003.

[7] C.M. Lee, M. Kieffer, and P. Duhamel. Soft decoding of vlc encoded data for robust transmission of packetized video. In *Proceedings of ICASSP*, pages 737–740, 2005.

[8] S. Kaiser and M. Bystrom. Soft decoding of variable-length codes. In *Proceedings of the IEEE International Conference on Communications*, volume 3, pages 1203–1207, New Orleans, 2000.

[9] J. Hagenauer and P. Hoeher. A viterbi algorithm with soft-decision outputs and its applications. In *Proc. Globecom*, pages 1680–1686, Dallas, TX, 1989.

[10] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv. Optimal decoding of linear codes for minimizing symbol error rate. *IEEE Trans. Info. Theory*, 20:284–287, 1974.

[11] G. R. Mohammad-Khani, M. Kieffer, and P. Duhamel. Simplification of vlc tables with application to ml and map decoding algorithms. *IEEE Transactions on Communications*, 2005. submitted.

[12] B. Vucetic and J. Yuan. *Turbo Codes - Principles and Applications*. Kluwer, Dordrecht, 2000.

[13] R. Bauer and J. Hagenauer. Turbo-FEC/VLC-decoding and its application to text compression. In *Proceedings of the Conference on Information Sciences and Systems (CISS'00)*, Princeton University, USA, 2000.

Une comparaison de récepteurs source-canal conjoint utilisant des codes LDPC

Charly POULLIAT[1,3]*, Catherine LAMY-BERGOT[2], Inbar FIJALKOW[1]

[1]ETIS - UMR8051, ENSEA/UCP/CNRS
6 avenue du Ponceau, F-95014 Cergy-Pontoise, FRANCE

[2]THALES Land & Joint Systems, EDS/SPM WaveForm Design Group
160 boulevard de Valmy, B.P. 82, F-92704 Colombes Cedex, FRANCE

[3]Coding Group, Dept. of Elec. Eng., University of Hawaii
2540 Dole Str., Honolulu, HI 96813, USA

{poulliat,fijalkow}@ensea.fr, catherine.lamy@fr.thalesgroup.com

Résumé – L'optimisation d'un récepteur utilisant un décodage source-canal conjoint (DSCC) impliquant un code LDPC sera considérée pour les systèmes itératifs suivant : (a) récepteur source-canal conjoint optimal, (b) récepteurs compatibles avec des applications non conjointes, (c) le récepteur tandem optimal supposant une compression de source parfaite. Des résultats d'optimisation et de simulation seront fournis pour différents rendements et longueurs de mots de code.

Abstract – In this paper, we compare different Joint Channel Decoding (JSCD) schemes involving LDPC codes: (i) JSCD optimised, (ii) non-JSCD compatible optimised, and finally (iii) an optimal tandem system assuming perfect source compression. Some Optimisation and simulation results are provided for different rates and codeword lengths.

1 Introduction

Les récepteurs source-canal conjoints sont des récepteurs de type "turbo" entre un décodeur de source et un décodeur de canal à entrées et sorties souples, dont le but est d'exploiter au mieux la redondance résiduelle et la structure de la source afin d'améliorer les performances en réception. Ainsi, par exemple, [1] considèrent un codeur source concaténé avec un code convolutif et [2] considèrent un système doublement itératif mettant en jeu des turbo-codes. [3] utilisent un code LDPC (Low Density Parity Check) comme code correcteur concaténé : ce choix fut motivé par la facilité d'optimisation de la structure du code permettant de donner un récepteur asymptotiquement optimal pour une source et un code source associé donnés. Ainsi les auteurs proposent-ils une méthode d'analyse et d'optimisation pour obtenir le récepteur optimal sous approximation gaussienne. Cependant, les codes obtenus par cette méthode, si ils permettent d'atteindre de bonnes performances pour le système conjoint, n'offrent pas forcément des performances satisfaisantes si ils sont considérés hors système conjoint. Il est d'un intérêt pratique d'avoir des systèmes qui offrent de bonnes performances à la fois dans un cadre source-canal conjoint et dans un cadre tandem. La question est alors de savoir quels peuvent être ces récepteurs et offre-t-ils un bon compromis de performances sur les deux types d'applications ? Ainsi, l'optimisation du récepteur source-canal conjoint sera considérée pour les systèmes itératifs suivant : (a) récepteur source-canal conjoint optimal, (b) récepteurs compatibles avec des applications non conjointes, (c) le récepteur tandem optimal supposant une compression de source parfaite.

La suite du papier est organisée comme suit : la section 2 présente le système itératif et les hypothèses considérés. En section 3, l'analyse asymptotique de convergence du récepteur conjoint est brièvement décrite. L'optimisation du récepteur conjoint dans les différents cas d'application précités est discutée en section 4. Enfin, des résultats d'optimisation et de simulations sont fournis section 5 et les conclusions et perspectives sont données section 6.

2 Description du récepteur et hypothèses

Dans la suite du papier, le canal de propagation considéré est le canal sans mémoire à Bruit Additif Blanc Gaussien (BABG). Pour des raisons pratiques d'implémentation, nous supposons le code LDPC systématique. Une itération globale de décodage est composée d'une itération de décodeur LDPC suivie d'une itération du décodeur source. Comme dans de nombreux travaux, nous supposerons ici la synchronisation des données effective au niveau paquet (mot de code). Le graphe factoriel correspondant au système proposé est représenté Figure 1.

L'algorithme de propagation de croyances (BP) est utilisé pour le décodage itératif du code LDPC et un Maximum A Posteriori (MAP) pour le décodeur de source. L'optimisation du récepteur conjoint étant basée sur une étude asymptotique de la convergence, on considérera dans un premier temps des mots de code infiniment longs. Cependant, des résultats de simulations seront fournis dans le cas de mots de code de tailles longues et moyennes.

*C. Poulliat est actuellement financé par une bourse postdoctorale du programme Lavoisier du Ministère des Affaires Etrangères.

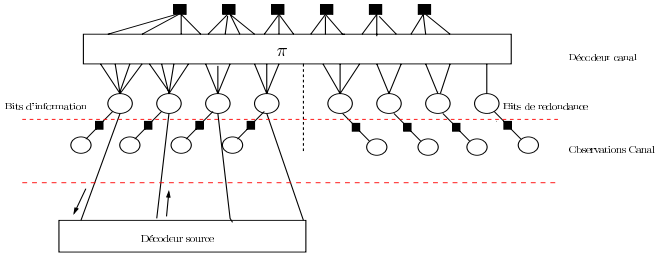


FIG. 1 – Représentation graphique du récepteur conjoint.

3 Analyse asymptotique de convergence

Le paramètre monodimentionnel considéré pour l'analyse asymptotique est l'information mutuelle associée à chaque message log-rapport de vraisemblance (Log-Likelihood ratio, LLR) transitant dans le graphe associée au récepteur (voir [4] [5]). Le décodeur source ne fournissant une information extrinsèque que sur les bits d'information, on distinguera, parmi les noeuds de donnée du codes LDPC, les noeuds de donnée associés aux bits d'information et ceux associés aux bits de redondance. A l'itération de décodage ℓ, on note $x_{cv}^{(\ell)}$, $x_{vc}^{I(\ell)}(i)$, $x_{vc}^{R(\ell)}(i)$, $x_{vs}^{(\ell)}(i)$ et $x_{sv}^{(\ell)}(i)$ respectivement l'information mutuelle en sortie des noeuds de parité, l'information mutuelle associée aux messages entre les noeuds de donnée associés aux bits d'information (I) de degré i et les noeuds de parité, l'information mutuelle associée aux messages entre les noeuds de donnée associés aux bits de redondance (R) de degré i et les noeuds de parité, l'information mutuelle associée aux messages entre les noeuds de donnée de degré i et le décodeur source et enfin l'information mutuelle en sortie de décodeur source. On définit $x_{vc}^{(\ell)}$ comme l'information mutuelle à l'entrée des noeuds de parité (après l'entrelaceur π). On a donc une mixture de $x_{vc}^{I(\ell)}(i)$ et $x_{vc}^{R(\ell)}(i)$. Dans le cadre d'une approximation gaussienne [6], la relation entre l'information mutuelle et la moyenne des messages LLR associés est donnée par la fonction $J(.)$ [4] :

$$J(m) = 1 - \frac{1}{\sqrt{4\pi m}} \int_{\mathbf{R}} \log_2\left(1+e^{-v}\right) \exp\left(\frac{-(v-m)^2}{4m}\right) dv.$$

Pour le code LDPC, $\underline{\rho}=[\rho_2,\ldots,\rho_{t_{r_{\max}}}]^\top$, $\underline{\lambda^I}=[\lambda_2^I,\ldots,\lambda_{t_{c_{\max}}}^I]^\top$ et $\underline{\lambda^R}=[\lambda_2^R,\ldots,\lambda_{t_{c_{\max}}}^R]^\top$ représentent respectivement la proportion de branches du graphe connectées avec des noeuds de parité de degré de connection $\{j, j=2,\ldots,t_{r_{max}}\}$, la proportion de branches du graphe connectées avec des noeuds de données d'information de degré de connection $\{i, i=2,\ldots,t_{c_{max}}\}$ et la proportion de branches du graphe connectées avec des noeuds de données de redondance de degré de connection $\{r, r=2,\ldots,t_{c_{max}}\}$. On définit également $\underline{\lambda}=[\underline{\lambda^I}^\top,\underline{\lambda^R}^\top]^\top$, $\underline{1/t_c}=[1/2,\ldots,1/t_{c_{\max}}]^\top$ et $\underline{1/t_r}=[1/2,\ldots,1/t_{r_{\max}}]^\top$. $t_{r_{max}}$ (resp. $t_{c_{max}}$) est le degré de connection maximum pour un noeud de parité (resp. pour un noeud de donnée). Le canal de propagation étant le canal BABG, la moyenne des messages LLR associés aux observations du canal est donnée par $\mu_0 = 2/\sigma^2$ avec σ^2 la variance du bruit du canal.

Dans le cadre d'une approximation gaussienne de la densité des messages pour le décodeur LDPC et le décodeur source, les équations décrivant l'évolution de l'information mutuelle au sein du récepteur conjoint sont les suivantes [3] :

- Mise à jour des messages pour les noeuds de donnée :

$$x_{vc}^{I(\ell)}(i) = \mathrm{J}(\mu_0 + (i-1)\mathrm{J}^{-1}(x_{cv}^{(\ell-1)}) + \mathrm{J}^{-1}(x_{sv}^{(\ell-1)}(i)))$$
$$x_{vc}^{R(\ell)}(r) = \mathrm{J}(\mu_0 + (r-1)\mathrm{J}^{-1}(x_{cv}^{(\ell-1)}))$$
$$x_{vc}^{(\ell)} = \sum_{i=2}^{t_{c_{\max}}} \lambda_i^I x_{vc}^{I(\ell)}(i) + \sum_{r=2}^{t_{c_{\max}}} \lambda_r^R x_{vc}^{R(\ell)}(r) \quad (1)$$

- Mise à jour des messages pour les noeuds de parité :

$$x_{cv}^{(\ell)} = 1 - \sum_{j=2}^{t_{r_{\max}}} \rho_j \mathrm{J}((j-1)\mathrm{J}^{-1}(1-x_{vc}^{(\ell)})) \quad (2)$$

- Mise à jour des messages du décodeur LDPC vers le décodeur source :

$$x_{vs}^{(\ell)}(i) = J(\mu_0 + iJ^{-1}(x_{cv}^{(\ell)})), \forall i = 2,\ldots,t_{c\max} \quad (3)$$

- Mise à jour des messages en sortie de décodeur source :

$$x_{sv}^{(\ell)}(i) = \mathrm{T}(x_{vs}^{(\ell)}(i)), \forall i = 2,\ldots,t_{c\max} \quad (4)$$

où T(.) la fonction EXIT chart associée au décodeur source, généralement obtenue par méthode de Monte Carlo comme dans [7] en utilisant une approximation gaussienne. La combinaison des équations (1), (2), (3) and (4) donne l'évolution complète de l'information mutuelle

$$x_{vc}^{(\ell+1)} = F([\underline{\lambda^I},\underline{\lambda^R}], x_{vc}^{(\ell)}, \mu_0) \quad (5)$$

et pour laquelle les conditions à l'initialisation sont $\forall i = 2\ldots t_{c_{\max}}$, $x_{sv}^{(0)}(i) = 0$ and $x_{cv}^{(0)} = 0$. La condition $F([\underline{\lambda^I},\underline{\lambda^R}], x, \mu_0) > x, \forall x \in [0,1]$ assure quant à elle la convergence au point fixe $x = 1$ de l'équation (5), correspondant en fait à une probabilité d'erreur nulle. En notant $M = \mathrm{J}^{-1}(\mathrm{T}(1))$, la condition de stabilité du système autour de ce point fixe peut être déterminée de manière semi-analytique et est donnée par [3][5] : (i) si $T(1) = 1 : \lambda_2^R < e^{\frac{1}{2\sigma^2}} / \sum_j \rho_j(j-1)$, (ii) si $T(1) < 1 : \lambda_2^I e^{-\frac{M}{4}} + \lambda_2^R < \lambda_2^*(\sigma^2, \underline{\rho}) = e^{\frac{1}{2\sigma^2}} / \sum_{j=2}^{t_{r_{\max}}} \rho_j(j-1)$.

4 Optimisation du récepteur joint

Dans cette section sont décrites différentes stratégies d'optimisation : le récepteur conjoint optimal (optimisation DSCC) et les récepteurs compatibles avec des applications non conjointes (optimisation compatible non-DSCC).

4.1 Optimisation DSCC

L'optimisation des paramètres du code LDPC irrégulier peut s'écrire sous la forme d'un problème d'optimisation que l'on peut résoudre par programmation linéaire lorsque l'on considère le rendement comme fonction de coût à maximiser [3]. Pour $\rho(x)$ et σ^2 fixés, le problème d'optimisation s'écrit comme suit :

$$\underline{\lambda}_{opt} = \max_{\underline{\lambda}} [\underline{1/t_c}^\top, \underline{1/t_r}^\top]^\top \underline{\lambda} \text{ avec les contraintes : } (6)$$

[C_1] proportions : $1^\top \underline{\lambda} = 1$ et $\underline{1/tc}^\top \underline{\lambda}^R = \underline{1/tr}^\top \underline{\rho}$,
[C_2] convergence : $\mathrm{F}(\underline{\lambda}, x, \mu_0) > x$,
[C_3] condition de stabilité : $\lambda_2^I e^{-\frac{M}{4}} + \lambda_2^R < \lambda_2^*(\sigma^2, \underline{\rho})$.

Pour un rendement de codage cible R, par recherche successive sur σ^2 et $\rho(x)$, on obtient les paramètres optimaux $(\lambda(x), \rho(x))$, donnant le meilleur seuil de convergence $\delta^* = (E_b/N_0)_{opt}$ sous approximation gaussienne.

4.2 Optimisation compatible non-DSCC

Pour des raisons système pratiques, il est souvent souhaité que le codeur canal conçu puisse offrir de bonnes performances sur un canal donné avec ou sans interactions avec le décodeur source. Pour un canal de transmission donné standard (BABG par exemple), il est possible de déterminer les paramètres du code LDPC permettant d'approcher le plus proche possible de la capacité du canal. Ce code est alors un candidat naturel pour notre récepteur compatible. D'un point de vue théorique (approche asymptotique utilisant le seuil de convergence comme critère de performance), tous les mappings d'information sont équivalents. D'un point de vue pratique cependant, le mapping d'information consistant à allouer les bits d'information aux noeuds les plus connectés permet de minimiser la probabilité d'erreur bit moyenne à nombre d'itérations fini. Nous notons ce mapping $\mathcal{M}_0$. Compte tenu de la dissymétrie information/redondance dans le cadre source-canal conjoint, différents mappings conduisent à des seuils de convergence différents et donc rien ne garantie que le mapping $\mathcal{M}_0$ soit optimal du point de vue asymptotique pour des applications compatibles. La question est alors : quel est le mapping optimal $\mathcal{M}^*$? Quels sont ses performances ?

En considérant la méthode d'optimisation (6), le mapping optimal $\mathcal{M}^*$ en terme de seuil de convergence est obtenu simplement par ajout de la contrainte suivante :

$$[\mathrm{C}_4]\forall k = 2...t_{c_{\max}}, \lambda_k^I + \lambda_k^R = \lambda_k^{(a)}$$

$\lambda_k^{(a)}$ est la proportion de branches connectées à des noeuds de donnée de degré k pour le code optimisé sur le canal considéré (ici le canal BABG).

5 Résultats

Sans perte de généralités, le décodeur source considéré est un décodeur de codes source de type Huffman à entrées et sorties souples (Variable Length Code, SISO-VLC). Le code de source considéré en exemple et la source de symboles correspondante sont issus de [1] : on considérera donc dans la suite le code $\mathcal{C} = (00, 11, 010, 101, 0110)$. Les probabilités associées sont $\mathcal{P} = (0.33, 0.30, 0.18, 0.10, 0.09)$. L'entropie associée et la longueur moyenne d'un mot de code source sont données respectivement par $H = 2.14$ et $\bar{l} = 2.46$ bits/symboles. La redondance résiduelle est alors $R_s = H/\bar{l} = 0.86992$. Le décodeur de source SISO-VLC est un décodeur MAP sur le treillis bit du code [8]. La fonction de transfert $T(.)$ du décodeur source est estimée par méthode de Monte Carlo [7]. Pour le code source utilisé, l'EXIT chart simulée donne $T(1) \simeq 1$. On considérera par ailleurs des degrés concentrés pour $\rho(x)$ [6]. L'optimisation pour différentes valeurs de $\overline{\rho} = \rho j + (1-\rho)(j+1)$ est réalisée pour obtenir le code avec le meilleur seuil de convergence pour $t_{c_{\max}} = 30$, $R = 1/2$ et $R = 2/3$. Le rendement global est $R_T = R_s R = 0.43496$ pour le système avec un code de rendement $R = 1/2$ et $R_T = 0.5799$ pour celui avec le code de rendement $R = 2/3$. La limite de Shannon pour le canal BABG donne les rapports signal à bruit théoriques optimaux $E_b/N_0 = -0.0957$ dB et $E_b/N_0 = 0.5734$ dB pour $R_T = 0.43496$ et $R_T = 0.5799$ respectivement.

$\lambda(x)$	$R=1/2$ $\mathcal{M}^*$		$R=1/2$ $\mathcal{M}_0$		$R=2/3$ $\mathcal{M}_0 = \mathcal{M}^*$	
	I	R	I	R	I	R
λ_2	0.0080	0.2022	/	0.2102	0.0348	0.1345
λ_3	0.1641	0.0301	0.1777	0.0165	0.2011	/
λ_5	0.0435	/	0.0435	/	0.0008	/
λ_7	0.1400	/	0.1400	/	0.1964	/
λ_8	0.1044	/	0.1044	/	0.0926	/
λ_{30}	0.3036	/	0.3036	/	0.3388	/

TAB. 1 – Mappings avec optimisation compatible non-DSCC.

5.1 Résultats d'optimisation

Dans le cadre de l'optimisation DSCC, le seuil de convergence théorique $\delta^* = 1/(2*R_T*\sigma^2)$ du récepteur conjoint global est une fonction de $\overline{\rho}$. Par optimisation successive, on peut montrer qu'il existe, à la fois pour les rendements $R = 1/2$ et $R = 2/3$, une valeur minimal δ^* qui nous donne alors les paramètres optimaux $(\lambda(x), \rho(x))$ pour le récepteur conjoint. Pour $R = 1/2$ et $R = 2/3$, les valeurs optimales de $\overline{\rho}$ minimisant le seuil de convergence sont respectivement $\overline{\rho} = 7.91$ et $\overline{\rho} = 12.46$. Pour $R = 1/2$, en utilisant une représentation polynomiale [6], le profile du code pour les distributions des branches connectées à des noeuds de données est donné par $\lambda(x) = \lambda^I(x) + \lambda^R(x)$ avec $\lambda^I(x) = 0.1130x + 0.0830x^3 + 0.1201x^4 + 0.0588x^8 + 0.1044x^9 + 0.2516x^{29}$ et $\lambda^R(x) = 0.2216x + 0.0475x^2$. De même, pour $R = 2/3$, nous obtenons $\lambda^I(x) = 0.1130x + 0.0830x^3 + 0.1201x^4 + 0.0588x^8 + 0.1044x^9 + 0.2516x^30$ et $\lambda^R(x) = 0.1608x$

Dans le cadre de l'optimisation compatible non DSCC, l'optimisation des profiles des codes est réalisée tout d'abord pour le canal BABG pour $t_{c_{\max}} = 30$, $R = 1/2$ et $R = 2/3$ sous approximation gaussienne [6] en utilisant l'évolution de l'information mutuelle [5]. Pour chaque rendement de codage, les mappings $\mathcal{M}_0$ et $\mathcal{M}^*$ sont ensuite déterminés, en utilisant pour $\mathcal{M}^*$ la méthode d'optimisation (6) avec la contrainte additionnelle C_4. Les résultats sont donnés Table 1. Il s'avère que, en fonction du rendement considéré, il est possible de trouver un mapping $\mathcal{M}^*$ différent de $\mathcal{M}_0$ donnant le meilleur seuil de convergence pour un système compatible. Cependant, la différence semble localisée sur les degrés les plus bas, comme on peut le voir Table 1 pour $R = 1/2$.

5.2 Résultats de simulation

Pour le décodage itératif, la règle suivante est adoptée : le décodage itératif est arrêté si l'on trouve un mot de code valide (calcul du syndrome après chaque itération) ou le nombre maximum d'itérations autorisé est atteint (contrainte de délai). Pour nos simulations, le nombre d'itérations maximum est fixé arbitrairement à 150. Examinons maintenant les résultats obtenus pour des tailles de mots de code longues et moyennes. Dans les deux cas, la longueur des mots de codes est choisie de façon à avoir le même nombre de symboles sources par trame. Cela nous permet alors de comparer le taux d'erreur paquet (TEP), ainsi que le taux d'erreur bit (TEB) pour les différents rendements et récepteurs considérés. Principalement, quatre systèmes pour deux rendements différents seront comparés :

(i) le récepteur optimal source canal conjoint (DSCC-opt).

(ii) le récepteur source-canal conjoint utilisant un code op-

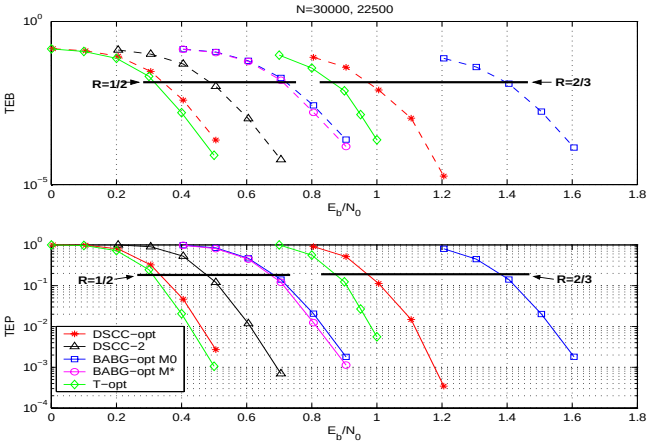


FIG. 2 – TEB et TEP pour $\{N = 30000, R = 1/2\}$ et $\{N = 22500, R = 2/3\}$.

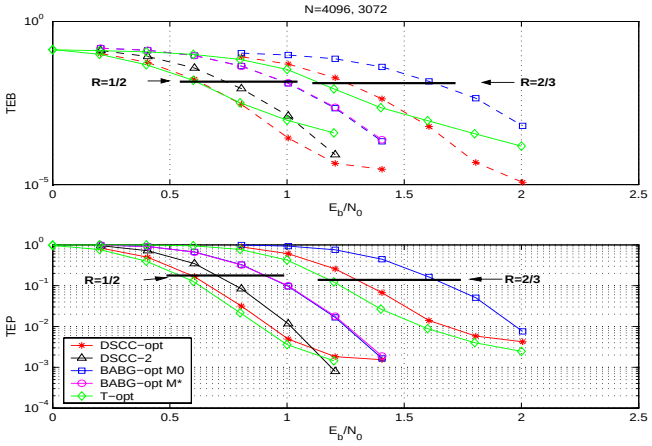


FIG. 3 – TEB et TEP pour $\{N = 4096, R = 1/2\}$ et $\{N = 3072, R = 2/3\}$.

timisé pour le canal BABG et utilisant le mapping $\mathcal{M}_0$ (BABG-opt $\mathcal{M}_0$).

(iii) le récepteur source-canal conjoint utilisant un code optimisé pour le canal BABG et utilisant le mapping $\mathcal{M}^*$ (BABG-opt $\mathcal{M}^*$) (seulement pour $R_T = 0.43496$).

(iv) le schéma tandem optimal (T-opt) : la source est supposée parfaitement compressée et codée avec un code de rendement R_T optimisé sur le canal BABG .

Par comparaison des courbes de TEP et TEB Figure 2, le récepteur DSCC-opt exhibe de meilleures performances que les récepteurs BABG-opt $\mathcal{M}^*$ et BABG-opt $\mathcal{M}_0$. Pour les deux rendements, le seuil de convergence est amélioré de 0.4 dB. Les récepteurs BABG-opt $\mathcal{M}^*$ et BABG-opt $\mathcal{M}_0$ ont des performances finalement proches. Pour le TEP, le récepteur DSCC-opt est à moins de 0.1 dB du récepteur T-opt pour $R_T = 0.43496$, et à 0.15 dB pour $R_T = 0.5799$: le but de notre récepteur source-canal conjoint semble atteint, *i.e* avoir une exploitation efficace de la redondance résiduelle pour s'approcher le plus près possible de la limite donnée par T-opt.

Les même observations que pour les tailles de mots de code longs peuvent être faîtes Figure 3. Cependant, contrairement à ce que l'on peut observer sur les courbes précédentes, un plancher d'erreur intervient aux alentours de 10^{-3} pour les courbes de TEP, et une réduction de la pente pour les courbes de TEB. Une étude des profiles des codes montre que la présence de noeuds de données de degrés 2 semble pénaliser la partie informative pour les schémas DSCC (faible robustesse). Pour vérifier cette hypothèse, l'optimisation DSCC pour $R = 1/2$ a été réalisée en imposant des degrés strictement supérieurs à 2 sur la partie informative. L'analyse de ce nouveau schéma DSCC-2 sur les Figures 2 et 3 montre une perte sur le seuil de convergence (degrés de liberté supprimés à l'optimisation), mais semble donner une amélioration du plancher d'erreur à taille moyenne.

6 CONCLUSION

Dans ce papier, différents schémas de décodage source-canal-conjoint impliquant un décodeur LDPC comme codeur canal sont comparés pour des tailles de mots de code et rendements différents. L'analyse des résultats montre que le récepteur optimal permet d'accroître les performances vis à vis des autres schémas. Pour les systèmes compatibles avec des applications non conjointes, le mapping naturel généralement utilisé semble offrir d'aussi bonnes performances que le mapping optimal. L'apparition de planchers d'erreur à taille moyenne incite à développer des études pour une optimisation à taille finie. Une future étude possible concerne l'allocation de la redondance entre source et code de canal de manière optimale.

Références

[1] R. Bauer, J. Hagenauer, "On variable length codes for iterative source/channel decoding," in IEEE Data Comp. Conf., pp. 272–282, Snowbird, UT, USA, Mar. 2001.

[2] X. Jaspar, L. Vandendorpe, "New iterative decoding of variable length codes with turbo codes," in IEEE Int. Conf. on Com., Paris, France, Jun. 2004.

[3] C. Poulliat, D. Declercq, C. LamyBergot, I. Fijalkow, "Analysis and optimization of irregular LDPC codes for joint sourcechannel decoding," IEEE Com. Letters, Accepted for publication, November 2004.

[4] S. ten brink, "Convergence behavior of iteratively de coded parallel concatenated codes," IEEE Trans. on Com., vol. 49, no. 10, pp. 1727–1737, Oct. 2001.

[5] A. Roumy, S. Guemghar, G. Caire, S. Verdù, "Design methods for irregular repeat-accumulate codes," IEEE Trans. on Inf. Theory, vol. 50, no. 8, pp. 1711–1727, Aug. 2004.

[6] S.Y. Chung, T.J. Richardson, R.L. Urbanke, "Analysis of sum product decoding of low density parity check codes using a gaussian approximation," IEEE Trans. on Inf. Theory, vol. 47, no.2, pp. 657–670, Feb. 2001.

[7] J. Hagenauer, R. Bauer, "The turbo principle in joint source channel decoding of variable length codes," in IEEE Inf. Theory Work., pp. 128–130, Cairns, Australia, Sept. 2001.

[8] V.B. Balakirsky, "Joint source and channel decoding with variable length codes," Prob. of Inf. Trans., vol. 27, no. 1, pp. 12–27, 2001.

Détection de phase à très faible rapport signal à bruit à l'aide d'un code de parité

Mathieu DERVIN[1], Marie-Laure BOUCHERET[2], Gilles MESNAGER[3], Alain DUCASSE[3]

[1] IRIT - ENST - TéSA
2 Rue Charles Camichel, BP 7122, 31071 Toulouse Cedex 7, France

[2] IRIT - ENSEEIHT
2 Rue Charles Camichel, BP 7122, 31071 Toulouse Cedex 7, France

[3]ALCATEL SPACE
26 av. Champollion, BP 1187, 31037 Toulouse, France
Mathieu.dervin@enst.fr

Résumé – De nombreuses études ont récemment montré l'avantage qu'on peut tirer du décodage canal pour l'amélioration de la synchronisation de phase porteuse à faible rapport signal à bruit. Lorsque la phase à estimer varie au court du temps (erreur de fréquence, bruit de phase), l'estimation de la phase avant la première étape de décodage pose toutefois problème lorsque les mots de code sont longs. Nous proposons ici un algorithme de synchronisation fondé sur l'introduction d'un code de parité simple et court. Nous montrons que pour une transmission continue et une modulation MDP8, le détecteur proposé permet de réduire sensiblement la gigue de phase à très faible point de fonctionnement et en présence de bruit de phase, par rapport aux algorithmes classiques. De faible complexité, il apparait notamment bien adapté au contexte des télécommunications par satellite.

Abstract – The improvement of carrier phase synchronization through the use of channel decoding information has been recently widely studied. However, an initial synchronization step is necessary prior to the first decoding iteration. When the phase to estimate is not constant (due to frequency error or phase noise), this first step is problematic on relatively long coded words. In this paper we propose to add a short single parity-check code to the transmission scheme for the synchronization purpose. The proposed phase detector is based on the soft symbol decoding of the additional code. Our simulation results show that the proposed synchronization scheme outperforms the classical non data-aided and decision directed algorithms for continuous transmissions with 8PSK modulation, in presence of phase noise. It allows indeed to reduce the phase jitter in considerable proportions. Requiring few additional complexity, it seems well adapted to satellite continuous transmissions.

1 Introduction

L'apparition de codes correcteurs d'erreurs très puissants tels que les turbo-codes et les codes LDPC (*low-density parity-check codes*) a eu des conséquences considérables sur la conception de systèmes de télécommunications pour des applications grand public, où tout gain sur le rapport signal à bruit de la transmission se traduit par une baisse du coût des terminaux. Les nouveaux standards de diffusion multimédia par satellite (DVB-RCS, DVB-S2 [1]) font ainsi intervenir des schémas de codage très performants, associés à des points de fonctionnement très faibles. En particulier, la norme DVB-S2 propose un code LDPC de rendement $3/5$ pour une modulation MDP8 au rapport signal à bruit $E_s/N_0 = 5.5\,dB$ sur le canal de transmission. Le problème de la synchronisation de porteuse est critique dans ce contexte : malgré la mauvaise qualité du signal reçu, sa démodulation doit être suffisament précise pour que le décodeur puisse converger.

Il est possible d'exploiter la redondance sur les données introduite par le codage canal afin d'améliorer la synchronisation [2, 3, 4, 5]. En présence de turbo-codes ou de codes LDPC, on s'est logiquement intéressé à des schémas itératifs où estimation de la porteuse et décodage sont imbriqués [2, 3]. De tels schémas ont montré de bonnes performances, à condition toutefois que le signal ait été suffisamment bien démodulé à la première itération pour permettre la convergence du décodeur. Ceci suppose que la phase à estimer est constante ou varie très faiblement, ce qui n'est pas le cas en pratique dans un contexte d'applications grand-public, où la contrainte de faible coût des terminaux se traduit notamment sur la précision des oscillateurs. En particulier dans des bandes de fréquences élevées (telles que la bande Ka envisagée pour de la diffusion multimédia), les imperfections des oscillateurs utilisés dans les terminaux utilisateurs se traduisent par du bruit de phase et une imprécision sur la fréquence porteuse. La variation de la phase qui en résulte ne peut pas être négligée sur des séquences codées relativement longues (8100 octets pour une trame DVB-S2). Il faut donc envisager une étape de synchronisation qui ne nécessite pas le décodage de la trame dans sa totalité. Pour cela, plusieurs approches sont possibles :

1 - Les techniques d'estimation classiques *non data aided* (NDA) et *decision directed* (DD) [6] n'exploitent aucune information de redondance pour la synchronisation de porteuse. A faible rapport signal à bruit, la gigue de phase en sortie de ce type d'estimateurs est trop importante pour garantir la convergence du décodeur.

2 - Les algorithmes dits *pilot aided* (PA), sont fondés sur la transmission de symboles de référence. Des symboles pilotes doivent être insérés dans les trames codées, en nombre suffi-

sant pour garantir une gigue acceptable, ce qui peut affecter fortement la capacité du système.
3 - Lorsqu'une information de décodage peut être obtenue à partir d'un nombre réduit de symboles reçus, on peut en tenir compte pour la synchronisation, tout en restant capable de suivre une variation de la phase relativement importante. En particulier, lorsque le codage canal met en jeu des codes convolutifs, des algorithmes utilisant des *tentative decisions* [4, 5] sur les symboles reçus ont montré de bonnes performances en présence de bruit de phase. En revanche, dans le cas général (par exemple, avec des codes LDPC), il n'est pas envisageable de décoder partiellement les données.
Afin d'assister la récupération de phase dans cette situation, nous nous proposons d'introduire dans la chaîne de transmission un code de parité simple et court [7]. Nous présentons ici un détecteur de phase qui tire profit de ce code supplémentaire. Nous traitons d'abord le problème du décodage souple d'une équation de parité à partir des symboles reçus. Introduite dans la vraisemblance de la phase, l'information de décodage permet d'obtenir un détecteur de phase relativement simple, que nous appellerons par la suite détecteur SDD (pour *soft decision directed*). Nous étudions ensuite les performances de ce détecteur dans des conditions réalistes, en prenant pour modèle l'interface-air définie par la norme DVB-S2 pour une transmission continue et une modulation de phase à 8 états (MDP8).

2 Détecteur de phase utilisant les équations de parité

On considère $k_p = 3p$ bits $b_0, \ldots, b_{3p-1}$ liés par une équation de parité : $b_0 \oplus \ldots \oplus b_{3p-1} = 0$. Après codage bit/symbole, on obtient p symboles complexes, notés $S_0, \ldots, S_{p-1}$. La somme modulo 2 des bits correspondant au symbole S_i est notée $\omega(S_i)$. Le canal additif gaussien peut être décomposé en deux composantes indépendantes de variance σ^2. Les symboles reçus en sortie du filtre adapté sont notés $y_0, \ldots, y_{p-1}$. En supposant la synchronisation d'horloge idéale, on a : $y_i = x_i e^{j\varphi} + g_i$, où φ représente l'erreur sur la phase porteuse (supposée constante sur les p symboles d'un mot de code) et g_i un échantillon de bruit gaussien.
On souhaite calculer les probabilités $P[S_i = s_k | y_0, \ldots, y_{p-1}]$ pour chacun des symboles transmis $(S_i)_{i=0,\ldots,p-1}$ et pour chaque symbole de la constellation $(s_k)_{k=0,\ldots,7}$.

Une démarche similaire à celle de Gallager dans ses travaux sur les codes LDPC [8] permet de montrer que :

$$P[S_i = s_k | y_0, \ldots, y_{p-1}] = \begin{cases} P_{i,k} F_{i,0} \text{ si } \omega(s_k) = 0 \\ P_{i,k} F_{i,1} \text{ si } \omega(s_k) = 1 \end{cases} \quad (1)$$

où

$$P_{i,k} = \lambda \exp\left(-\frac{|y_i - s_k|^2}{2\sigma^2}\right) \quad (2)$$

$$F_{i,0} = \frac{1 + \prod_{j \neq i} (1 - 2\Psi_{j,1})}{2} \quad (3)$$

$$F_{i,1} = \frac{1 - \prod_{j \neq i} (1 - 2\Psi_{j,1})}{2} \quad (4)$$

avec

$$\Psi_{i,1} = \sum_{k | \omega(s_k) = 1} P_{i,k} \quad (5)$$

– $P_{i,k}$ représente la probabilité canal $P[y_i | S_i = s_k]$. Le facteur multiplicatif λ est déterminé selon l'hypothèse

$$\sum_{k=0}^{7} P_{i,k} = 1 \quad (6)$$

– $F_{i,0}$ et $F_{i,1}$ représentent des probabilités *extrinsèques* sur le i-ème symbole transmis, calculées à partir des autres symboles impliqués dans la même équation de parité.

Une démarche inspirée de l'algorithme *Expectation-Maximisation* [2] et consistant à estimer la vraisemblance de la phase $\Lambda(\varphi)$ à partir de probabilités a posteriori sur les symboles transmis (S_i) conduit à l'expression suivante, à faible rapport signal à bruit [9] :

$$\Lambda(\varphi) = \text{Re}\left(\sum_{i=0}^{p-1} \delta_i^* y_i e^{-j\varphi}\right) \quad (7)$$

où les coefficients complexes δ_i sont des barycentres des symboles de la constellation, pondérés par les probabilités a posteriori issues du décodage du code de parité :

$$\delta_i = \sum_{k=0}^{7} P[S_i = s_k | y_0, ..., y_{p-1}]\, s_k \quad (8)$$

En dérivant la fonction de vraisemblance (7), on obtient l'expression d'un detecteur de phase :

$$u_{SDD}(i) = \text{Im}\left(\delta_i^* y_i e^{-j\varphi}\right) \quad (9)$$

3 Performances de l'algorithme proposé

3.1 Conditions de simulation

Afin de tester notre algorithme dans des conditions réalistes, nous avons considéré pour nos simulations l'interface-air définie dans la norme DVB-S2 [1] pour la voie aller d'une transmission satellite, à laquelle nous avons ajouté un code de parité supplémentaire, de rendement $\frac{k_p - 1}{k_p}$ (Fig. 1). En sortie du filtre adapté, la correction de phase est effectuée à l'aide d'une boucle du second ordre comprenant le détecteur proposé (9). Nous avons considéré pour notre étude la constellation MDP8 représentée en Fig. 2.

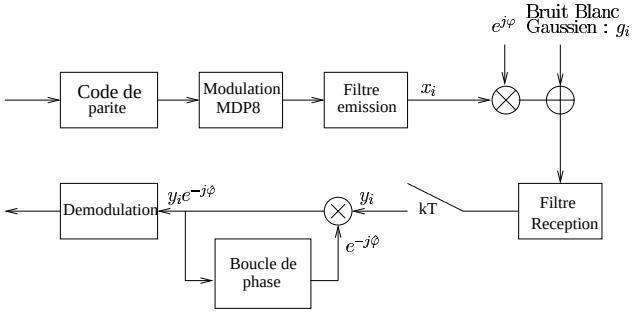


FIG. 1 – Modèle de simulation

Les performances de notre détecteur sont comparées ici à celles des détecteurs classiques [6] *non data aided* (NDA) et *decision directed* (DD) , dont les expressions respectives sont :

$$u_{NDA}(i) = \frac{1}{8} |y_k|^2 \sin\left(8 \arg\{y_k e^{-j\varphi}\}\right) \quad (10)$$

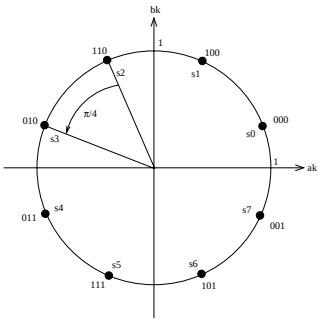


FIG. 2 – Constellation MDP8 avec codage de Gray

$$u_{DD}(i) = \mathrm{Im}\left(\hat{d}_i^* y_i e^{-j\varphi}\right) \tag{11}$$

où $\hat{d}_i$ est le symbole de la constellation obtenu en prenant une décision dure sur le symbole reçu y_i.
Remarque : le détecteur NDA proposé ici, inspiré de l'estimateur de Viterbi et Viterbi [6], est préféré au détecteur issu du maximum de vraisemblance [10], moins performant.

3.2 Influence du code de parité

L'apport du code de parité dans l'estimation de la phase peut se comprendre en observant la répartition des coefficients δ_i dans le plan complexe (Fig. 3) : lorsque le symbole reçu $S_{r,i}$ est fiable, δ_i tend vers le point de la constellation le plus vraisemblable ; le détecteur SDD (9) est alors équivalent au détecteur DD (11). En revanche, s'il y a une ambiguïté importante sur la phase du symbole $S_{r,i}$, le module de δ_i est atténué, ce qui modère l'influence de ce symbole peu fiable dans l'estimation de la phase.

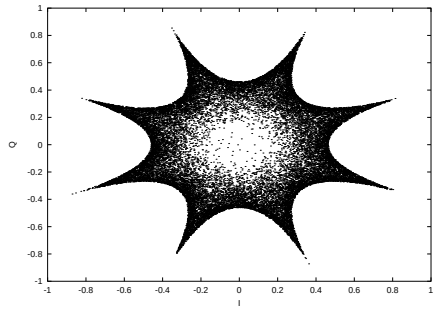


FIG. 3 – Constellation des coefficients δ_k ($k_p = 6$)

3.3 Réponse en boucle ouverte

La caractéristique du détecteur proposé (Fig. 4) est une courbe en S classique, dont la pente à l'origine (c'est à dire le gain) dépend du rapport signal à bruit E_s/N_0, où E_s représente l'énergie d'un symbole transmis et $N_0 = 2\sigma^2$ la densité spectrale de puissance (monolatérale) de bruit dans le canal. Une estimation du rapport signal à bruit est donc nécessaire pour régler la bande de bruit équivalente en boucle fermée.

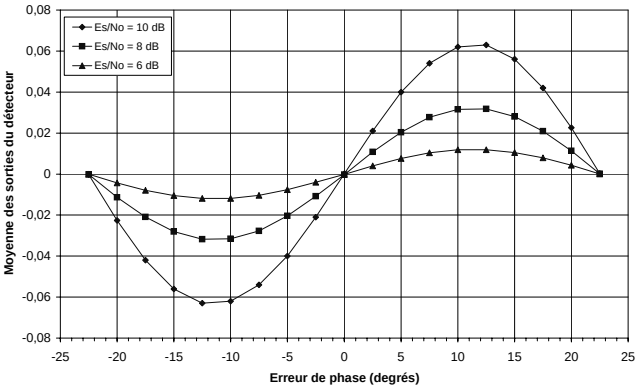


FIG. 4 – Détecteur SDD : courbes en S ($k_p = 6$)

3.4 Estimation d'une phase constante

Des simulations avec une erreur de phase constante permettent de comparer le comportement des estimateurs en présence de bruit gaussien (Fig. 5). Les boucles de phase sont comparées pour une bande de bruit équivalente $B_l T_s = 10^{-4}$, où T_s représente la durée d'un symbole. Il apparaît que le détecteur SDD est plus performant que les détecteurs DD et NDA, et que la gigue de phase se rapproche de la borne de Cramer-Rao lorsque le rendement du code de parité diminue. Notons qu'à $E_s/N_0 = 5.5\,dB$, avec un code de rendement $5/6$, l'algorithme SDD permet de réduire la variance d'estimation d'un facteur 11 par rapport à la boucle NDA et d'un facteur 16 par rapport à l'algorithme DD.

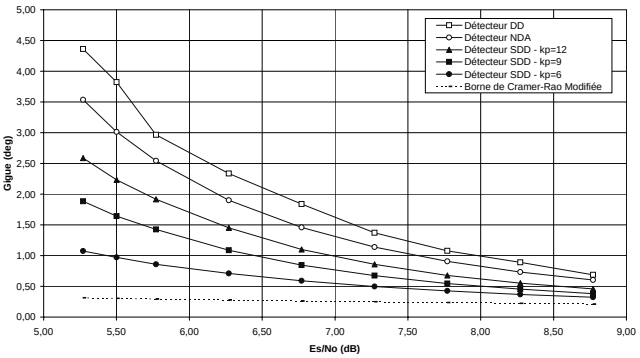


FIG. 5 – Estimation d'une phase constante : écart-type des estimées en degrés en fonction de E_s/N_0 ($B_l T_s = 10^{-4}$)

3.5 Suivi du bruit de phase

Lorsque la phase à estimer varie au cours du temps, un compromis sur la bande de bruit dans la boucle s'avère nécessaire. En effet, si une boucle étroite atténue mieux les effets du bruit gaussien, une large bande de bruit est préférable pour suivre les variations de la phase. Nous avons considéré dans nos simulations un modèle de bruit de phase conforme au gabarit à deux pentes proposé dans la norme DVB-S2 [1] pour l'évaluation de la synchronisation de phase dans une chaine de transmission à 25 Mbauds. Le détecteur SDD ayant montré de bonnes performances en présence de bruit gaussien, il permet d'élargir la bande de bruit en présence de bruit de phase. L'écart-type des erreurs d'estimation est représenté en fonction de la bande de

bruit en Fig. 6 pour un rapport signal à bruit de $7.8\ dB$ dans le canal de transmission. On vérifie que la bande de bruit optimale est plus large avec le détecteur SDD qu'avec les détecteurs DD et NDA. On gagne alors, dans le meilleur des cas (lorsque $k_p = 6$) un facteur 2 sur la variance de l'erreur par rapport à la boucle NDA, et un facteur 2.6 par rapport à la boucle DD.

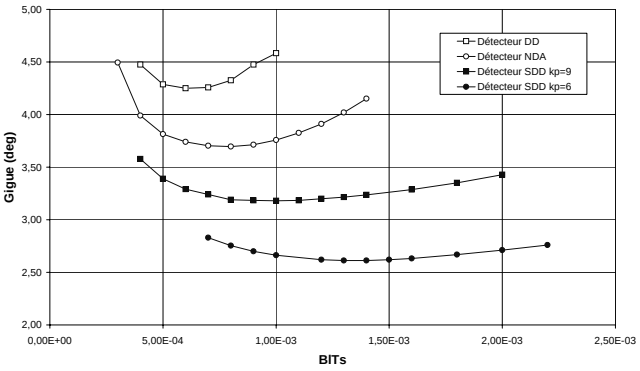


FIG. 6 – Performances en présence de bruit de phase : écart-type des estimées en fonction de la bande de bruit - $E_s/N_0 = 7.8\ dB$

Afin de comparer les performances des détecteurs en présence de bruit de phase et pour plusieurs rapports signal à bruit, nous avons représenté (Fig.7) les gigues de phases obtenues avec les boucles NDA, DD et SDD lorsque la bande de bruit est à sa valeur optimale. Afin de tenir compte de la perte d'efficacité en puissance due au code de parité supplémentaire, les rapports signal à bruit sont exprimés par rapport aux bits non codés. Pour un code de rendement $\frac{k_p-1}{k_p}$, on a :

$$\frac{E_b}{E_0} = \frac{E_s}{E_0} - 10\log_{10}\left(\frac{k_p-1}{k_p}\right) - 10\log_{10}(3) \qquad (12)$$

Comparé aux algorithmes classiques et à efficacité en puissance constante, l'algorithme SDD avec $k_p = 6$ permet de gagner 1.5 à $2\ dB$ sur le rapport signal à bruit en présence de bruit de phase. Notons que tout gain sur la variance de l'erreur d'estimation a pour conséquence une réduction significative de la probabilité de saut de cycle, qui est exponentielle par rapport à la gigue et l'ordre de la modulation.

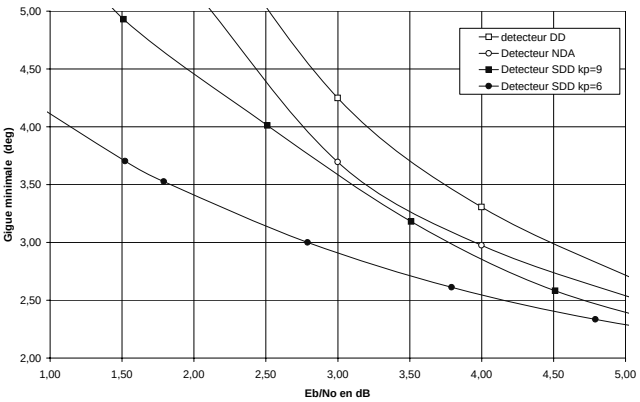


FIG. 7 – Performances en présence de bruit de phase : gigues minimales en fonction de E_b/N_0

4 Conclusion

Nos résultats montrent qu'il est possible d'améliorer sensiblement la synchronisation de phase porteuse à très faible rapport signal à bruit en tirant profit du décodage souple d'un code de parité. Le détecteur SDD présente de bonnes performances dans des conditions de réception réalistes et en présence de bruit de phase. Le prix à payer en terme d'efficacité spectrale n'est pas rédhibitoire, étant donnés les faibles rendement de codage requis pour la correction d'erreur à des points de fonctionnement aussi bas. La technique proposée semble notamment bien adaptée au contexte des télécommunications par satellite, compte tenu de la limitation des performances des oscillateurs dans des bandes de fréquences toujours plus élevées (bande Ka pour des systèmes DVB-S2).

Références

[1] Digital Video Broadcasting (DVB) - Second Generation Framing Structure, Channel Coding and Modulation Systems for Broadcasting, Interactive Services, News Gathering and other Broadband Satellite Applications, ETSI EN 302 307 v1.1.1 (2004-01).

[2] V. Lottici, M. Luise, "Embedding Carrier Phase Recovery Into Iterative Decoding of Turbo-Coded Linear Modulations", IEEE Transactions on Communications, Vol. 52, N° 4, avril 2004.

[3] H. Steendam, N. Noels, M. Moeneclaey, "Iterative Carrier Phase Synchronization for Low-Density Parity-Check Coded Systems", International Conference on Communications 2003, ICC'03, Anchorage, Alaska, 11-15 mai 2003, pp. 3120-3124.

[4] C. Morlet, "Démodulateur embarqué multiporteuses pour services multimédia par satellites", mémoire de thèse ENST, sept. 2000, ref. 00-E-027.

[5] C. Langlais, M. Hélard, "Using tentative decisions for carrier phase recovery of turbo-coded transmissions", Electronics Letters, 26/04/01, vol. 37, N° 9.

[6] H. Meyr, M. Moeneclaey, S.A. Fechtel, *Digital Communication Receivers*, Wiley&Sons ed. 1998.

[7] D.M.Rankin, T.A.Gulliver, "Single parity check product codes", IEEE Transactions on Communications, vol. 49, N° 8, pp. 1354-1362.

[8] R.G. Gallager, *Low-density parity-check codes* Cambridge, MA : MIT Press, 1963.

[9] M. Dervin, M.L. Boucheret, G. Mesnager, A. Ducasse, "Phase detection involving parity-check equations and suited to transmissions at low signal-to-noise ratio", International Conference on Communications 2003, ICC'05, Séoul, 16-20 mai 2005.

[10] M. Moeneclaey, G. de Jonghe, "ML oriented NDA carrier synchronization for general rotationally symmetric signal constellations", IEEE Transactions on Information Theory, vol. 29, pp. 543-551, 1983.

Séparation aveugle de signaux de type DS-CDMA à l'aide de techniques algébriques

Joséphine CASTAING, Lieven DE LATHAUWER

ETIS, UMR 8051 (CNRS, ENSEA, UCP)
6, avenue du Ponceau, BP 44, F 95014 Cergy-Pontoise Cedex, France
castaing@ensea.fr, delathau@ensea.fr

Résumé – Nous avons établi une nouvelle borne sur le nombre d'utilisateurs dans une cellule beaucoup moins contraignante que la borne actuelle. Nous proposons un algorithme pour résoudre le problème de l'extraction de signaux CDMA qui s'appuie sur la résolution d'un système de matrices à diagonaliser conjointement. Par ailleurs, dans le cas où les symboles d'information appartiennent à une constellation circulaire, nous avons constaté qu'il était possible de combiner la structure des données CDMA avec la propriété du module constant. Nous avons développé deux techniques à cet effet.

Abstract – We established a new, more relaxed bound on the number of users in a cell. We propose an algorithm to solve the problem of blind extraction of DS-CDMA signals. This algorithm is based on a joint diagonalisation problem. Furthermore, in the case where the transmitted symbols belong to a circular constellation, it is possible to combine the CDMA stucture with the constant modulus property. We developed two algorithms with this aim.

1 Introduction

Dans la troisième génération de téléphones mobiles, on utilise la solution d'étalement de spectre par codage DS-CDMA (Direct Sequence Code Division Multiple Access) afin de séparer les utilisateurs. De cette manière, tous les utilisateurs communiquent en même temps et dans la même bande de fréquence. D'autre part, cette technique possède l'avantage d'être robuste à un éventuel éblouissement. Nous nous plaçons dans le cadre particulier de l'extraction aveugle. L'objectif est d'extraire les signaux émis par les différents utilisateurs sans connaissance préalable sur ces signaux, les codes d'étalement ou les canaux. Pour cela, nous nous appuyons sur la structure algébrique très particulière des signaux reçus. Cette méthode a l'avantage de fonctionner sans exploiter les propriétés statistiques des signaux, c'est-à-dire par exemple dans une situation où les signaux émis sont courts, s'ils ne sont pas indépendants, ou encore si les canaux varient rapidement dans le temps.

2 Formulation du problème

Nous supposons que R utilisateurs émettent des séquences d'information de longueur K. Ces signaux sont étalés à l'aide de codes d'étalement de facteur d'étalement J. Le signal émis par un utilisateur r à l'instant $kJ + j$ s'écrit comme le produit du $k$ème symbole d'information s_{kr} et du $j$ème symbole de sa séquence d'étalement h_{jr} :

$$x_{kj,r} = s_{kr} h_{jr}. \tag{1}$$

Les signaux émis sont reçus sur un réseau de I antennes. Nous supposons que le mélange produit par le canal est linéaire et instantané, le signal reçu par l'antenne i à l'instant $kJ + j$ s'écrit par conséquent comme une combinaison linéaire des signaux émis :

$$y_{ijk} = \sum_{r=1}^{R} a_{ir} x_{kj,r} = \sum_{r=1}^{R} a_{ir} s_{kr} h_{jr} \tag{2}$$

Ecrivons cette équation sous forme tensorielle : on peut regarder l'élément y_{ijk} comme l'élément d'indice ijk d'un tenseur $\mathcal{Y}$ d'ordre 3 et de taille $I \times J \times K$. $\mathcal{Y}$ s'écrit alors :

$$\mathcal{Y} = \sum_{r=1}^{R} A_r \circ H_r \circ S_r \tag{3}$$

Où A_r, H_r, S_r désignent les vecteurs contenant la séquence d'étalement, les coefficients du canal et les symboles d'information correspondant à l'utilisateur r. L'opérateur $\circ$ désigne le produit externe. On appelle la décomposition (3) d'un tenseur de rang R en une somme de tenseurs de rang 1 (c'est-à-dire s'écrivant comme un produit externe de vecteurs) décomposition PARAFAC (Parallel Factor Analysis) ou encore CANDECOMP (Canonical Decomposition) [1, 2, 4]. Son utilisation pour résoudre le problème de l'extraction de données CDMA a été décrite dans [4], [3]. Les auteurs ont par ailleurs montré que cette décomposition était unique (à une matrice diagonale et une matrice de permutation près) tant que l'équation suivante était vérifiée :

$$k(\mathbf{A}) + k(\mathbf{H}) + k(\mathbf{S}) \geq 2(R + 1). \tag{4}$$

Où $\mathbf{A}$ (resp. $\mathbf{H}$, resp. $\mathbf{S}$) désigne la matrice contenant les vecteurs A_r, pour $r \in [1, R]$ (resp. H_r, resp. S_r) et où $k(\mathbf{A})$ désigne le "rang de Kruskal" de la matrice $\mathbf{A}$ défini comme le nombre maximal tel que toute sous-matrice de $\mathbf{A}$ de k colonnes ait des colonnes linéairement indépendantes.

Nous avons donné une preuve constructive qu'il était possible d'extraire les signaux pour une condition moins sévère sur le nombre d'utilisateurs. Dans le cas où le nombre d'antennes et le facteur d'étalement sont inférieurs au nombre d'utilisateurs , lui même inférieur au nombre d'échantillons ($I, J \leq R \leq K$), la borne de Kruskal sur le nombre d'utilisateurs dépend linéairement du nombre d'antennes et du facteur d'étalement, nous verrons que la nouvelle borne dépend de manière quadratique de ces deux paramètres.

Pour un développement plus étendu, il est possible de se référer à [8, 10]. Ordonnons dans un premier temps les colonnes du tenseur $\mathcal{Y}$ afin d'obtenir une expression des observations sous forme matricielle. Pour cela on extrait tout d'abord la tranche d'indice r de $\mathcal{Y}$ dans sa troisième dimension, on range les colonnes de cette matrice les unes derrière les autres, puis on place le vecteur obtenu dans la colonne d'indice r dans la matrice $\mathbf{Y}$. Cette matrice peut s'écrire sous la forme :

$$\mathbf{Y} = (\mathbf{A} \odot \mathbf{H})\ \mathbf{S}^T, \tag{5}$$

où $\odot$ désigne le produit de Khatri-Rao encore appelé produit de Kronecker par colonnes.

Nous savons qu'il existe une autre décomposition de $\mathbf{Y}$, sous la forme du produit d'une matrice unitaire, d'une matrice diagonale positive et de la transposée conjuguée d'une matrice unitaire (décomposition en valeurs singulières ou svd) :

$$\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^H. \tag{6}$$

Nous pouvons déduire des équations (5) et (6) qu'il existe une matrice $\mathbf{F}$ non singulière et a priori inconnue qui vérifie :

$$\left\{ \begin{array}{rcl} \mathbf{A} \odot \mathbf{H} & = & \mathbf{U}\mathbf{D}\mathbf{F} \\ \mathbf{S}^T & = & \mathbf{F}^{-1}\mathbf{V}^H \end{array} \right. , \tag{7}$$

Il suffit par conséquent d'estimer la matrice $\mathbf{F}$ pour retrouver les matrices $\mathbf{A}$, $\mathbf{H}$ et $\mathbf{S}$ recherchées. En effet, on obtient la matrice $\mathbf{S}$ à partir de l'équation (7) de manière directe par $\mathbf{S} = \mathbf{V}^*\mathbf{F}^{-T}$. D'autre part, définissons l'opérateur $vec()$ qui transforme une matrice en un vecteur de la manière suivante : soit $\mathbf{X} = [X_1, X_2, \ldots, X_N]$ une matrice de taille $M \times N$, alors $vec(X) = \left[X_1^T, X_2^T, \ldots, X_N^T\right]^T$. $Unvec(.)$ désigne l'opérateur inverse de $vec(.)$.
On range chaque colonne de $\mathbf{A} \odot \mathbf{H}$ dans une matrice $\mathbf{N_i}$ de taille $R \times R$:

$$\mathbf{N_i} = unvec\left(A_i \otimes H_i\right) = H_i A_i^T$$

Cette matrice est de rang un. A un facteur d'échelle près, H_i est le vecteur singulier de gauche de $\mathbf{N_i}$ correspondant à la plus grande valeur singulière, et A_i est le conjugué du vecteur singulier de droite correspondant à la plus grande valeur singulière.
Il s'agit maintenant d'évaluer la matrice $\mathbf{F}$. Nous allons pour cela utiliser la première égalité du système (7) et exploiter la structure particulière de la matrice $\mathbf{A} \odot \mathbf{H}$.

Soit $\mathbf{E_r}$ la matrice de taille $I \times J$ obtenue en extrayant la $r$ème colonne de la matrice $\tilde{\mathbf{U}} = \mathbf{U}\mathbf{D}$ et en la réordonnant sous forme d'une matrice :

$$\mathbf{E_r} = unvec(\tilde{\mathbf{U}}_r) \tag{8}$$

On utilise l'équation (7) pour exprimer $\mathbf{E_r}$ en fonction de $\mathbf{A}$, $\mathbf{H}$ et $\mathbf{F}$.

$$\begin{aligned} \mathbf{E_r} &= unvec\left(\left((\mathbf{A} \odot \mathbf{H})\,\mathbf{F}^{-1}\right)_r\right) && (9) \\ &= \sum_{k=1}^{R} \left(H_k A_k^T\right)\left(\mathbf{F}^{-1}\right)_{kr}. && (10) \end{aligned}$$

Dans la suite, nous allons utiliser le théorème suivant.
Théorème 1
Soit la fonction Φ *:* $(\mathbf{X}, \mathbf{Y}) \in \mathbb{C}^{I\times J} \times \mathbb{C}^{I\times J} \longmapsto \Phi(\mathbf{X}, \mathbf{Y}) \in \mathbb{C}^{I\times J\times I\times J}$ *définie pr :*
$(\Phi(\mathbf{X}, \mathbf{Y}))_{ijkl} = x_{ij}y_{kl} + y_{ij}x_{kl} - x_{il}y_{kj} - y_{il}x_{kj}$ *pour tout* $(i, j, k, l) \in \mathbb{N}_I \times \mathbb{N}_J \times \mathbb{N}_I \times \mathbb{N}_J$.
Soit $\mathbf{X} \in \mathbb{C}^{I\times J}$*, alors* $\Phi(\mathbf{X}, \mathbf{X}) = 0$ *si et seulement si le rang de* $\mathbf{X}$ *est au plus un .*

Preuve
Le cas $\mathbf{X} = 0$ est évident.
Soit $\mathbf{X}$ une matrice de rang un. Il existe deux vecteurs $\mathbf{u}$ et $\mathbf{v}$ tels que $x_{ij} = u_i v_j$. Alors $(\Phi_{xx})_{ijkl} = 2(u_i v_j u_k v_l - u_i v_l u_k v_j) = 0$.
Soit maintenant $\mathbf{X}$ une matrice vérifiant $\Phi(\mathbf{X}, \mathbf{X}) = 0$. La décomposition en valeurs singulières de $\mathbf{X}$ est donnée par $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H$. Nous obtenons donc :

$$\begin{aligned} x_{ij}x_{kl} - x_{il}x_{kj} &= \sum_{r,s} \sigma_r\sigma_s u_{ir}u_{ks}(v_{jr}v_{ls} - v_{js}v_{lr})^* \\ &= \sum_{r\neq s} \sigma_r\sigma_s u_{ir}u_{ks}(v_{jr}v_{ls} - v_{js}v_{lr})^*. \end{aligned}$$

Les matrices $\mathbf{U}$ et $\mathbf{V}$ sont unitaires, donc les tenseurs dont les éléments sont les $u_{ir}u_{ks}(v_{jr}v_{ls})^*$ et $u_{ir}u_{ks}(v_{js}v_{lr})^*$, $r \neq s$ sont linéairement indépendants. Par conséquent, $\sigma_r\sigma_s = 0$ si $r \neq s$ et $\mathbf{X}$ et $\mathbf{\Sigma}$ sont de rang un.

La fontion Φ permet de construire un ensemble de R^2 tenseurs Φ_{rs} définis par :

$$\begin{aligned} \Phi_{rs} &= \Phi\left(\mathbf{E_r}, \mathbf{E_s}\right) \\ &= \Phi\left(\sum_{p=1}^{R} H_p A_p^T (\mathbf{F}^{-1})_{pr},\ \sum_{q=1}^{R} H_q A_q^T (\mathbf{F}^{-1})_{qs}\right). \end{aligned}$$

En vertu de la bilinéarité de Φ, nous pouvons écrire :

$$\Phi_{rs} = \sum_{p,q=1}^{R} (\mathbf{F}^{-1})_{pr}(\mathbf{F}^{-1})_{qs}\Phi\left(H_p A_p^T,\ H_q A_q^T\right). \tag{11}$$

Soit $\mathbf{B}$ une matrice symétrique de taille $R \times R$ telle que :

$$\sum_{r,s=1}^{R} \Phi_{rs}\mathbf{B}_{rs} = 0. \tag{12}$$

En remplaçant Φ_{rs} par son expression, nous obtenons :

$$\sum_{r,s=1}^{R}\sum_{p,q=1}^{R} (\mathbf{F}^{-1})_{pr}(\mathbf{F}^{-1})_{qs}\Phi\left(H_p A_p^T,\ H_q A_q^T\right)\mathbf{B}_{rs} = 0.$$

D'après le théorème 1, $\Phi\left(H_pA_p^T,\ H_pA_p^T\right) = 0$ pour tout p in $[1,\ldots,R]$, donc :

$$\sum_{r,s=1}^{R}\sum_{\substack{p,q=1\\p\neq q}}^{R}(\mathbf{F}^{-1})_{pr}(\mathbf{F}^{-1})_{qs}\mathbf{B}_{rs}\Phi\left(H_pA_p^T,\ H_qA_q^T\right)=0.$$

D'autre part, Φ and $\mathbf{B}$ étant symétriques, nous pouvons écrire :

$$\sum_{r,s=1}^{R}\sum_{\substack{p,q=1\\p<q}}^{R}(\mathbf{F}^{-1})_{pr}(\mathbf{F}^{-1})_{qs}\mathbf{B}_{rs}\Phi\left(H_pA_p^T,\ H_qA_q^T\right)=0. \tag{13}$$

Supposons maintenant que les tenseurs $\left(\Phi\left(H_pA_p^T,\ H_qA_q^T\right)\right)_{p<q}$ soient linéairement indépendants. Alors nous pouvons déduire de l'équation (13) le résultat suivant :

$$\sum_{r,s=1}^{R}(\mathbf{F}^{-1})_{pr}(\mathbf{F}^{-1})_{qs}\mathbf{B}_{rs}=\lambda_{pq}\delta_{pq}, \tag{14}$$

où δ désigne le symbole de Kronecker ($\delta_{pq}=1$ si $p=q$, $\delta_{pq}=0$ sinon).
L'équation (14) peut être écrite sous forme matricielle :

$$\mathbf{B}=\mathbf{F}\mathbf{\Lambda}\mathbf{F}^T, \tag{15}$$

Dans cette expression, $\mathbf{\Lambda}$ est une matrice diagonale dont les éléments sont les λ_{pp}, $p\in[1,R]$.
Par ailleurs, toute matrice $\mathbf{B}$ s'écrivant sous la forme $\mathbf{F}\mathbf{\Lambda}\mathbf{F}^T$ avec $\mathbf{\Lambda}$ une matrice diagonale quelconque vérifie l'équation (12). Le noyau de $\mathbf{P}=[vec(\Phi_{11}),vec(\Phi_{12}),\ldots,vec(\Phi_{RR})]$ contient R matrices.
La matrice $\mathbf{F}$ peut alors être évaluée à l'aide d'une diagonalisation simultanée du système

$$\mathbf{B_r}=\mathbf{F}\mathbf{\Lambda_r}\mathbf{F}^T,\forall r\in[1,R], \tag{16}$$

dans lequel les matrices $\mathbf{\Lambda_r},\ r\in[1,\ldots,R]$ sont des matrices diagonales inconnues a priori.
S'il est possible d'évaluer $\mathbf{F}$ à partir de deux équations de ce système (en effet $\mathbf{F}$ est la matrice des vecteurs propres de la matrice $\mathbf{B_i}{\mathbf{B_j}}^{-1}, i\neq j$), il est bien plus judicieux de choisir de conserver l'ensemble des équations et de diagonaliser conjointement toutes les matrices $\mathbf{B_r}, r\in[1,R]$. Il existe différentes manières de procéder, nous parlerons de l'algorithme des moindres carrés alternés et de l'itération de QZ étendue [7]. Cette approche est une généralisation de l'algorithme de Cardoso pour l'analyse en composantes indépendantes [10].
D'autre part, on montre que l'unicité de la matrice $\mathbf{F}$ est assurée si le nombre d'utilisateurs R vérifie :

$$R(R-1)\leq\frac{1}{2}(I^2-I)(J^2-J). \tag{17}$$

Cela signifie que, si le nombre d'utilisateurs respecte cette condition, il est toujours possible d'extraire les signaux, et cela même si la borne de Kruskal (4) est dépassée.

3 Contrainte du module constant

Nous avons pu voir dans le paragraphe précédent qu'il était possible d'estimer la matrice $\mathbf{F}$ à partir de la première équation du système (7). Une autre solution consiste à choisir d'exploiter la seconde équation. En effet, cette expression est celle d'un problème de séparation aveugle de sources dans lequel les observations seraient contenues dans la matrice $\mathbf{V}^H$, et où $\mathbf{F}$ serait la matrice de mélange. Si l'on suppose que les sources sont de module constant, un tel problème peut être résolu à l'aide de l'algorithme analytique du module constant (ACMA) décrit par Van der Veen et Paulraj dans [7]. Cet algorithme a été préféré à d'autres car il possède la particularité de mener à un problème de diagonalisation simultanée :

$$\mathbf{M_r}=\mathbf{F}^{-H}\mathbf{\Omega_r}\mathbf{F}^{-1},\forall r\in[1,R], \tag{18}$$

dans lequel les matrices $(\mathbf{M_r})_{r\in[1,R]}$ sont obtenues à partir des données et où les matrices $(\mathbf{\Omega_r})_{r\in[1,R]}$ sont des matrices digonales.
Ce système est très similaire au système (16), nous proposons donc de résoudre les deux problèmes de manière conjointe :

$$\left\{\begin{array}{rcl}\mathbf{B_r}&=&\mathbf{F}\mathbf{\Lambda_r}\tilde{\mathbf{F}}\\ \mathbf{M_r}&=&\tilde{\mathbf{F}}^{-*}\mathbf{\Omega_r}\mathbf{F}^{-1}\end{array}\right.,\forall r\in[1,R]\ . \tag{19}$$

Nous nous inspirons pour cela de l'algorithme des moindres carrés alternés ou encore de l'itération de QZ [4, 7].

4 Simulations

Nous illustrons dans cette section les performances de l'algorithme développé. Les simulations ont été réalisées pour I=5 antennes, un facteur d'étalement J=4, K=200 symboles QPSK, et R=7 utilisateurs. Dans ce cas, la borne de Kruskal est dépassée.

Sur la figure (1) nous avons représenté le Taux d'Erreurs Symbole médian en fonction du Rapport Signal sur Bruit pour l'égaliseur non aveugle MMSE, pour l'algorithme consistant à appliquer un algorithme ALS directement aux données (DALS) et pour notre algorithme utilisant une diagonalisation simultanée par itération de QZ étendue (SD-QZ). Les deux derniers algorithmes donnent des résulats équivalents.
Sur la figure (2), nous avons représenté le Taux d'Erreurs Symbole moyen. Pour un RSB supérieur à 10dB, l'algorithme DALS ne converge pas vers le minimum global dans environ 10 % des cas, ce qui explique la différence entre la courbe médiane et la courbe moyenne.
La figure (3) représente le coût de calcul nécessaire pour l'algorithme DALS et l'algorithme SD-QZ. Afin de converger avec de plus grandes chances vers l'optimum, l'algorithme DALS est initialisé à l'aide de 10 valeurs, ce qui augmente son coût de calcul. Sur la figure, on a représenté le temps de calcul nécessaire pour l'initialisation donnant la meilleure peformance. D'autre part, on dispose pour l'algorithme SD-QZ d'une valeur initiale qui permet de se rapprocher rapidement du résultat.

5 Conclusion

Nous avons montré qu'il était possible d'extraire les séquences d'information d'un mélange de signaux étalés à l'aide d'un code DS-CDMA en s'appuyant sur la structure algébrique de ces signaux. L'algorithme proposé à cette fin mène à un problème de diagonalisation simultanée que l'on peut résoudre de différentes manières. Afin d'améliorer les performances, on

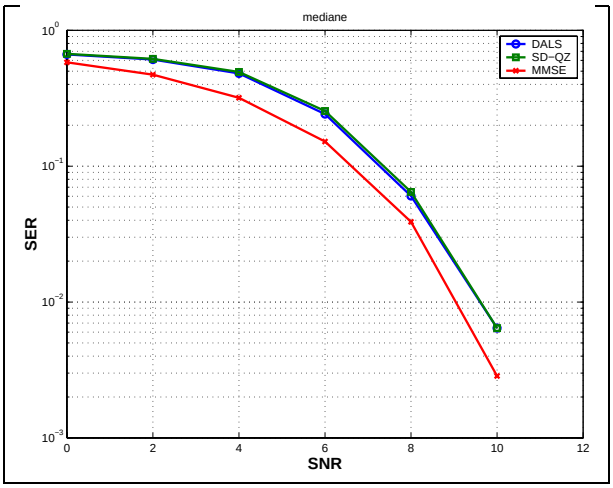


FIG. 1 – SER versus SNR pour I=4, J=4, K=200, R=7

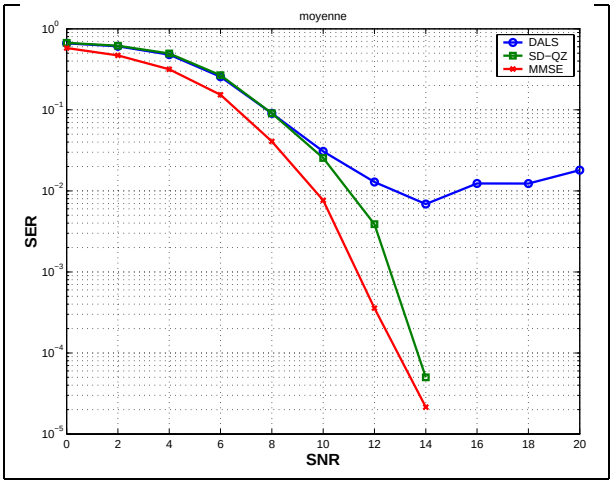


FIG. 2 – SER versus SNR pour I=4, J=4, K=200, R=7

peut par ailleurs combiner le système d'équations obtenu par cet algorithme avec le système d'équations obtenu en imposant la constance du module des sources. Nous avons d'autre part établi une nouvelle condition sur le nombre d'utilisateurs beaucoup moins contraignante que la condition de Kruskal.

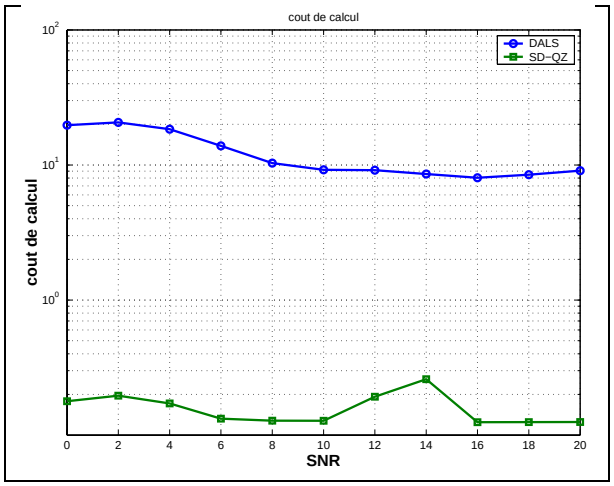


FIG. 3 – temps de calcul versus SNR pour I=4, J=4, K=200, R=7

Références

[1] L. De Lathauwer, B. De Moor and J. Vandewalle, "Computation of the Canonical Decomposition by Means of a Simultaneous Generalized Schur Decompositition," *SIAM J. Matrix Anal. Appl.*, Vol. 26, pp. 295–327, 2004.

[2] R.A. Harshman and M.E. Lundy, "The PARAFAC Model for Three-Way Factor Analysis and Multidimensional Scaling," in : H.G. Law, C.W. Snyder, J.A. Hattie and R.P. McDonald (Eds.), *Research Methods for Multimode Data Analysis*, pp. 122-215. Address :Praeger, N.Y., 1984.

[3] J.B. Kruskal, "Three-way Arrays : Rank and Uniqueness of Trilinear Decompositions, with Application to Arithmetic Complexity and Statistics," *Lin. Alg. Appl.*, Vol. 18, pp. 95–138, 1977.

[4] N. Sidiropoulos, G. Giannakis and R. Bro, "Blind PARAFAC Receivers for DS-CDMA systems," *IEEE Trans. on Signal Processing*, vol. 48, pp. 810–823, Mar. 2000.

[5] N. Sidiropoulos, R. Bro and G. Giannakis, "Parallel Factor Analysis in Sensor Array Processing," *IEEE Trans. on Signal Processing*, Vol. 48 (8), pp. 2377–2388, Aug. 2000.

[6] N. Sidiropoulos and X. Liu, "Identifiability Results for Blind Beamforming in Incoherent Multipath with Small Delay Spread," *IEEE Trans. on Signal Processing*, vol.49 (1), pp. 228–236, Jan. 2001.

[7] A.J. van der Veen and A. Paulraj, "An Analytical Constant Modulus Algorithm," *IEEE Trans. on Signal Processing*, vol. 44 (5), pp. 1136–1155, May 1996.

[8] L. De Lathauwer "A link between the canonical decomposition in multilinear algebra and simultaneous matrix diagonalization", *Internal Report*, ETIS Cergy France, 2004.

[9] J-F. Cardoso, A. Souloumiac, "Blind Beamforming for Non Gaussian Signals," *IEEE-Proceedings-F*, vol.140, no6, pp. 362–370, dec. 1993.

[10] J-F. Cardoso, "Super-symmetric decomposition of the fourth-order cumulant tensor. Blind identification of more sources than sensors," *Proc. ICASSP*, pp. 3109–3112, 1991.

Reconnaissance des formes par méthodes à noyau dans le domaine temps-fréquence

Paul HONEINE[1], Cédric RICHARD[2], Patrick FLANDRIN[3]

[1]Sonalyse, Pist Oasis, 131 impasse des palmiers, 30319 Alès, France

[2]ISTIT (FRE CNRS 2732) - Équipe Modélisation et Sûreté des Systèmes
Université de Technologie de Troyes, 12 rue Marie Curie, BP 2060, 10010 Troyes cedex, France

[3]Laboratoire de Physique (UMR CNRS 5672), École Normale Supérieure de Lyon
46 allée d'Italie, 69364 Lyon, France

paul.honeine@mail.sonalyse.fr, cedric.richard@utt.fr, patrick.flandrin@ens-lyon.fr

Résumé –
Au cours de la dernière décennie, de nombreuses méthodes pour l'analyse et la classification de données fondées sur la théorie des noyaux reproduisants ont été proposées. Elles constituent une source de progrès, tant au niveau de la complexité algorithmique que des performances atteintes. On montre dans cet article qu'un choix approprié de noyau reproduisant permet à ces méthodes d'opérer sur des représentations temps-fréquence. Ce lien ouvre de multiples perspectives au domaine de l'analyse des signaux non-stationnaires puisqu'il lui permet d'accéder aux plus récentes avancées méthodologiques et théoriques en matière de reconnaissance des formes et de théorie de l'apprentissage.

Abstract –
During the last decade, many important advances have been made in the field of pattern recognition with the theory of reproducing kernels. This unified view has led to new algorithms with improved performance and lower computational complexity. In this paper, we show that some specific reproducing kernels allow these algorithms to operate on time-frequency representations. This link offers new perspectives in the field of non-stationary signal analysis since it provides an access to the most recent methodological and theoretical developments of pattern recognition and statistical learning theory.

1 Introduction

L'analyse des signaux non-stationnaires nécessite la mise en œuvre d'outils spécifiques parmi lesquels émergent les distributions temps-fréquence et temps-échelle. Celles-ci constituent une alternative intéressante aux techniques d'analyse spectrale classiques en incluant explicitement une dimension supplémentaire d'évolution temporelle. Elles offrent de plus une grande variété d'espaces de représentation, à la mesure de la diversité des objectifs visés par les utilisateurs. Par exemple, il est possible de privilégier l'intelligibilité de l'information délivrée par une représentation en y limitant les manifestations de bruits et autres termes interférentiels nuisant à sa lisibilité [1, 2, 3]. On peut encore être à la recherche d'un espace de représentation favorisant la résolution d'un problème de décision [4], point de rencontre entre l'analyse des signaux non-stationnaires et les théories de la décision et de l'apprentissage.

Au cours de la dernière décennie, de multiples algorithmes pour l'analyse et la classification de données fondés sur la théorie des espaces de Hilbert à noyau reproduisant ont été développés [5]. Ils reposent sur un principe central appelé *coup du noyau*, exploité pour la première fois avec les Support Vector Machines ou SVM [6], qui confère un caractère non-linéaire à nombre de méthodes originellement linéaires [7]. Parmi elles, on compte par exemple des généralisations non-linéaires de l'analyse en composantes principales [8] et des analyses discriminantes de Fisher [9] et généralisée [10], respectivement désignées dans la littérature par kernel-PCA, kernel-FDA et kernel-GDA. Les méthodes à noyau constituent une source de progrès, tant au niveau du coût calculatoire que des performances atteintes, ces dernières étant par ailleurs encadrées par des bornes théoriques sans cesse affinées dans le cadre de la théorie statistique de l'apprentissage [11, 12]. Clairement, le domaine de l'analyse temps-fréquence ne s'est pas encore approprié ces techniques et la théorie qui les accompagne bien que, paradoxalement, la recherche d'espaces de représentation mettant en évidence une information pertinente y fasse l'objet de toutes les attentions. Aujourd'hui encore, peu de travaux combinent en effet méthodes à noyau et analyse temps-fréquence. Citons toutefois [13, 14] où, pour résoudre un problème de classification de signaux non-stationnaires, les auteurs associent les SVM à un noyau défini à partir de représentations temps-fréquence paramétrées selon [2].

A caractère plus générique, le présent article a pour objectif de montrer que les méthodes à noyau les plus performantes et les plus diverses peuvent être mises en œuvre dans le plan temps-fréquence grâce à un choix approprié de noyau reproduisant. L'analyse temps-fréquence des informations extraites par ces algorithmes est également abordée, contribuant ainsi à justifier le choix de ces espaces de représentation. Mais au préalable, il convient de rappeler les principaux éléments de théorie soutenant les méthodes à noyau.

2 Rappels du cadre théorique

Les méthodes à noyau sont pour la plupart issues d'algorithmes linéaires auxquels on a pu appliquer les deux résultats clés que sont le coup du noyau et le théorème de représentation [15]. Leur présentation fait ici suite à celle de la notion de noyau.

2.1 Noyau défini positif, noyau reproduisant

On considère un sous-espace $\mathcal{X}$ de $\mathcal{L}_2(\mathbb{C})$, l'espace des signaux d'énergie finie à valeurs dans $\mathbb{C}$, que l'on munit du produit scalaire canonique $\langle x_i, x_j \rangle = \int_t x_i(t)\,\overline{x}_j(t)\,dt$ et de la norme associée. Un noyau désigne une fonction $\kappa(x_i, x_j)$ de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{C}$, à symétrie hermitienne. On rappelle les deux définitions fondamentales suivantes [5, 11].

Définition 1. *Un noyau κ est dit défini positif sur $\mathcal{X}$ s'il vérifie*

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i\,\overline{a}_j\,\kappa(x_i, x_j) \geq 0 \qquad (1)$$

pour tout $n \in \mathbb{N}$, $x_1, \ldots, x_n \in \mathcal{X}$ et $a_1, \ldots, a_n \in \mathbb{C}$.

Définition 2. *Soit $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$ un espace de Hilbert constitué de fonctions de $\mathcal{X}$ dans $\mathbb{C}$. La fonction $\kappa(x_i, x_j)$ de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{C}$ est le noyau reproduisant de $\mathcal{H}$, sous réserve qu'il en admette un, si et seulement si*

- *la fonction $\kappa_{x_i} : x_j \mapsto \kappa_{x_i}(x_j) = \kappa(x_i, x_j)$ appartient à $\mathcal{H}$, quel que soit $x_i \in \mathcal{X}$ fixé ;*
- *on a $\psi(x_i) = \langle \psi, \kappa_{x_i} \rangle_{\mathcal{H}}$ pour tout $x_i \in \mathcal{X}$ et $\psi \in \mathcal{H}$.*

On démontre que tout noyau défini positif κ est le noyau reproduisant d'un espace de Hilbert de fonctions de $\mathcal{X}$ dans $\mathbb{C}$ en exhibant directement celui-ci. On considère pour cela l'espace vectoriel $\mathcal{H}_0$ engendré par les fonctions $\{\kappa_x\}_{x \in \mathcal{X}}$, auquel on associe le produit scalaire

$$\langle \psi, \phi \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i\,\overline{b}_j\,\kappa(x_i, x_j), \qquad (2)$$

avec $\psi = \sum_{i=1}^{n} a_i\,\kappa_{x_i}$ et $\phi = \sum_{j=1}^{m} b_j\,\kappa_{x_j}$ appartenant à l'espace $\mathcal{H}_0$. Il s'agit donc là d'un espace pré-hilbertien, que l'on complète conformément à [5] de sorte que toute suite de Cauchy y converge. On aboutit ainsi à l'espace de Hilbert $\mathcal{H}$ à noyau reproduisant κ recherché. Réciproquement, on peut démontrer que tout noyau reproduisant est défini positif [5]. A titre d'exemple, parmi les noyaux reproduisants classiques, citons le noyau gaussien $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2/2\sigma_0^2)$, où σ_0^2 est un paramètre à fixer. D'autres noyaux, ainsi que des règles permettant de les combiner, sont présentés dans [12, 16].

2.2 Coup du noyau, théorème de représentation

Du second point de la définition 2 résulte une propriété fondamentale des espaces de Hilbert à noyau reproduisant. En remplaçant ψ par κ_{x_j}, on aboutit en effet à

$$\kappa(x_i, x_j) = \langle \kappa_{x_j}, \kappa_{x_i} \rangle_{\mathcal{H}} \qquad (3)$$

quels que soient x_i et $x_j \in \mathcal{X}$. Ainsi le noyau κ fournit-il le produit scalaire des images dans $\mathcal{H}$ de toute paire d'éléments de $\mathcal{X}$, sans qu'il soit nécessaire d'expliciter ces images. Ce principe est appelé *coup du noyau*. Il permet d'élaborer des méthodes non-linéaires de traitement de données à partir d'approches linéaires, sous réserve que celles-ci puissent s'exprimer uniquement en fonction de produits scalaires des observations. Il suffit en effet de remplacer alors chacun de ces produits scalaires par un noyau non-linéaire. Ainsi la structure des algorithmes demeure-t-elle inchangée et le surcoût calculatoire dû à l'évaluation des noyaux négligeable.

Pour qu'il soit opérationnel, le coup du noyau nécessite souvent d'être associé au théorème de représentation [15]. Il en est par exemple ainsi pour les méthodes kernel-PCA, kernel-FDA et kernel-GDA. Ce théorème établit que toute fonction ψ^* d'un espace de Hilbert à noyau reproduisant $\mathcal{H}$ qui minimise un coût

$$J((x_1, y_1, \psi(x_1)), \ldots, (x_n, y_n, \psi(x_n))) + g(\|\psi\|^2_{\mathcal{H}}), \qquad (4)$$

impliquant n sorties $\psi(x_i)$ obtenues pour des entrées x_i et éventuellement n sorties désirées y_i, avec g une fonction monotone croissante sur $\mathbb{R}+$, peut s'écrire sous la forme

$$\psi^* = \sum_{i=1}^{n} a_i^*\,\kappa_{x_i}. \qquad (5)$$

On démontre ceci en notant que toute fonction ψ de $\mathcal{H}$ se décompose selon $\psi = \sum_{i=1}^{n} a_i\,\kappa_{x_i} + \psi^{\perp}$, avec $\langle \psi^{\perp}, \kappa_{x_i} \rangle_{\mathcal{H}} = 0$ pour tout $i = 1, \ldots, n$. Puisque $\psi(x_j) = \langle \psi, \kappa_{x_j} \rangle_{\mathcal{H}}$, la valeur de $\psi(x_j)$ n'est donc pas affectée par $\psi^{\perp}$, pour $j = 1, \ldots, n$.

3 Mises en œuvre temps-fréquence

L'objectif de cette section est de montrer que les méthodes à noyau peuvent être mise en œuvre dans le domaine temps-fréquence grâce à un choix approprié de noyau reproduisant. On s'intéresse d'abord à la distribution de Wigner, par soucis de clarté, avant d'étendre l'étude à d'autres distributions. On suppose disposer dans la suite d'un ensemble $\mathcal{A}_n$ comprenant n signaux $x_i \in \mathcal{X}$, éventuellement accompagnés d'une étiquette ou autre sortie désirée $y_i \in \mathcal{Y}$.

3.1 Distribution de Wigner

On désigne par W_x la distribution de Wigner de x, c'est-à-dire

$$W_x(t, f) = \int x(t + \tau/2)\,\overline{x}(t - \tau/2)\,e^{-2j\pi f \tau}\,d\tau. \qquad (6)$$

Les classiques méthodes paramétriques de reconnaissance des formes, appliquées ici dans le plan temps-fréquence, reposent sur l'estimation de Ψ^* de sorte que la statistique

$$\psi^*(x) = \langle \Psi^*, W_x \rangle = \iint \Psi^*(t, f)\,W_x(t, f)\,dt\,df \qquad (7)$$

optimise un critère J donné, par exemple celui de Fisher pour une analyse factorielle discriminante. En pratique,[1] la résolution d'un tel problème est rendue difficile par la taille des représentations W_{x_i} manipulées. Le coup du noyau et le théorème de représentation peuvent ici jouer un rôle déterminant. On considère pour cela le noyau suivant

$$\kappa_W(x_i, x_j) = \langle W_{x_i}, W_{x_j} \rangle. \qquad (8)$$

[1] Les méthodes de discrétisation de la distribution de Wigner étant multiples, on a pris le parti de raisonner sur la définition (6) dans cet article. En pratique, on rappelle toutefois que les représentations temps-fréquence discrètes de signaux composés de l échantillons sont des matrices de taille $(l \times l)$.

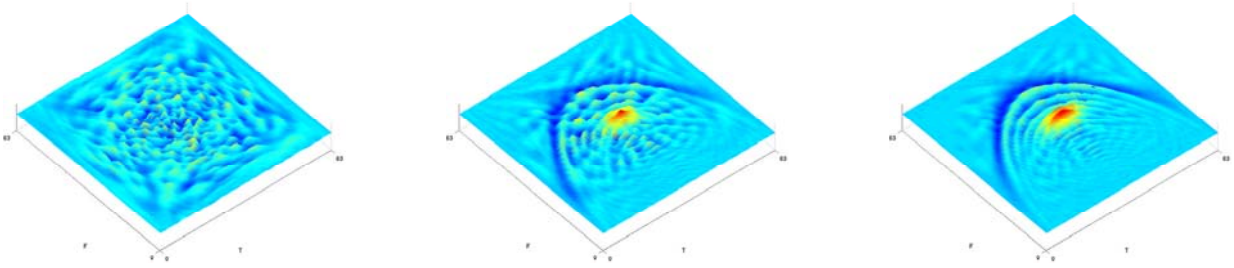


Figure 1: Problème de classification concernant deux familles de 200 signaux de longueur 64, gaussiens ($\mathcal{C}_1$) et à modulation de fréquence parabolique ($\mathcal{C}_2$), noyés dans un bruit blanc. Résultats issus d'une kernel-FDA, sans régularisation à gauche et avec régularisation au milieu, et de l'application d'un algorithme SVM à droite. Le noyau quadratique (9) et la relation de reconstruction (10) ont été utilisés.

Notons immédiatement qu'il ne nécessite aucun calcul de distributions de Wigner puisque, par la relation de Moyal, on a

$$\kappa_W(x_i, x_j) = |\langle x_i, x_j \rangle|^2. \tag{9}$$

On vérifie qu'il s'agit d'un noyau défini positif. La condition (1), qui peut s'exprimer sous la forme $\| \sum_i a_i \, W_{x_i} \|^2 \geq 0$, est en effet satisfaite. Il lui correspond donc un espace de Hilbert à noyau reproduisant $\mathcal{H}_W$ unique. On l'obtient en complétant l'espace fonctionnel $\mathcal{H}_0$ défini ci-dessous de sorte que toute suite de Cauchy y converge, ce que l'on note $\mathcal{H}_W = \overline{\mathcal{H}}_0$.

$$\mathcal{H}_0 = \{\psi : \mathcal{X} \to \mathbb{R} \mid \psi = \sum_i a_i \, |\langle \cdot \, , x_i \rangle|^2, a_i \in \mathbb{R}, x_i \in \mathcal{X}\}$$

Rien ne s'oppose désormais à la mise en œuvre du coup du noyau et à l'application du théorème de représentation, ou plus généralement à l'association du noyau (8)-(9) à tout algorithme à noyau décrit dans la littérature. Le résultat (5) admet alors une interprétation temps-fréquence implicite, que l'on exhibe en identifiant Ψ^* figurant dans l'expression (7). Ainsi, on a

$$\Psi^* = \sum_{i=1}^{n} a_i^* \, W_{x_i} \tag{10}$$

où les échantillons x_i désignent les éléments de l'ensemble d'apprentissage $\mathcal{A}_n$. On insiste sur le fait que l'estimation des paramètres a_i ne nécessite aucun calcul de distributions temps-fréquence puisqu'elle repose sur le noyau (9). En revanche, si on le souhaite, un calcul itératif permet de déterminer Ψ^* dans un second temps sans qu'il ait jamais été nécessaire de conserver en mémoire l'ensemble des distributions de Wigner des éléments de la base d'apprentissage.

3.2 Extension à d'autres distributions

3.2.1 Transformée de Fourier à court-terme

On s'intéresse tout d'abord à une distribution temps-fréquence linéaire : la transformée de Fourier à court-terme, notée F_x. Elle est unitaire à condition que la fenêtre d'analyse soit d'énergie unité, ce qui signifie

$$\langle F_{x_i}, F_{x_j} \rangle = \langle x_i, x_j \rangle. \tag{11}$$

Le noyau associé, $\kappa_F(x_i, x_j) = \langle x_i, x_j \rangle$, est évidemment défini positif puisque la condition $\| \sum_i a_i \, x_i \|^2 \geq 0$ est satisfaite. Toute méthode à noyau peut donc être mise en œuvre sur la transformée de Fourier à court-terme à partir de celui-ci. Dans un second temps, on peut accéder à l'information temps-fréquence extraite grâce à l'expression (5) dans laquelle W_{x_i} aura fait place à F_{x_i}.

3.2.2 Classe de Cohen

On se concentre à présent sur les distributions C_x de la classe de Cohen, caractérisées par une fonction de paramétrisation Φ que l'on a choisie de considérer dans le plan Doppler-retard

$$C_x(t, f) = \iint \Phi(\nu, \tau) \, A_x(\nu, \tau) \, e^{-2j\pi(f\tau + \nu t)} \, d\nu \, d\tau, \tag{12}$$

où A_x désigne la fonction d'ambiguïté à bande étroite du signal x. Notons en passant que la distribution A_x est unitaire. Tout comme la distribution de Wigner, lui est associé le noyau quadratique (9), et par conséquent le même espace de Hilbert à noyau reproduisant. En outre, leurs signatures Ψ^* respectives sont liées par une transformée de Fourier bidimensionnelle, les paramètres a_k^* obtenus au terme de la minimisation du critère (4) étant identiques dans les deux cas en vertu de l'égalité de Parseval-Plancherel appliquée à (7). Concernant les distributions de la classe de Cohen en général, le noyau reproduisant associé s'exprime ainsi

$$\kappa(x_i, x_j) = \iint |\Phi(\nu, \tau)|^2 \, A_{x_i}(\nu, \tau) \, \overline{A}_{x_j}(\nu, \tau) \, d\nu \, d\tau. \tag{13}$$

On retrouve à nouveau le noyau quadratique (9) si la condition d'unitarité $|\Phi(\nu, \tau)| = 1$ est satisfaite. Parmi les distributions concernées, on peut citer par exemple celles de Page-Levin et de Rihaczek. Dans les autres cas, les calculs ne se présentent pas aussi favorablement puisqu'il est nécessaire d'expliciter les distributions C_{x_i} pour évaluer leurs produits scalaires deux à deux préalablement à l'application d'une méthode à noyau.

4 Applications

La première application concerne un problème de classification de deux familles de 200 signaux de longueur 64, gaussiens ($\mathcal{C}_1$) et à modulation de fréquence parabolique ($\mathcal{C}_2$), noyés dans un bruit blanc gaussien. La figure 1 présente les signatures Ψ^* obtenues au terme de la mise en œuvre de deux algorithmes de classification non-linéaire, kernel-FDA et SVM, sur la base du noyau quadratique (9). Comme le montre la figure de gauche, l'application directe de la méthode kernel-FDA ne met pas en évidence la configuration temps-fréquence des signaux étudiés. Des problèmes d'instabilité numérique identifiés dans [9] en sont la cause, que l'on peut contenir à l'aide d'une technique de régularisation de type Tikhonov comme l'illustre la figure du centre. La figure de droite met en évidence la propriété régularisante des SVM, propriété intégrante du critère de marge

Méthode	Taux d'erreur (%)
kernel-FDA non-régularisée	50
kernel-FDA régularisée	7
SVM	1.5

Figure 2: Comparaison des taux d'erreur obtenus avec les méthodes SVM et kernel-FDA, non-régularisée et régularisée, pour le problème décrit en figure 1.

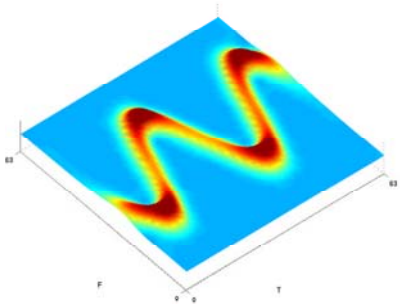


Figure 3: Module-carré de la 1ère composante principale issue d'un algorithme kernel-PCA appliqué à 300 signaux à modulation de fréquence sinusoïdale, noyés dans un bruit blanc.

sur lequel elles reposent et qui a fait leur succès. Le tableau en figure 2 fournit les taux d'erreur associés.

La seconde application est relative à la mise en œuvre de la méthode kernel-PCA, qui effectue une analyse en composantes principales non-linéaires d'un ensemble d'individus. Celle-ci a été appliquée à une famille de 300 signaux de 256 composantes, à modulation de fréquence sinusoïdale et noyés dans un bruit blanc. La figure 3 représente la première composante principale obtenue à partir du noyau (11) et de la transformée de Fourier à court-terme. Les caractéristiques temps-fréquence des signaux étudiés sont clairement identifiables. Notons qu'une analyse en composantes principales appliquée directement aux transformées de Fourier à court-terme des signaux aurait fourni le même résultat. Elle aurait cependant été très difficile à mettre en œuvre puisqu'elle aurait nécessité la manipulation de 300 individus constitués de 256^2 composantes, et notamment la diagonalisation de leur matrice de covariance de taille $256^2 \times 256^2$. L'approche proposée a permis de surmonter cet inconvénient puisqu'elle n'a nécessité que la diagonalisation d'une matrice de Gram, ici de taille 300×300.

5 Conclusion

La théorie des noyaux reproduisants a permis le développement fulgurant d'une classe d'algorithmes de reconnaissance des formes dont la formulation ne dépend pas de de la nature des données traitées, ni de l'espace de représentation adopté pour résoudre les problèmes. Au delà de ce caractère universel, celles que l'on range désormais sous le qualificatif de méthodes à noyau doivent également leur succès à l'essor de la théorie statistique de l'apprentissage, au sein de laquelle la prédiction de leurs performances en généralisation fait aujourd'hui encore l'objet d'études approfondies. L'analyse temps-fréquence traitant d'espaces de représentation particuliers, nous avons montré dans cet article qu'elle peut bénéficier des plus récents développements des méthodes à noyau. Il s'agit à présent de poursuivre cet effort et de profiter de l'éclairage nouveau apporté sur elles par la théorie de l'apprentissage et celle des noyaux reproduisants.

Références

[1] F. Auger et P. Flandrin, "Improving the readability of time-frequency and time-scale representations by reassignment methods," *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1068–1089, 1995.

[2] D. Jones et R. Baraniuk, "An adaptive optimal-kernel time-frequency representation," *IEEE Transactions on Signal Processing*, vol. 43, no. 10, pp. 2361–2371, 1995.

[3] J. Gosme, C. Richard, et P. Gonçalvès, "Adaptive diffusion of time-frequency and time-scale representations: a review." *IEEE Transactions on Signal Processing (à paraître)*, 2005.

[4] C. Doncarli et N. Martin, *Décision dans le plan temps-fréquence*. Paris: Hermès Sciences, Traité IC2, 2004.

[5] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, pp. 337–404, 1950.

[6] B. Boser, I. Guyon, et V. Vapnik, "An training algorithm for optimal margin classifiers," in *Proc. Fifth Annual Workshop on Computational Learning Theory*, 1992, pp. 144–152.

[7] J. Vert, K. Tsuda, et B. Schölkopf, "A primer on kernel methods," in *Kernel Methods in Computational Biology*, B. Schölkopf, K. Tsuda, and J. Vert, Eds. Cambridge, MA: MIT Press, 2004, pp. 35–70.

[8] B. Schölkopf, A. J. Smola, et K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299–1319, 1998.

[9] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, et K. R. Müller, "Fisher discriminant analysis with kernels," in *Advances in neural networks for signal processing*, Y. H. Hu, J. Larsen, E. Wilson, and S. Douglas, Eds. San Mateo, CA: Morgan Kaufmann, 1999, pp. 41–48.

[10] G. Baudat et F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, no. 10, pp. 2385–2404, 2000.

[11] F. Cucker et S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, no. 1, pp. 1–49, 2002.

[12] V. N. Vapnik, *The nature of statistical learning theory*. New York, NY: Springer, 1995.

[13] M. Davy, A. Gretton, A. Doucet, et P. Rayner, "Optimised support vector machines for nonstationary signal classification," *IEEE Signal Processing Letters*, vol. 9, no. 12, pp. 442–445, 2002.

[14] A. Gretton, M. Davy, A. Doucet, et P. Rayner, "Nonstationary signal classification using support vector machine," in *Proc. IEEE Workshop on Statistical Signal Processing*, 2001.

[15] B. Schölkopf, R. Herbrich, et R. Williamson, "A generalized representer theorem," NeuroCOLT, Royal Holloway College, University of London, UK, Tech. Rep. NC2-TR-2000-81, 2000.

[16] R. Herbrich, *Learning kernel classifiers. Theory and algorithms*. Cambridge, MA: The MIT Press, 2002.

Une Méthode à Noyaux pour l'Approximation Parcimonieuse des Représentations Temps-Fréquence Bilinéaires

Manuel DAVY[1], Patrick WOLFE[2]

[1]Laboratoire d'Automatique, de Génie Informatique et Signal
BP 48, Cité Scientifique, 59651 Villeneuve d'Ascq cedex, France

[2]Harvard University, Cambridge, MA, USA

Manuel.Davy@ec-lille.fr,p.wolfe@ieee.org

Résumé – Nous présentons une méthode d'approximation des représentations temps-fréquence (RTF) reprenant le formalisme des méthodes à noyaux. Certains noyaux temps-fréquence étant définis positifs, la RTF est une fonction de l'espace de Hilbert à noyau reproduisant correspondant, qui est un espace de recherche naturel pour l'approximation. Cette dernière est recherchée avec une complexité minimale, et elle est parcimonieuse. Nous étendons le formalisme de la regression à vecteurs support pour résoudre le problème.

Abstract – We introduce a Time-Frequency Representation (TFR) approximation method, within the kernel machines framework. Some time-frequency kernels being positive definite, the TFR is a member of the related reproducing kernel Hilbert space which is the natural approximating function space. The approximation function is sought with minimum complexity and it is sparse. We solve the problem by an extention to the support vector regression formalism.

1 Introduction

Les représentations temps-fréquence (RTF) sont largement utilisées dans de multiples champs du traitement du signal, comme l'analyse des signaux audio, mécaniques, astrophysiques... Elle sont cependant largement redondantes, ce qui en fait des objets de grande dimension, et dont la manipulation intensive peut être coûteuse. Dans cet article, nous proposons une méthode d'approximation parcimonieuse des RTFs. L'objectif est de trouver une représentation aussi proche que possible de la représentation initiale, mais pouvant être résumée à peu de paramètres non nuls.

Depuis les travaux de Cohen [1], nous savons que toute représentation bilinéaire et covariante $\mathsf{C}_z^{\phi}(t,f)$ d'un signal peut s'écrire sous la forme d'une double convolution temps-fréquence de la représentation de Wigner-Ville, donnée par

$$\mathsf{WV}_z(t,f) \;=\; \int_{-\infty}^{\infty} x(t+\frac{\tau}{2})\, x^{\star}(t-\frac{\tau}{2})\; e^{-j2\pi f\tau} \qquad (1)$$

par un noyau $\phi(t,f)$, comme suit :

$$\mathsf{C}_z^{\phi}(t,f) \;=\; \iint \mathsf{WV}_z(\tau,\nu)\phi(t-\tau,f-\nu)\,d\tau\,d\nu \qquad (2)$$

En pratique, nous utilisons des version discrétisées de ces expressions, selon une grille de N_t points temporels t_i, $i = 1,\ldots,N_t$ et N_f points fréquentiels f_i, $i = 1,\ldots,N_f$ (c'est-à-dire, $\mathsf{C}_z^{\phi}(t,f)$ est représenté sous la forme d'une matrice $N_f \times N_t$).

L'idée de cet article est de construire une approximation de $\mathsf{C}_z^{\phi}(t,f)$, recherchée dans la classe de fonctions de la forme

$$\mathsf{f}(t,f) \;=\; \sum_{i=1}^{N_t}\sum_{j=1}^{N_f} \alpha_{i,j} K\big((t,f),(t_i,f_j)\big) + b \qquad (3)$$

où $K(\cdot,\cdot)$ est un noyau défini positif défini sur $\mathbb{R}^2 \times \mathbb{R}^2$, et où chaque point (t_i,f_j) de la grille temps-fréquence porte un coefficient $\alpha_{i,j}$, $i = 1,\ldots,N_t$, $j = 1,\ldots,N_f$. En outre, la solution recherchée doit être telle que les $\alpha_{i,j}$ soient nuls pour la plupart – fournissant ainsi une approximation parcimonieuse – et positifs – conservant ainsi une interprétation en termes de représentation d'énergie.

Cet article est organisé de la façon suivante. La section 2 développe une méthode originale de régression à vecteurs support dans laquelle la contrainte de positivité des coefficients est imposée. La section 3 propose quelques commentaires et évoque le réglage des paramètres. La section 4 présente des résultats, et la dernière section propose quelques conclusions et des pistes de recherche future.

2 Formulation proposée

Dans la suite, nous noterons par x_i, $i = 1,\ldots,m$ (avec $m = N_f N_t$) les points en dimension 2 de la grille temps-fréquence définie par t_k, $k = 1,\ldots,N_t$ et $f_{k'}$, $k = 1,\ldots,N_f$. Ainsi, à chaque couple (k,k'), $k = 1,\ldots,N_t$, $k' = 1,\ldots,N_f$, correspond-il un indice i, tels que $x_i = [t_k, f_{k'}]^{\mathsf{T}}$. L'équation (3) peut être réécrite sous la forme

$$\mathsf{f}(x) \;=\; \sum_{i=1}^{m} \alpha_i K(x,x_i) + b \qquad (4)$$

où $x = [t,f]^{\mathsf{T}}$. Nous supposons pour l'instant que les α_i et b sont dans $\mathbb{R}$ et que $K(\cdot,\cdot)$ est donné. Enfin, nous notons dans la suite y la valeur de la RTF en x, c'est-à-dire $y \;=\; \mathsf{C}_z^{\phi}\big((t,f) = x\big)$ pour z et $\phi(\cdot,\cdot)$ donnés.

Approcher $y = \mathsf{C}_z^{\phi}(x)$ par une fonction $\mathsf{f}(x)$ telle que celle de l'équation (4), avec une contrainte de parcimonie revient à résoudre un problème variationnel. Avant de le développer, nous introduisons une fonction de coût et les risques correspondants.

2.1 Fonction de coût et risque

Reprenant le formalisme introduit dans le cadre de la régression par les méthodes à vecteurs support, nous utilisons dans la suite la fonction de coût $c(\cdot, \mathsf{f}(\cdot))$ de la figure 1, dite ϵ-insensible [2]. Le coût $c(\cdot, \mathsf{f}(\cdot))$ indique l'importance que l'on accorde à un écart entre la fonction $\mathsf{f}(x)$ et la valeur y de la RTF.

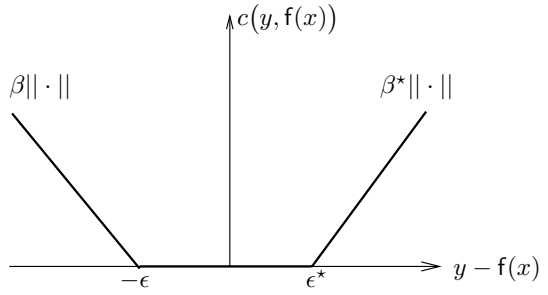


FIG. 1: *Une fonction de coût de type ϵ-insensible, non symétrique. Elle ne pénalise pas les erreurs se trouvant dans l'intervalle $[-\epsilon, \epsilon^\star]$, et pénalise linéairement les erreurs plus importantes. Cela assure la parcimonie de la l'approximation trouvée.*

Le risque empirique associé s'écrit comme la moyenne empirique de la fonction de coût

$$R_{\text{emp}}[\mathsf{f}] \;=\; \frac{1}{m}\sum_{i=1}^{m} c\big(y_i, \mathsf{f}(x_i <)\big) \tag{5}$$

2.2 Problème variationnel

La fonction $\mathsf{f}(\cdot)$ de l'équation (4) appartient à un espace de fonctions de type *espace de Hilbert à noyau reproduisant* $K(\cdot,\cdot)$ (car ce dernier est défini positif) $\mathcal{H}$, dans lequel il existe une norme induite notée $||\cdot||_{\mathcal{H}}$. A l'aide de cette norme, nous définissons le problème variationnel suivant, à résoudre numériquement :

$$\begin{array}{rl} \text{Minimiser} & \dfrac{1}{2}||\mathsf{f}||^2_{\mathcal{H}} + C R_{\text{emp}}[\mathsf{f}] \\ \text{Avec} & \alpha_i \geq 0 \;, \forall i = 1, \ldots, m \end{array}$$

La solution de ce problème assure à la fois une bonne approximation de la RTF par la fonction f (faible risque empirique), ainsi qu'une solution de faible complexité (norme faible de f dans $\mathcal{H}$), évitant ainsi le sur-apprentissage, et favorisant la parcimonie. Le problème peut se réécrire comme suit (voir [2] pour une dérivation similaire) : g

$$\begin{array}{rl} \text{Minimiser} & \dfrac{1}{2}\displaystyle\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) + \frac{C}{m}\sum_{i=1}^{m}(\beta\xi_i + \beta^\star \xi_i^\star) \\ & \text{par rapport à } \; \{\alpha_i, \xi_i, \xi_i^\star\}_{i=1,\ldots,m} \in \mathbb{R}^{3m} \\ \text{Avec} & \big(< \mathbf{w}, \mathbf{x}_i > + b\big) - y_i \leq \epsilon + \xi_i \;, \\ & y_i - \big(< \mathbf{w}, \mathbf{x}_i > + b\big) \leq \epsilon^\star + \xi_i^\star \;, \\ & \xi_i \geq 0 \;,\; \xi_i^\star \geq 0 \;,\; \alpha_i \geq 0 \text{ pour } i = 1, \ldots, m \end{array}$$

Ce problème quadratique et convexe admet une formulation duale qui fournit les coefficients α_i. Pour cela, nous introduisons les multiplicateurs de Lagrange η_i, $\eta_i^\star$, γ_i, $\gamma_i^\star$ et λ_i, positifs, et nous formons le Lagrangien L qui doit être minimisé par rapport aux variables primales b, $\{\alpha_i, \xi_i, \xi_i^\star\}_{i=1,\ldots,m}$ et maximisé par rapport aux variables duales η_i, $\eta_i^\star$, γ_i, $\gamma_i^\star$ et λ_i

$$\begin{array}{rl} L = & \dfrac{1}{2}\displaystyle\sum_{i,j=1}^{m} \alpha_i \alpha_j K(x_i, x_j) + \frac{C}{m}\sum_{i=1}^{m}(\beta\xi_i + \beta^\star \xi_i^\star) \\ & -\displaystyle\sum_{i=1}^{m}(\eta_i \xi_i + \eta_i^\star \xi_i^\star + \lambda_i \alpha_i) \\ & -\displaystyle\sum_{i=1}^{m}\gamma_i\big(\epsilon + \xi_i + y_i - \sum_{j=1}^{m}\alpha_j K(x_i, x_j) - b\big) \\ & -\displaystyle\sum_{i=1}^{m}\gamma_i^\star\big(\epsilon^\star + \xi_i^\star + \sum_{j=1}^{m}\alpha_j K(x_i, x_j) + b - y_i\big) \end{array} \tag{6}$$

Sous forme vectorielle, il s'écrit (où les notations en gras représentent les vecteurs et matrices, par exemple, $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_m]^\mathsf{T}$ et $\mathbf{K} = \big[K(x_i, x_j)\big]_{(i,j=1,\ldots,N)}$

$$\begin{array}{rl} L \;=\; & \frac{1}{2}\boldsymbol{\alpha}^\mathsf{T}\mathbf{K}\boldsymbol{\alpha} + \frac{C}{m}\big(\beta \cdot \mathbf{1}_m^\mathsf{T}\boldsymbol{\xi} + \beta^\star \cdot \mathbf{1}_m^\mathsf{T}\boldsymbol{\xi}^\star\big) \\ & -\boldsymbol{\eta}^\mathsf{T}\boldsymbol{\xi} - \boldsymbol{\eta}^{\star\mathsf{T}}\boldsymbol{\xi}^\star - \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{\lambda} \\ & -\boldsymbol{\gamma}^\mathsf{T}\big((\epsilon - b)\mathbf{1}_m + \boldsymbol{\xi} + \mathbf{y} - \mathbf{K}\boldsymbol{\alpha}\big) \\ & -\boldsymbol{\gamma}^{\star\mathsf{T}}\big((\epsilon^\star + b)\mathbf{1}_m + \boldsymbol{\xi}^\star + \mathbf{K}\boldsymbol{\alpha} - \mathbf{y}\big) \end{array} \tag{7}$$

A l'optimum, dit *point selle*, les dérivées par rapport aux variables primales s'annulent,

$$\frac{\partial L}{\partial b} \;=\; \big(\boldsymbol{\gamma} - \boldsymbol{\gamma}^\star\big)^\mathsf{T}\,\mathbf{1}_m \;=\; 0 \tag{8}$$

$$\frac{\partial L}{\partial \boldsymbol{\alpha}} \;=\; \mathbf{K}\big(\boldsymbol{\alpha} + \boldsymbol{\gamma} - \boldsymbol{\gamma}^\star\big) - \boldsymbol{\lambda} \;=\; 0 \tag{9}$$

$$\frac{\partial L}{\partial \boldsymbol{\xi}} \;=\; \frac{C}{m}\beta \cdot \mathbf{1}_m - \boldsymbol{\eta} - \boldsymbol{\gamma} \;=\; 0 \tag{10}$$

$$\frac{\partial L}{\partial \boldsymbol{\xi}^\star} \;=\; \frac{C}{m}\beta^\star \cdot \mathbf{1}_m - \boldsymbol{\eta}^\star - \boldsymbol{\gamma}^\star \;=\; 0 \tag{11}$$

d'où l'on extrait ($\mathbf{K}$ est inversible)

$$\boldsymbol{\alpha} \;=\; \mathbf{K}^{-1}\boldsymbol{\lambda} + \boldsymbol{\gamma}^\star - \boldsymbol{\gamma} \tag{12}$$

Afin d'écrire ce problème sous une forme classique, nous introduisons $\boldsymbol{\omega}^\mathsf{T} \;=\; [\boldsymbol{\lambda}^\mathsf{T}; \boldsymbol{\gamma}^\mathsf{T}; \boldsymbol{\gamma}^{\star\mathsf{T}}]$ et

$$\mathbf{H} \;=\; \begin{bmatrix} \mathbf{K}^{-1} & -\mathbf{I}_m & \mathbf{I}_m \\ -\mathbf{I}_m & \mathbf{K} & -\mathbf{K} \\ \mathbf{I}_m & -\mathbf{K} & \mathbf{K} \end{bmatrix} \tag{13}$$

$$\mathbf{C} \;=\; \Big[\mathbf{0}_m^\mathsf{T} \;;\; \mathbf{y} + \epsilon\,\mathbf{1}_m^\mathsf{T} \;;\; -\mathbf{y} + \epsilon^\star\,\mathbf{1}_m^\mathsf{T}\Big]^\mathsf{T} \tag{14}$$

$$\mathbf{A} \;=\; \Big[\mathbf{0}_m^\mathsf{T} \;;\; \mathbf{1}_m^\mathsf{T} \;;\; -\mathbf{1}_m^\mathsf{T}\Big]^\mathsf{T} \tag{15}$$

Enfin, soient $\mathbf{l} = \mathbf{0}_{3m}$, $\mathbf{u} \;=\; \big[+\infty\mathbf{1}_m^\mathsf{T} \;;\; \frac{C}{m}\beta\mathbf{1}_m^\mathsf{T} \;;\; \frac{C}{m}\beta^\star\mathbf{1}_m^\mathsf{T}\big]^\mathsf{T}$ et $b = 0$. Le problème d'optimisation à résoudre est quadratique sous contraintes linéaires, à résoudre par rapport à $\boldsymbol{\omega}$

$$\begin{array}{ll} \text{minimiser} & \frac{1}{2}\boldsymbol{\omega}^\mathsf{T}\mathbf{H}\boldsymbol{\omega} + \mathbf{C}^\mathsf{T}\boldsymbol{\omega} \\ \text{avec} & \mathbf{A}^\mathsf{T}\boldsymbol{\omega} \;=\; b \\ & \mathbf{l} \leq \boldsymbol{\omega} \leq \mathbf{u} \end{array} \tag{16}$$

Ce problème d'optimisation ne présente pas de difficultés numériques, et peut être traité par un algorithme de point intérieur [3] ou encore itérativement, en prenant en compte les points temps-fréquence l'un après l'autre.

3 Discussion

Le principe d'approximation proposé ici repose sur le choix d'un noyau défini positif. Ce choix peut être effectué de manière arbitraire, par exemple un noyau gaussien

$$K\big((t_1,f_1),(t_2,f_2)\big) = e^{-0.5\left[\frac{(t_1-t_2)^2}{\sigma_t^2}+\frac{(f_1-f_2)^2}{\sigma_f^2}\right]} \qquad (17)$$

Dans certains cas, le noyau temps-fréquence ϕ lui-même est défini positif (par exemple, le noyau du spectrogramme à fenêtre gaussienne correspond au noyau gaussien ci-dessus), et il est naturel de choisir

$$K\big((t_1,f_1),(t_2,f_2)\big) \;=\; \phi(t_1-t_2,f_1-f_2) \qquad (18)$$

Ainsi, la fonction f obtenue est proche de la représentation initiale, tout en comportant beaucoup de coefficients nuls. En d'autres termes, la fonction f appartient au même espace de Hilbert à noyau reproduisant que $\mathsf{C}_z^{\phi}(t,f)$. Cela permet à la fois un débruitage, et une compression de la RTF, permettant par exemple la mise en œuvre d'algorithmes de classification ou de détection. Le coefficient de régularisation permet de fixer la proportion de coefficients nuls: pour C petit, beaucoup de coefficients α_i sont nuls, mais l'approximation effectuée est moins précise que si C est grand, et que beaucoup de coefficients sont non nuls. De même, les paramètres ϵ et $\epsilon^\star$ permettent de régler le niveau de précision souhaité pour l'approximation en fixant un «horizon »à partir duquel on pénalise l'écart entre la RTF originale et son approximation.

L'ensemble des points temps-fréquence pour lesquels $-\epsilon^\star \leq \mathsf{f}(t,f)-\mathsf{C}_z^{\phi}(t,f) \leq \epsilon$ ont un coefficient α nul et ne contribuent pas à former l'approximation.

La procédure de calcul des α_i est la suivante. Dans un premier temps, la RTF de noyau ϕ est calculée, puis la matrice $\mathbf{K}$ est calculée. Enfin, le problème d'optimisation 16 est résolu numériquement. Les α_i significativement non nuls sont conservés et servent de base à l'approximation. Enfin, le paramètre b est calculé grâce aux conditions de Karush-Kuhn-Tucker, voir [2].

L'approche proposée ici montre de nombreux avantages. Tout d'abord, elle est d'une grande simplicité algorithmique. Ensuite, l'influence des paramètres de réglage est facile à interpréter, et ils sont faciles à régler. Enfin,le coût calculatoire peut être très réduit lorsqu'on utilise une procédure d'optimisation adaptée. Il faut noter toutefois que cette méthode ne prend pas en compte d'*a priori* structurels pouvant être définis dans certains problèmes, contrairement à l'approche bayésienne développée dans [4], où un champ de Markov modélise les relations entre coefficients voisins.

4 Résultats

Les résultats présentés concernent des signaux synthétiques comportant deux chirps se croisant. Le noyau $\phi(t,f)$ est gaussien avec $\sigma_t^2 = \sigma_f^2 = 2$, et le signal traité comporte $N = 40$ points (cette longueur est relativement courte car nous avons mis en oeuvre un algorithme de résolution du problème (16) non itératif, simple mais coûteux; un algorithme itératif permettrait de traiter des signaux beaucoup plus long en un temps comparable). Nous avons réglé $C\beta = C\beta^\star = 5000$, $\epsilon = \epsilon^\star = 0.2$. La figure 2 représente la RTF originale, et la figure 3 présente son approximation. On constate une excellente

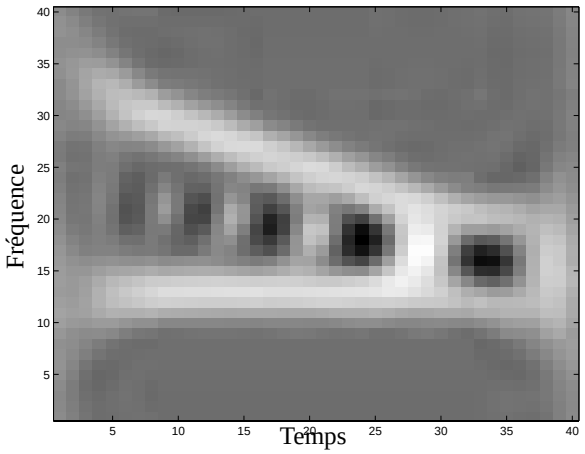


FIG. 2: *Représentation temps-fréquence de deux chirps calculée pour un noyau gaussien.*

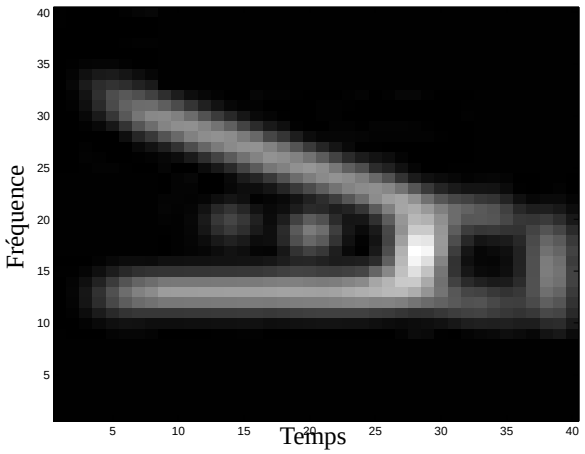


FIG. 3: *Représentation temps-fréquence approchée de la RTF de la figure 2 obtenue par la méthode proposée.*

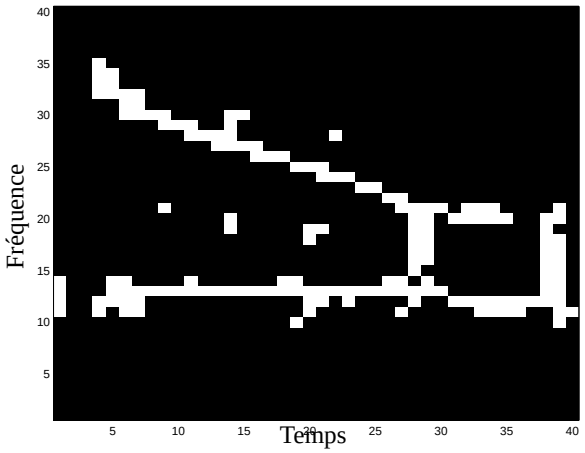


FIG. 4: *Positions temps-fréquence des coefficients α_i significativement non nuls et utilisés pour calculer la RTF approchée de la figure 3 obtenue par la méthode proposée.*

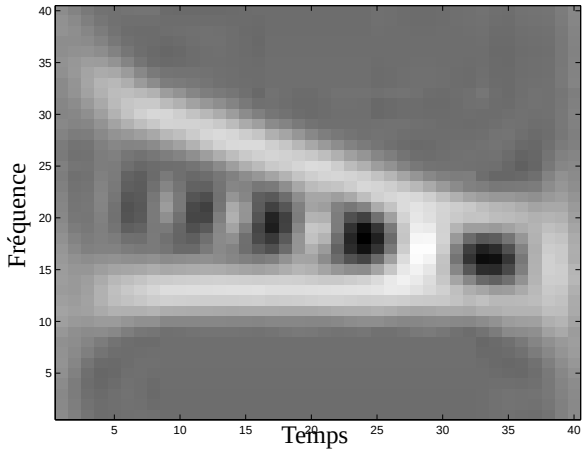


FIG. 5: *Représentation temps-fréquence de deux chirps calculée pour un noyau gaussien.*

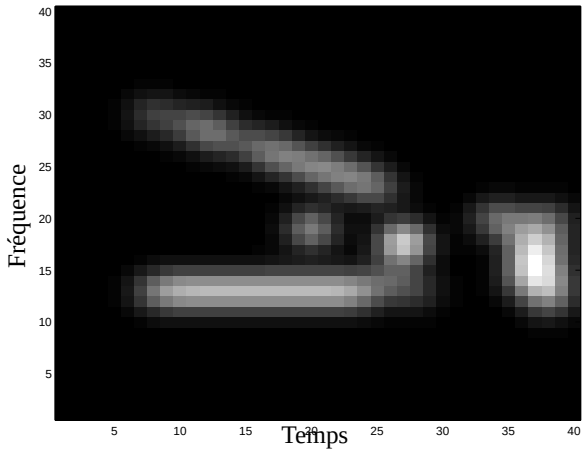


FIG. 6: *Représentation temps-fréquence approchée de la RTF de la figure 2 obtenue par la méthode proposée. La régularisation est plus forte qu'à la figure 3.*

adéquation. Enfin, la figure 4 montre les points temps-fréquence où les coefficients α_i sont non nuls. Ils représentent 9.75% du nombre total des points temps-fréquence.

Les figures 5, 6 et 7 ci-dessous montrent le résultat obtenu lorsque $C\beta = C\beta^\star = 500$, c'est-à-dire avec une régularisation plus importante, les autres paramètres étant inchangée. Les coefficients α_i non nuls ne représentent plus que 2.62% du nombre total de coefficients temps-fréquence.

5 Conclusions

La méthode présentée permet de réduire significativement l'information temps-fréquence par un problème d'optimisation bien posé et sans difficultés numériques. Elle permet une compression de l'information pouvant atteindre plus de 97% sans trop dénaturer l'information temps-fréquence. Par ailleurs, les coefficients α_i étant positifs, ils peuvent être interprétés comme une RTF épurée de l'énergie du signal qui pourra être employée dans des problèmes de décision.

Un algorithme itératif spécifique va maintenant être développé

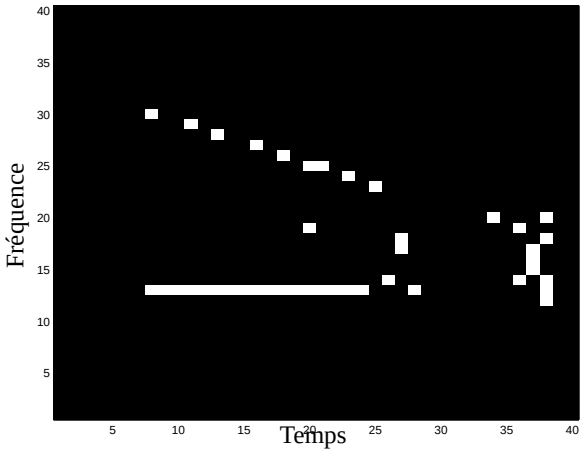


FIG. 7: *Positions temps-fréquence des coefficients α_i significativement non nuls et utilisés pour calculer la RTF approchée de la figure 6 obtenue par la méthode proposée. La régularisation est plus forte qu'à la figure 4*

et la méthode pourra être appliquée à la classification temps-fréquence de signaux et à la compression audio.

Références

[1] P. Flandrin, *Temps-Fréquence*, Série Traitement du signal. Hermès, 1993.

[2] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT press, Cambridge, MA, 2002.

[3] R. J. Vanderbei, "Loqo: An interior point code for quadratic programming", Tech. Rep. TR SOR-94-15, Department of Civil Engineering and Operations Research, Princeton University, 1995.

[4] P. J. Wolfe, S. J. Godsill, and W.J. Ng, "Bayesian variable selection and regularisation for time-frequency surface estimation", *Journal of the Royal Statistical Society: Series B*, vol. 66, no. 3, pp. 575–589, Aug. 2004.

Estimation et Reconstruction des Signaux Courts Multicomposantes Modulées Non-linéairement à la fois en Amplitude et en Fréquence

Meryem JABLOUN[1], Michelle VIEIRA[1], Nadine MARTIN[1], François LÉONARD[2]

[1]Laboratoire des Images et des Signaux
961 rue de la Houille Blanche, BP 46, 38 402 St Martin d'Hères, France

[2]Institut Hydro-Québec
Varennes, Canada

meryem.jabloun@lis.inpg.fr, michelle.vieira@lis.inpg.fr,
nadine.martin@lis.inpg.fr,leonard.francois@ireq.ca

Résumé – Dans ce papier, nous considérons des signaux, courts, non-stationnaires, modulés non linéairement en amplitude et en fréquence. Nous étendons une approche locale, développée dans le cadre des signaux monocomposante et dont l'efficacité et la robustesse ont été prouvées [3, 4]. Nous utilisons un modèle polynomial pour la fréquence et l'amplitude instantanées (FI / AI). Les paramètres du modèle sont ensuite estimés en maximisant la Vraisemblance par une technique d'optimisation stochastique : le Recuit Simulé. Basés sur la même démarche et dans le contexte des signaux multicomposantes, nous comparons deux approches différentes. Une première approche, optimale, et qui s'avère coûteuse en temps de calcul, consiste à estimer tous les paramètres du modèle à la fois. La deuxième approche, sous-optimale, reconstruit itérativement le signal composante par composante. Des simulations de Monte Carlo et une comparaison avec les Bornes de Cramer Rao illustrant les bonnes performances seront présentées. Nous obtenons une bonne estimation dans le cas où les fréquences instantanées se croisent, ce qui constitue une bonne performance compte tenu du faible nombre d'échantillons. Les deux approches sont testées ensuite sur des données réelles.

Abstract – In this paper, we consider nonstationary signals with nonlinear amplitude and frequency modulation on short time-windows. Motivated by published works [3, 4], where we obtain high performances for monocomponent signals, we propose to extend the approach to multicomponent signals. Both the instantaneous amplitude and frequency were modeled by polynomial functions. The maximization of the likelihood function was achieved using a stochastic optimization technique: the Simulated Annealing. We compare two different strategies. In the first one, we simultaneously estimate all model parameters. This is a highly computational strategy. The second one consists in iteratively recontructing the signal component by component. At each iteration, the parameters of only one component are estimated. Monte Carlo simulations are presented and compared to the appropriate Cramer-Rao Bounds. It proves the efficiency of the algorithm. Moreover it underscores the performance to estimate crossing frequency trajectories which is a great challenge related to the low sample number.

1 Introduction

L'analyse paramétrique des signaux multicomposantes, modulés en fréquence et en amplitude, a suscité un grand intérêt depuis quelques années [1, 2, 6]. Parmi les techniques proposées, nous citons la HAF (Higher Ambiguity Function) [1, 2] et la P-HAF (Product HAF) [6]. Malgré leur simplicité, ces méthodes sous optimales présentent plusieurs inconvénients. En effet, la présence de composantes multiples entraîne l'apparition d'interférences, qui dégradent l'estimation. Elles s'avèrent aussi très sensibles au bruit et aux modulations fortement non linéaires.

Afin de surmonter ces inconvénients, nous avons développé lors de précédents travaux [3, 4] une approche locale. Il s'agissait de considérer le signal sur une suite de courts segments temporels. Sur chacun d'eux, l'Amplitude et la Fréquence Instantanées (AI/FI) sont modélisées par des polynômes. Ensuite, en vue de préserver l'optimalité, la vraisemblance relative à ce modèle est maximisée à l'aide des techniques d'optimisation stochastique tel que le Recuit Simulé (RS). Ces segments, estimés un par un, sont fusionnés afin de reconstruire le signal dans sa totalité. Cette approche a permis une meilleure adaptabilité du modèle à tout type de modulation. L'efficacité de cette méthode a été prouvée dans le cadre d'un signal moncomposante. La précision de l'estimation, ainsi que la robustesse au bruit, étaient meilleures par rapport à celles obtenues par la HAF.

Dans ce papier, nous poursuivons cette approche dans le cadre des signaux multicomposantes et nous ne présentons qu'une étude locale. Nous proposons et comparons deux nouvelles méthodes, s'appuyant chacune sur les résultats obtenus dans [3, 4]. La première méthode estime toutes les composantes simultanément. La seconde est itérative : le signal est estimé composante par composante. Nous présentons des simulations de Monte Carlo et les comparons aux Bornes de Cramer Rao appropriées, pour illustrer l'efficacité de chacune de ces méthodes. Nous soulignerons, également, leur aptitude à gérer certains problèmes délicats tels que la présence de trajectoires de fréquence proches ou qui se croisent.
Nous considérons dans la suite le signal multicomposantes $s[n]$, noyé dans du bruit blanc gaussien additif $e[n]$, centré et de variance σ^2, défini comme suit :

$$\begin{array}{c} s[n] = \sum_{i=1}^{K} A_i[n] \exp\left(j\, \Phi_i[n]\right) \\ y[n] = s[n] + e[n], \text{ pour } \frac{-N}{2} \leq n \leq \frac{N}{2} \end{array} \quad (1)$$

$A_i[n]$ et $\Phi_i[n]$ désignent respectivement l'amplitude et la phase de la i-ème composante. $F_i[n] = \frac{1}{2\pi}\frac{\partial \Phi_i[n]}{\partial n}$ représente sa fréquence instantanée. Les amplitudes sont supposées strictement positives et les phases ne présentent pas de discontinuité. $y[n]$ est le signal bruité. $N+1$ désigne le nombre d'échantillons, supposé impair pour simplifier le calcul.

2 Modélisation

Dans [3, 4], grâce à la courte durée du segment considéré (une trentaine de points), nous supposons une base de polynômes de degré deux, définie sur $[\frac{-N}{2}, \frac{N}{2}]$, suffisante pour modéliser la non linéarité de la modulation. Sept paramètres sont à estimer pour chaque composante : la phase initiale $\theta_{i,0}$, trois paramètres pour l'amplitude $\{a_{i,0}, a_{i,1}, a_{i,2}\}$ et trois pour la fréquence $\{f_{i,0}, f_{i,1}, f_{i,2}\}$. Soit $g_0[n], g_1[n]$ et $g_2[n]$ la base considérée sur $[\frac{-N}{2}, \frac{N}{2}]$ ayant respectivement les degrés $0, 1$ et 2. Les décompositions de la FI et de la AI sont données par

$$\begin{array}{c} A_i[n] = \sum_{k=0}^{2} a_{i,k}\, g_k[n], \\ F_i[n] = \sum_{k=0}^{2} f_{i,k}\, g_k[n], \\ \Phi_i[n] = \theta_{i,0} + 2\pi(\sum_{k=\frac{-N}{2}}^{n} F_i[k] - \sum_{k=\frac{-N}{2}}^{0} F_i[k]) \end{array} \quad (2)$$

Le problème se résout à estimer $\theta = \{\theta_1, ..., \theta_K\}$ où $\theta_i = \{a_{i,0}, a_{i,1}, a_{i,2}, \theta_{i,0}, f_{i,0}, f_{i,1}, f_{i,2}\}$. Dans [3], nous avons proposé une base orthonormale discrète afin de découpler les paramètres de l'amplitude d'un signal monocomposante. Dans la suite, nous considérons la même base.

Pour estimer θ, nous allons maximiser la vraisemblance. Etant donnée l'hypothèse du bruit gaussien, cela revient à minimiser l'erreur des moindres carrés donnée ci-dessous :

$$\widehat{\theta} = \underbrace{argmin}_{\theta} \sum_{n=\frac{-N}{2}}^{\frac{N}{2}} |y[n] - \tilde{s}[n]|^2 \quad (3)$$

$\tilde{s}[n]$ est obtenu en remplaçant (2) dans (1) pour chaque θ. Vu la nonlinéarité de l'équation (3) ($7K$ paramètres inconnus), les techniques d'optimisation classiques, tel que la descente du gradient, convergent généralement vers des minima locaux, d'où l'usage de techniques d'optimisation stochastique.

3 Algorithmes

Dans [3, 4], nous avons comparé des techniques d'optimisation stochastique pour résoudre (3). Le RS est le meilleur compromis entre un faible biais et une petite variance d'estimation. Dans la suite, nous proposons d'étendre son utilisation à l'estimation des signaux multicomposantes. Nous présentons deux approches différentes et qui seront comparées.

3.1 Algorithme optimal

Il s'agit d'estimer les $7K$ paramètres simultanément afin de conserver le caractère optimal de l'estimateur de Maximum de Vraisemblance. L'algorithme proposé est une version modifiée du RS, décrit dans [4]. On initialise θ via le spectrogramme. Ensuite, on répète I fois les étapes (a), (b) et (c) décrites ci-dessous. I est un nombre déterminé expérimentalement pour accélérer la convergence. Il peut prendre des valeurs entre 1500 et 3000.

a- On génère un nombre de candidats θ_C selon une loi gaussienne de moyenne θ et de variance δ. δ peut être vue comme une agitation évitant de stagner sur un minimum local.

b- Si θ_C minimise (3), alors $\theta = \theta_C$, sinon θ reste inchangé.

c- Ensuite, on génère u selon une loi binomiale avec une probabilité $\rho \in [0, 0.5]$. Si $u = 1$, alors on diminue $\delta = 0.97\text{x}\delta$. Ainsi, on augmente la vitesse de convergeance.

d- Une fois les I itérations achevées, un test de blancheur sur le signal résiduel $\tilde{e}[n] = y[n] - \tilde{s}[n]$ permet d'arrêter l'algorithme. Dans le cas contraire, on relance l'estimation à partir de l'étape (a).

Considérons $r = \sum_{n=\frac{-N}{2}}^{\frac{N}{2}} |\tilde{e}[n]|^2$. r est une variable aléatoire de loi χ^2 et de degré de liberté $2(N+1)$, de moyenne et de variance respectivement $\sigma^2(N+1)$ et $\sigma^4(N+1)$. Etant donné que N est de l'odre de 30, le critère de blancheur adopté est la vérification de la relation suivante :

$$\text{r} \in \left[\sigma^2(N+1) - \sigma^2\sqrt{N+1}, \sigma^2(N+1) + \sigma^2\sqrt{N+1}\right]$$

3.2 Algorithme sous optimal

Il s'agit d'un algorithme itératif. En effet, $A_i[n]$ et $F_i[n]$ de la i-ème composante sont estimées en résolvant l'équation (3) par le recuit simulé. A chaque itération, une seule composante est reconstruite à la fois. Grâce à cette approche, on réduit le temps du calcul. Les étapes de l'algorithme se déroulent comme suit :

1. $i = 1$ et $r[n] = y[n]$,
2. Initialiser les paramètres de la i-ème composante. Par exemple, on calcule l'amplitude, la fréquence et la phase du maximum de la transformée de Fourier du signal bruité $r[n]$: $\theta_i = \{a_{FFT}, 0, 0, \theta_{FFT}, f_{FFT}, 0, 0\}$. Les autres paramètres sont tous mis à zéro.
3. Afin d'estimer θ_i, on itère I fois les étapes (a) (b) et (c) décrites dans le paragraphe (3.1).
4. Une fois $F_i[n]$ et $A_i[n]$ de la i-ème composante estimées, on reconstruit cette composante : $s_i[n] = A_i[n].e^{j\,\Phi_i[n]}$. Ensuite, on la soustrait du signal bruité afin d'obtenir un signal résiduel : $r[n] = r[n] - s_i[n]$.
5. Si $r[n]$ vérifie le test de blancheur (cf. paragraphe (3.1)), l'estimation est achevée. Dans le cas contraire, $i = i + 1$ et on estime la composante suivante à partir de l'étape 2.

Nous notons que l'algorithme sous-optimal fournit une bonne estimation du nombre de composantes compte tenu du niveau du bruit.

4 Bornes de Cramer Rao

Dans [2], la Matrice d'Information de Fisher (MIF) est déterminée pour les paramètres d'une phase polynomiale. En adaptant le calcul à la définition de nos paramètres (θ), on obtient :

$$MIF(\theta) = \frac{2}{\sigma^2}\, Re\left\{ \begin{array}{cc} \mathbf{A}_i^H\, \mathbf{A}_j & \mathbf{A}_i^H\, \phi_j \\ \phi_i^H\, \mathbf{A}_j & \phi_i^H\, \phi_j \end{array} \right\} \begin{array}{l} 1 \le i \le K \\ 1 \le j \le K \end{array} \quad (4)$$

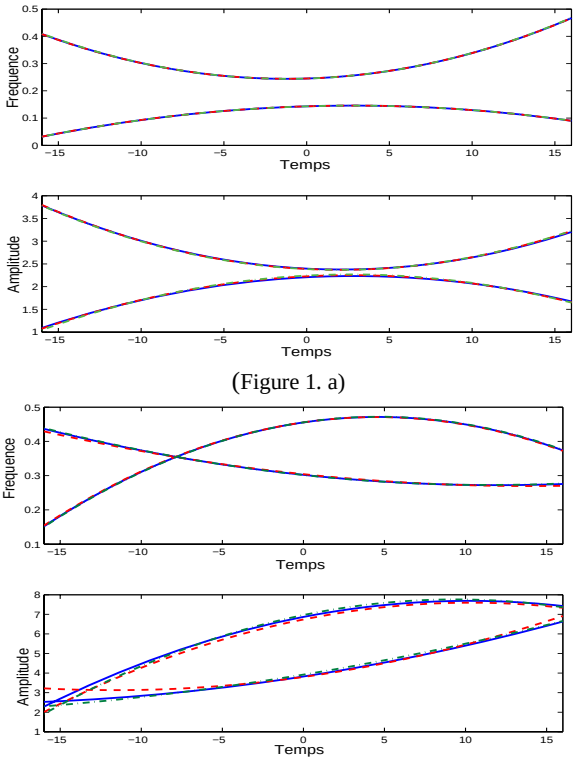
où $\mathbf{A}_i = [\, g_0(\underline{n}) \bullet e^{j\,\Phi_i(\underline{n})},\ g_1(\underline{n}) \bullet e^{j\,\Phi_i(\underline{n})},\ g_2(\underline{n}) \bullet e^{j\,\Phi_i(\underline{n})}]$
$\phi_i = j\,[\eta_{-1}(\underline{n}) \bullet s_i(\underline{n}),\ \eta_0(\underline{n}) \bullet s_i(\underline{n}),\ \eta_1(\underline{n}) \bullet s_i(\underline{n}),\ \eta_2(\underline{n}) \bullet s_i(\underline{n})]$

$\Phi_i(\underline{n})$ et $s_i(\underline{n})$ sont les vecteurs temporels de la phase et du signal de la i-ème composante, calculés à chaque instant n. $s_i(\underline{n})$ est égal à $A_i(\underline{n}) \bullet e^{(j\,\Phi_i(\underline{n}))}$. On note $\underline{n} = [\frac{-N}{2}, \frac{-N}{2}+1, ..., \frac{N}{2}]$, $\eta_{-1}[n] = 1$ et $\eta_i[n] = 2\pi(\sum_{k=\frac{-N}{2}}^{n} g_i[k] - \sum_{k=\frac{-N}{2}}^{0} g_i[k])$ pour $i = 0, 1, 2$ et $n \in [\frac{-N}{2}, \frac{N}{2}]$. $(\bullet)$ désigne la multiplication terme à terme.
La borne de Cramer Rao de θ est l'inverse de la matrice MIF donnée par (4). Notons que lorsque les trajectoires de fréquence deviennent très proches, la MIF tend vers une matrice singulière, ceci explique la difficulté de l'estimation dans ce cas. Dans le cadre des signaux multicomposantes, l'utilisation d'une base orthonormale permet le découplage entre les paramètres de l'amplitude d'une même composante. Toutefois, une forte corrélation subsiste entre les paramètres intercomposantes.

5 Résultats

Dans cette partie, nous utilisons des signaux simulés, formés de 33 points. La fréquence d'échantillonnage est de 1Hz. Les signaux ont deux composantes à phases polynomiales cubiques. Nous proposons d'étudier deux cas : dans le premier cas (I), les trajectoires de fréquences sont séparées. Dans le deuxième cas (II), elles se coupent.

Sur les figure 1 et 2, on a tracé les estimations de la FI et de la AI superposées avec les courbes réelles respectivement pour 20 dB et 10 dB. Nous obtenons une bonne estimation de la FI dans les deux cas (I) et (II). En revanche, nous remarquons que les performances d'estimation de l'AI sont meilleures dans le premier cas que dans le deuxième. En effet, le problème d'estimation est plus délicat en raison du croisements des fréquences. Sur la figure 3, des simulations de Monte Carlo illustrent l'efficacité de ces deux méthodes et leur robustesse au bruit. On a comparé les Erreurs Quadratiques Moyennes (EQM) aux BCR (4). Les EQM des paramètres d'amplitude sont très proches des BCR. En revanche, celles de la fréquence s'écartent des BCR. Cela s'explique d'une part par la forte corrélation entre les différents paramètres (Les termes en dehors de la diagonale de la FIM^{-1} ne sont pas négligeables) et d'autre part par la forte non linéarité en fonction des paramètres de fréquence.
Enfin, ces deux méthodes sont testées sur un signal réel. Il s'agit d'un signal émis par une baleine. Il comporte 31 échantillons. Toutes les figures sont présentées en fréquence réduite. Sur la figure 4, on peut voir que les deux méthodes donnent un résultat quasiment équivalent. En effet, on peut voir trois composantes de fréquences dont une est très basse fréquence. On voit également que deux composantes ont une énergie équivalente (ce qui ne facilite pas l'estimation). Les courbes estimées de la FI sont ensuite reportées sur le spectrogramme (figure 5). Le signal estimé, via les deux méthodes, est tracé sur la figure 6. L'autocorrélation du signal résiduel dans les deux cas est reportée sur la figure 7. En ce qui concerne l'algorithme sous optimal, la première composante estimée est la plus énergétique. Ensuite, comme les deux composantes restantes ont une énergie équivalente, à chaque réalisation de l'estimation, on peut estimer l'une ou l'autre en premier lieu. Le RSB est estimé à 14 dB. Finalement, sachant que l'hypothèse du bruit blanc n'est pas vrai en milieu sous-marin, les résultats obtenus restent relativement corrects.



(Figure 1. a)

(Figure 1. b)

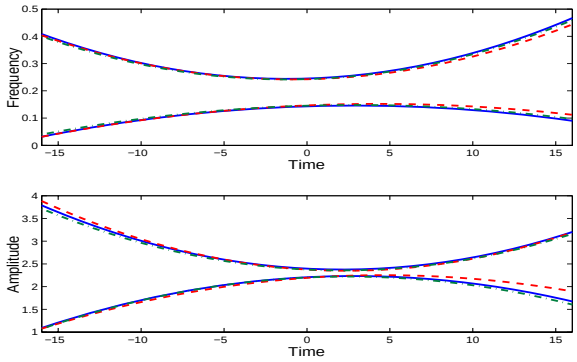
FIG. 1 – Exemples de FI : (a) cas I, (b) cas II, pour un RSB de 20 dB. Les courbes estimées de la FI et AI par l'algorithme optimal $(--)$ et celui sous optimal (-.) contre les courbes originales (-).

6 Conclusion

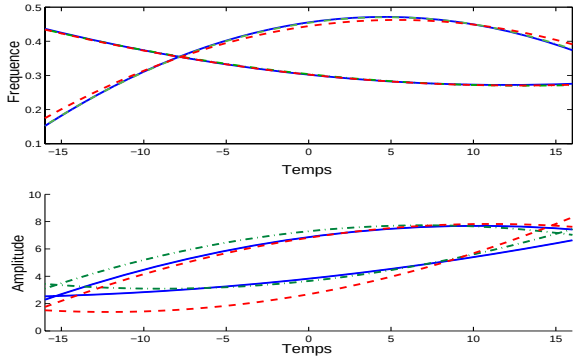
Cette communication vient compléter une étude locale faite sur des signaux monocomposante et étendre le résultat au cas multicomposantes. Nous avons proposé deux méthodes basées sur la maximisation de la vraisemblance et utilisant le recuit simulé. La première, optimale, opère simultanément sur tous les paramètres du modèle complet. La deuxième, sous optimale, reconstruit le signal composante par composante. La comparaison des deux méthodes montre qu'elles donnent des résultats proches. Toutefois la deuxième est plus avantageuse en temps de calcul (au moins trois plus rapide). La comparaison avec les BCR montrent une bonne performance pour l'estimation des paramètres d'amplitude. En revanche, le couplage entre les différents paramètres entraine une dégradation plus rapide pour les paramètres de fréquence. Un signal réel a été analysé pour valider cette étude. Nous envisageons de traiter des signaux longs en fusionnant les traitements locaux par la suite.

Références

[1] A. Francos and M. Porat, "Analysis and Synthesis of Multicomponent Signals Using Positive Time-Frequency Distributions," *IEEE Trans. On Signal Processing*, vol. 47, pp. 493–504, Feb. 1999.



(Figure 2. a)



(Figure 2. b)

FIG. 2 – Exemples de FI : (a) cas I, (b) cas II, pour un RSB de 10 dB. Les courbes estimées de la FI et AI par l'algorithme optimal $(--)$ et celui sous optimal (-.) contre les courbes originales (-).

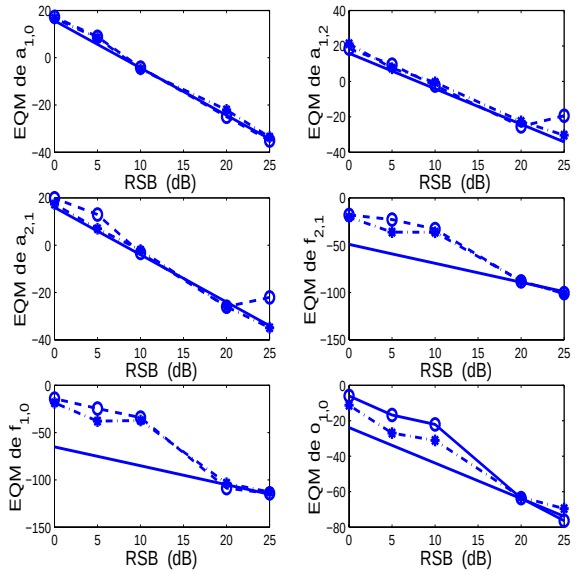


FIG. 3 – Les EQM de l'estimation des paramètres $\{a_{1,0}, a_{1,2}, a_{2,1}, f_{1,0}, f_{2,0}, \theta_{1,0}\}$ via l'algorithme optimal (-o), et celui sous optimal (-*) en fonction du RSB. Les EQM sont comparées aux BCR (-).

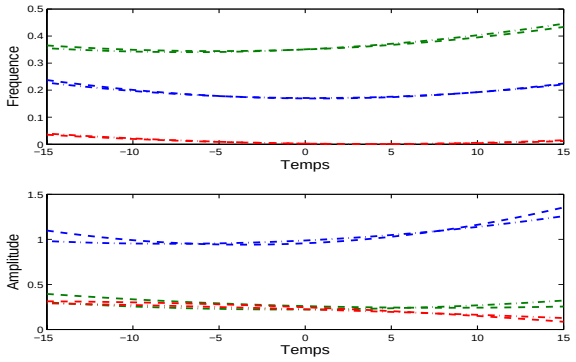


FIG. 4 – Signal réel : estimé via l'algorithme optimal (-.), et celui sous optimal (- -).

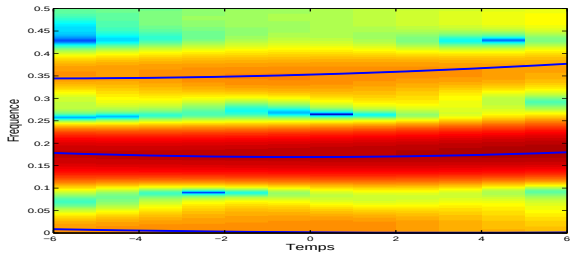


FIG. 5 – Superposition des fréquences estimées avec le spectrogramme du signal réel.

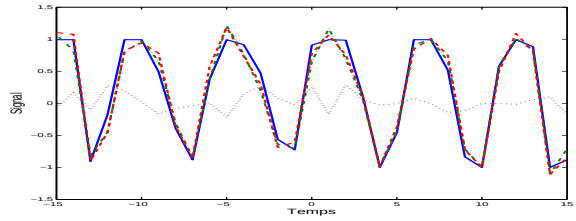


FIG. 6 – Superposition du signal réel avec celui estimé via l'algorithme optimal (- -) et le sous optimal (-.). signal résiduel (:).

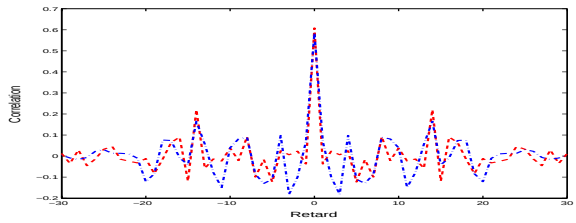


FIG. 7 – La corrélation du signal résiduel après avoir enlevé celui estimé avec l'algorithme : optimal (- -) et le sous optimal (-.).

rencester, United Kingdom, Dec. 14-16. 2004.

[4] M. Vieira, F. Leonard, M. Jabloun and N. Martin "Short Time Single Polynomial Phase Using Legendre functions,"in *EUSIPCO 2004*, Vienna, austria, Sept. 6-10. 2004, pp. 793–796.

[5] M. Z. Ikram and G. T Zhou, " Estimation of Multicomponent Polynomial signals of Mixed Orders," *IEEE Trans. On Signal Processing*.vol. 81, pp. 2293–2038, 2001.

[6] S. Barbarossa, A. Scaglione and G. B Giannakis, " Product Higher Order Function for Multicomponent Polynomial Phase Signal Modeling," *IEEE Trans. On Signal Processing*.vol. 49, pp. 691–708, Mar. 1998.

[2] B. Friedlander and J. M. Francos, " Estimation of Amplitude and Phase Parametetrs of Multicomponent Signals," *IEEE Trans. On Signal Processing*, vol. 43, pp. 917–926, Apr. 1995.

[3] M. Jabloun, M. Vieira, F. Leonard and N. Martin, " Local Orthonormal Decomposition for both Instantaneous Amplitude and Frequency of Highly Nonstationary discrete Signals," *IMA*. Ci-

Algorithme de zéro-diagonalisation conjointe pour la séparation de sources déterministes

El Mostafa FADAILI, Nadège THIRION-MOREAU et Eric MOREAU

STD, ISITV, Université du Sud Toulon Var
avenue G. Pompidou, BP56, 83162 La Valette du Var Cedex
{fadaili, thirion, moreau}@univ-tln.fr

Résumé – Nous considérons le problème de la séparation aveugle de sources déterministes par zéro-diagonalisation conjointe d'un ensemble particulier de matrices issues des représentations temps-fréquence spatiales quadratiques des observations. Nous proposons un nouvel algorithme de zéro-diagonalisation conjointe sous transformation non unitaire, obtenu par optimisation paramétrique d'un critère que nous donnons. L'avantage d'une telle approche est de ne requérir aucun blanchiment préalable des observations. Nous proposons également un détecteur de points temps-fréquence permettant de construire l'ensemble des matrices à zéro-diagonaliser. Enfin, des simulations informatiques illustrent l'amélioration des performances en comparant l'algorithme proposé à d'autres algorithmes basés soit sur de la diagonalisation conjointe sous transformation unitaire ou non unitaire soit sur de la zéro-diagonalisation sous transformation unitaire.

Abstract – The problem of blind separation of deterministic sources based on joint zero-diagonalization of a particular set of spatial quadratic time-frequency matrices is considered. We propose a new non unitary joint zero-diagonalization algorithm. It is obtained thanks to the parametric optimization of a criterium we give. The advantage of such an approach is that it does not require a pre-whitening stage any more. We also detail a time-frequency points selection procedure used in order to build the set of matrices to be joint zero-diagonalized. Finally, computer simulations illustrate the effectiveness of the proposed algorithm by comparing it with other ones based on unitary or non unitary joint-diagonalization and on unitary joint zero-diagonalization.

1 Introduction

Dans ce papier, nous nous intéressons au problème de la séparation aveugle de mélanges instantanés de signaux sources basée sur l'utilisation de distributions temps-fréquence spatiales quadratiques (DTFSQ). Ces dernières années, ces méthodes ont fait l'objet de nombreux travaux [1]-[8][10] du fait qu'elles permettent de considérer une classe plus large de sources que la classe habituelle des sources aléatoires et indépendantes. Elles ont notamment permis d'aborder le cas de sources non stationnaires, même s'il existe d'autres manières de traiter ce problème [9]. Une partie des algorithmes proposés [1][3][7] reposent, après blanchiment, sur la diagonalisation conjointe d'un ensemble particulier de matrices.

Plus récemment, on s'est intéressé à ces deux problèmes que sont la suppression de l'étape de blanchiment [4]-[6][10] et la prise en compte d'un nouveau type de matrices temps-fréquence (*t-f*) se devant cette fois d'être zéro-diagonalisées conjointement [2][10]. Dans tous les cas, ces méthodes requièrent une étape préalable de sélection de points *t-f* "utiles" en ce sens qu'ils permettent la construction d'ensembles de matrices devant être diagonalisées conjointement et/ou zéro-diagonalisées conjointement. Il a d'abord été proposé dans [1] de diagonaliser conjointement les matrices issues des DTFSQ calculées en des points *t-f* ne correspondant qu'à des auto-termes sur les sources, cette méthode s'appliquant après blanchiment des observations. Par la suite, il a été montré dans [4]-[6], qu'en utilisant des algorithmes de diagonalisation conjointe (DC) non unitaire l'étape de blanchiment pouvait être supprimée ce qui présente de nombreux avantages parmi lesquels le fait que l'on obtienne généralement de meilleures performances et que l'on puisse considérer le cas de sources "corrélées".

Dans le même temps, une approche complémentaire fut proposée dans [2], consistant cette fois à zéro-diagonaliser conjointement (encore après une étape préalable de blanchiment) un autre type de matrices issues des DTFSQ : celles calculées en des points *t-f* ne correspondant qu'à des inter-termes sur les sources. Dans [10], nous avons proposé un premier algorithme de zéro-diagonalisation conjointe (ZDC) sous transformation non unitaire basé sur un schéma d'optimisation de type algébrique, l'avantage d'une telle approche étant à nouveau de ne plus requérir de blanchiment préalable des observations. Dans ce papier, nous en proposons un nouveau fondé cette fois sur un schéma d'optimisation de type Newton et basé sur une modélisation paramétrique des colonnes de la matrice de séparation recherchée du même type que celle utilisée dans [12]. Au préalable, nous présentons le critère sur lequel est fondé l'algorithme. Nous détaillons également une procédure de sélection automatique des points *t-f* permettant la construction d'un ensemble de matrices devant être zéro-diagonalisées conjointement.

2 Modèle & hypothèses

Considérons le problème de la séparation aveugle de sources dans lequel $N \in \mathbb{N}^+ \setminus \{0,1\}$ signaux baptisés sources sont reçus au niveau d'une antenne de $P \in \mathbb{N}^+ \setminus \{0,1\}$ capteurs. Plaçons-nous de plus dans le cas $P \geq N$, *i.e.* sur-déterminé (pour le cas sous-déterminé [8]). En notations matricielles, la relation qui lie l'entrée à la sortie du système mélangeant, est

donnée dans le cas non bruité par :

$$\mathbf{x}(t) = \mathbf{A}\,\mathbf{s}(t) \tag{1}$$

où $\mathbf{A}$ est la matrice de mélange de dimensions $P \times N$, $\mathbf{x}(t) = (x_1(t), \ldots, x_P(t))^{\mathrm{T}}$ le vecteur $P \times 1$ des observations et $\mathbf{s}(t) = (s_1(t), \ldots, s_N(t))^{\mathrm{T}}$ le vecteur $N \times 1$ des sources et $(\cdot)^{\mathrm{T}}$ l'opérateur de transposition. Dans toute la suite on se placera dans le cas de signaux sources réels et une matrice de mélange réelle.

Nous considérons les hypothèses suivantes :

- La matrice de mélange instantané $\mathbf{A}$ est inconnue mais de rang plein.
- Le nombre de sources N est connu.
- Les sources $s_i(t)$, $i = 1, \ldots, N$ sont des signaux déterministes.

Supposant les signaux sources inobservables et la matrice de mélange inconnue, l'objectif est d'estimer une "inverse" de la matrice de mélange afin de retrouver les contributions des différentes sources. En général, celles-ci ne sont restituées qu'à une amplitude et une permutation près. C'est pourquoi, on considère généralement que la matrice de mélange $\mathbf{A}$ a été estimée quand la matrice obtenue, notée $\hat{\mathbf{A}}$, se factorise sous la forme $\hat{\mathbf{A}} = \mathbf{ADP}$ avec $\mathbf{D}$ matrice diagonale inversible et $\mathbf{P}$ matrice de permutation.

En ce qui concerne finalement la séparation de sources à proprement parler, quand $P \geq N$, *i.e.* dans le cas sur-déterminé, une estimée des sources $\hat{\mathbf{s}}(t)$ s'obtient au moyen de $\hat{\mathbf{A}}^{\dagger}\mathbf{x}(t)$ ($\hat{\mathbf{A}}^{\dagger}$ désignant la pseudo-inverse de la matrice $\hat{\mathbf{A}}$). Quand $P < N$, *i.e.* dans le cas sous-déterminé, il n'y a plus une solution unique et la séparation effective des sources nécessite des hypothèses et des développements supplémentaires.

3 Transformée spatiale quadratique

Une transformée spatiale quadratique (TSQ) est définie [11] comme :

$$\begin{aligned}(\mathbf{x}(t), \mathbf{x}(t)) &\overset{\mathsf{TSQ}}{\longmapsto} \mathbf{D}_{xx}(t, \nu; R) = \mathbf{D}_x(t, \nu) \\ &= \int_{\mathbb{R}^2} \mathbf{x}(\theta)\mathbf{x}^{\mathrm{H}}(\theta') R(\theta, \theta'; t, \nu) d\theta d\theta'\end{aligned} \tag{2}$$

où R est le noyau de la transformation et où $(\cdot)^{\mathrm{H}}$ désigne l'opérateur de transposition conjugaison. Les termes diagonaux de cette matrice sont appelés "auto-termes", ils correspondent aux transformées quadratiques associées à chacune des composantes du vecteur $\mathbf{x}(t)$. Les termes hors diagonale sont quant à eux appelés "inter-termes", ils correspondent aux transformations bilinéaires associées à des couples de composantes distinctes de ce même vecteur. Pour des raisons physiques, on contraint souvent le noyau de la TSQ à satisfaire la propriété suivante

$$R(\theta, \theta'; t, \nu) = R^{*}(\theta', \theta; t, \nu) \tag{3}$$

où $(\cdot)^{*}$ est l'opérateur de conjugaison. Alors, la TSQ possède la propriété de symétrie hermitienne suivante :

$$\mathbf{D}_x(t, \nu) = \mathbf{D}_x{}^{\mathrm{H}}(t, \nu). \tag{4}$$

Il s'avère en fait que cette condition n'est pas très restrictive dans la mesure où elle est satisfaite par de nombreuses transformations dont les distribution de Wigner-Ville (WV), de Choï-Williams (CW), de Pseudo Wigner-Ville (PWV), Pseudo Wigner-Ville lissée (moyennant l'adjonction de conditions sur l'une des deux fenêtres qui doit être réelle), celle de Born-Jordan (BJ), etc...

On suppose également dans ce qui suit que les représentations *t-f* de chaque signaux sources ne se recouvrent pas trop ceci dans le but de pouvoir sélectionner des points *t-f* en vue de la construction des ensembles de matrices à diagonaliser ou à zéro-diagonaliser conjointement.

On montre enfin [13] que la DTFSQ $\mathbf{D}_x(t, \nu)$ du vecteur des observations admet naturellement la décomposition suivante :

$$\mathbf{D}_x(t, \nu) = \mathbf{A}\mathbf{D}_s(t, \nu)\mathbf{A}^{\mathrm{T}} \tag{5}$$

où $\mathbf{D}_s(t, \nu)$ est la DTFSQ du vecteur des signaux sources.

4 Zéro-diagonalisation non unitaire

Soit un ensemble $\{\mathbf{Z}_i, i = 1, \ldots, N_m\}$ de matrices hermitiennes tel qu'il existe une matrice $\mathbf{B}$ telle que, pour tout i,

$$\mathbf{B}\mathbf{Z}_i\mathbf{B}^{\mathrm{T}}$$

soit une matrice zéro-diagonale, *i.e.* une matrice dont les éléments diagonaux sont tous nuls. Dans le but d'estimer la matrice $\mathbf{B}$ à partir de la factorisation précédente, on propose l'utilisation de la fonction objectif suivante :

$$\mathcal{C}(\mathbf{B}) = \sum_{i=1}^{N_m} \|\mathsf{Diag}\{\mathbf{B}\mathbf{Z}_i\mathbf{B}^{\mathrm{T}}\}\|^2 \tag{6}$$

où l'opérateur $\mathsf{Diag}\{\cdot\}$ renvoie une matrice diagonale construite à partir de la diagonale de la matrice en argument. Il s'agit donc de trouver une matrice en minimisant par rapport à $\mathbf{B}$ la somme des carrés des éléments de la diagonale de $\mathbf{B}\mathbf{Z}_i\mathbf{B}^{\mathrm{T}}$.

Afin d'effectuer cette minimisation, nous proposons d'utiliser un algorithme de gradient ou de Newton. Cependant, avant de calculer le gradient de $\mathcal{C}$, il convient de remarquer que l'on peut toujours écrire $\mathcal{C}(\mathbf{B})$ sous la forme

$$\mathcal{C}(\mathbf{B}) = \sum_{i=1}^{N_m}\sum_{\ell=1}^{N} |\mathbf{b}_\ell\mathbf{Z}_i\mathbf{b}_\ell{}^{\mathrm{T}}|^2 = \sum_{\ell=1}^{N} \mathcal{C}_l(\mathbf{b}_\ell) \tag{7}$$

où les vecteurs $\mathbf{b}_\ell$, $\ell = 1, \ldots, N$ sont les vecteurs lignes de la matrice $\mathbf{B}$ et où

$$\mathcal{C}_l(\mathbf{b}) = \sum_{i=1}^{N_m} |\mathbf{b}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}}|^2 \,. \tag{8}$$

Ainsi la minimisation de $\mathcal{C}(\mathbf{B})$ peut se faire ligne par ligne à partir de $\mathcal{C}_l(\mathbf{b})$. Donc considérons un vecteur ligne quelconque noté $\mathbf{b}$ en lui imposant une norme égale à 1 et supposons que celui-ci dépende de paramètres regroupés dans un $(K, 1)$ vecteur ($K = N - 1$) noté $\boldsymbol{\theta} = (\theta_k)$. Une telle paramétrisation peut par exemple se trouver dans [12]. Comme les matrices $\mathbf{Z}_i$ sont hermitiennes, le terme $\mathbf{b}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}}$ est réel et on a

$$\frac{\partial \mathcal{C}_l(\mathbf{b})}{\partial \theta_k} = 2\sum_{i=1}^{N_m} (\mathbf{b}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}}) \frac{\partial(\mathbf{b}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}})}{\partial \theta_k} \,. \tag{9}$$

Comme par ailleurs on a

$$\frac{\partial(\mathbf{b}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}})}{\partial \theta_k} = 2\frac{\partial \mathbf{b}}{\partial \theta_k}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}} \tag{10}$$

en introduisant la matrice $\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}}$ définie par

$$\left(\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}}\right)_{l,m} = \frac{\partial b_m}{\partial \theta_l} \tag{11}$$

on obtient

$$\frac{\partial(\mathbf{b}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}})}{\partial \boldsymbol{\theta}} = 2\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}} \tag{12}$$

et donc en utilisant en plus (10) dans (9)

$$\mathbf{g}_{\mathcal{C}_l}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{\partial \mathcal{C}_l(\mathbf{b})}{\partial \boldsymbol{\theta}} = 4\sum_{i=1}^{N_m}(\mathbf{b}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}})\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}}. \tag{13}$$

Le calcul de $\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}}$ dépend de la paramétrisation considérée. Nous considérons dans la suite celle explicitée dans [12]. Un algorithme de gradient s'écrit alors

$$\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1} - \lambda\, \mathbf{g}_{\mathcal{C}_l}(\boldsymbol{\theta}^{i-1}) \tag{14}$$

où λ est une constante positive.

Cependant dans notre cas le Hessien peut aussi être calculé afin d'obtenir un algorithme de Newton. D'abord, à partir de (9), on a dans le cas scalaire

$$\frac{\partial^2 \mathcal{C}_l(\mathbf{b})}{\partial\theta_k\partial\theta_\ell} = 2\sum_{i=1}^{N_m}\frac{\partial(\mathbf{b}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}})}{\partial\theta_k}\frac{\partial(\mathbf{b}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}})}{\partial\theta_\ell} + (\mathbf{b}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}})\frac{\partial^2(\mathbf{b}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}})}{\partial\theta_k\partial\theta_\ell}. \tag{15}$$

Ecrivons maintenant l'expression précédente sous une forme vectorielle. A cette fin, on pose

$$\frac{\partial^2 \mathbf{b}}{\partial \boldsymbol{\theta}^2} = \left(\left(\frac{\partial^2 \mathbf{b}}{\partial\boldsymbol{\theta}\partial\theta_1}\right)^{\mathrm{T}}, \cdots, \left(\frac{\partial^2 \mathbf{b}}{\partial\boldsymbol{\theta}\partial\theta_K}\right)^{\mathrm{T}}\right)^{\mathrm{T}} \tag{16}$$

et

$$\frac{\partial^2 \mathcal{C}_l(\mathbf{b})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} = \left(\frac{\partial^2 \mathcal{C}_l(\mathbf{b})}{\partial\theta_k\partial\theta_\ell}\right) \tag{17}$$

On montre alors que

$$\mathbf{H}_{\mathcal{C}_l}(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \frac{\partial^2 \mathcal{C}_l(\mathbf{b})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} = 4\sum_{i=1}^{N_m}\Big(\mathbf{C}_i(\boldsymbol{\theta}) + (\mathbf{b}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}})\mathbf{D}_i(\boldsymbol{\theta})\Big) \tag{18}$$

où

$$\mathbf{C}_i(\boldsymbol{\theta}) = 2\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}}\mathbf{b}\mathbf{Z}_i\left(\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}} \tag{19}$$

et

$$\mathbf{D}_i(\boldsymbol{\theta}) = \mathsf{Unvec}\{\frac{\partial^2 \mathbf{b}}{\partial \boldsymbol{\theta}^2}\mathbf{Z}_i\mathbf{b}^{\mathrm{T}}\} + \frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}}\mathbf{Z}_i\left(\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}}\right)^{\mathrm{T}} \tag{20}$$

et où l'opérateur $\mathsf{Unvec}\{\cdot\}$ renvoie une matrice carrée à partir de son argument vectoriel. Finalement l'algorithme de Newton s'écrit sous la forme

$$\boldsymbol{\theta}^i = \boldsymbol{\theta}^{i-1} - \lambda\, \mathbf{H}_{\mathcal{C}_l}^{-1}(\boldsymbol{\theta}^{i-1})\, \mathbf{g}_{\mathcal{C}_l}(\boldsymbol{\theta}^{i-1}) \tag{21}$$

où λ est une constante positive. Notons que dans le cas rectangulaire, pour éviter des problèmes d'inversion du Hessien, on passe à un schéma d'optimisation de type Levenberg-Marquardt.

5 Sélection automatique des points *t-f*

Il existe 4 types distincts de points *t-f* : 1- ceux qui ne correspondent qu'à des auto-termes sur les sources (en ces points la DTFSQ des sources est diagonale, propriété perdue sous l'effet du mélange), 2- ceux qui ne correspondent qu'à des inter-termes sur les sources (en ces points la DTFSQ des sources est zéro-diagonale, propriété perdue sous l'effet du mélange), 3- ceux qui correspondent à la fois à des inter- et à des auto-termes sur les sources (en ces points la DTFSQ des sources ne possède pas de structure algébrique particulière), et enfin 4- ceux où il n'y a ni inter- ni auto-termes sur les sources ; seuls les deux premiers s'avèrent intéressant d'un point de vue de la séparation de sources. Dans cette partie, nous expliquons la sélection automatique des points *t-f* permettant la construction de l'ensemble des matrices à zéro-diagonaliser. En outre, pour pouvoir comparer les algorithmes de ZDC unitaire et non unitaire sur des ensembles identiques de points, nous présentons une procédure de sélection des points *t-f* à partir des DTFSQ des observations blanchies. Pour la sélection automatique des points *t-f* permettant la construction de l'ensemble des matrices à diagonaliser conjointement, on procède comme dans [3] car les points *t-f* sont sélectionnés, pour les mêmes raisons que précédemment, à partir des DTFSQ des observations blanchies (dans [4] ou [6], on décrit deux autres procédures de sélection automatique à partir des DTFSQ des observations non blanchies cette fois).

Etant donné les hypothèses faites ($\mathbf{A}$ réelle, DTFSQ à symétrie hermitienne), on montre que les DTFSQ des sources, $\mathbf{D}_s(t,\nu)$, et des observations blanchies, $\underline{\mathbf{D}}_x(t,\nu)$, en des points *t-f* ne correspondant qu'à des auto-termes sur les sources sont des matrices réelles et non nulles alors qu'elles sont généralement complexes en des points *t-f* ne correspondant qu'à des inter-termes sur les sources. La trace ($\mathsf{trace}\{.\}$) étant invariante sous transformation unitaire, on a par ailleurs $\mathsf{trace}\{\underline{\mathbf{D}}_x(t,\nu)\} = \mathsf{trace}\{\mathbf{D}_s(t,\nu)\}$. L'une des manières les plus naturelles de sélectionner les matrices issues des DTFSQ (des observations blanchies) à zéro-diagonaliser consiste à choisir celles qui sont telles que la valeur absolue de leur trace ($|\mathsf{trace}\{.\}|$) soit faible (inférieure à un seuil ε_2) tandis que la norme euclidienne de leur partie imaginaire ($\|\Im\{.\}\|$) soit grande (supérieure à un seuil ε_1). Ce qui donne en résumé :

• pour la ZDC choisir (t,ν) tels que :

$$\begin{cases} \|\Im\{\underline{\mathbf{D}}_x(t,\nu)\}\| > \varepsilon_1 \\ |\mathsf{trace}\{\underline{\mathbf{D}}_x(t,\nu)\}| < \varepsilon_2 \end{cases}. \tag{22}$$

6 Simulations informatiques

Dans cette partie, nous illustrons au moyen de simulations informatiques l'intérêt de cet algorithme en comparant les performances obtenues avec celles d'autres algorithmes basés sur de la DC (unitaire ou non) et de la ZDC (unitaire). On utilise la distribution de Pseudo-Wigner Ville spatiale. Dans cet exemple, on considère le cas de 2 sources (une somme de sinusoïdes, une modulation sinusoïdale de fréquence) reçues sur 2 capteurs. La partie réelle de la distribution de PWV spatiale des sources est représentée au niveau de la figure 1, celle de la distribution de PWV spatiale des observations au niveau de la figure 2. Les 651 points *t-f* détectés et utilisés par tous les algorithmes que nous comparons sont représentés par un "plus" au niveau de la figure 3 (haut), à gauche pour la DC et à droite pour la ZDC. Pour faciliter la compréhension, ils sont superposés à la trace de la DTFSQ des sources pour la DC et à la

somme des termes hors diagonale pour la ZDC. Ils ont été obtenus en fixant $\varepsilon_1 = 5.9$ et $\varepsilon_2 = 4.5$. En figure 3 (bas) les points ont été choisis à la main. En utilisant un indice de performance classique [6], les résultats obtenues sont résumés au niveau du Tableau 1. On compare la DC (JD) et la ZDC (JZD) des matrices sélectionnées après une étape de blanchiment. On donne aussi le résultat obtenu par l'algorithme classique JADE. On compare finalement la DC ($\mathsf{JD}_{\mathsf{NO,YER}}$) et l'algorithme proposé basé sur la ZDC ($\mathsf{JZD}_{\mathsf{NO,PAR}}$) des matrices sélectionnées, ces deux derniers algorithmes opérant sans étape de blanchiment. Dans cet exemple, les résultats obtenus sont meilleurs avec l'algorithme proposé et dans tous les cas bien meilleurs que lorsqu'une étape de blanchiment est considérée.

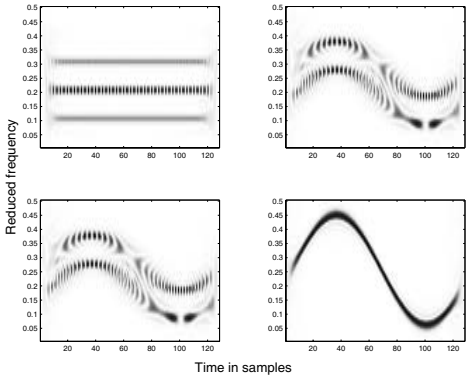


FIG. 1: *Partie réelle de la distribution de* PWV *spatiale des sources*

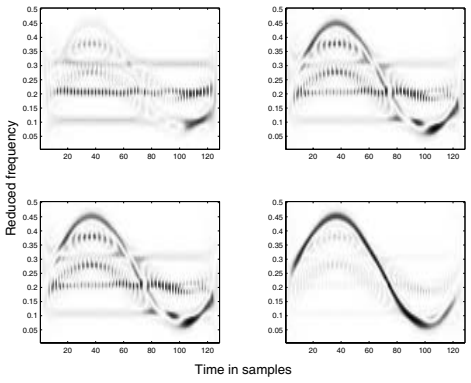


FIG. 2: *Partie réelle de la distribution de* PWV *spatiale du mélange*

TAB. 1: Comparaison des performances obtenues au moyen de : la méthode proposée ($\mathsf{JZD}_{\mathsf{NO,PAR}}$), JADE, DC non orthogonale ($\mathsf{JD}_{\mathsf{NO,YER}}$), DC orthogonale (JD), ZDC orthogonale (JZD) sur des ensembles de points obtenus en mode automatique ou choisis à la main.

Méthodes	Perf. (mode auto.)	Perf. (main)
JD (dB)	-29.3	-28.76
JZD (dB)	-29.44	-29.49
$\mathsf{JD}_{\mathsf{NO,YER}}$ (dB)	-37.56	-42.01
$\mathsf{JZD}_{\mathsf{NO,PAR}}$ (dB)	-40.31	-42.1
JADE (dB)	-21.48	

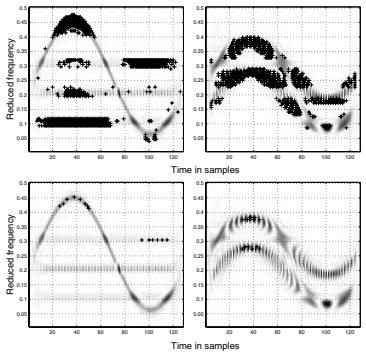


FIG. 3: *Haut : les 651 points t-f sélectionnés de manière automatique pour la* DC *(gauche) ; les 651 retenus pour la* ZDC *(droite); Bas : points choisis à la main (14 pour la* DC *et 13 pour la* ZDC*)*

Références

[1] A. Belouchrani, M. G. Amin, "Blind source separation based on time-frequency signal representations," *IEEE Trans. Signal Processing*, Vol. 46, No. 11, pp. 2888-2897, November 1998.

[2] A. Belouchrani, K. Abed-Meraim, M. G. Amin, A. M. Zoubir, "Joint anti-diagonalisation for blind source separation", In Proc. ICASSP'2001, Salt Lake City, USA, pp. 2789-2792, May 2001.

[3] L. Giulieri, N. Thirion-Moreau, P.-Y. Arquès, "Blind sources separation based on bilinear time-frequency representations: a performance analysis", In Proc. ICASSP'2002, Orlando, USA, pp. 1649-1652, May 2002.

[4] L. Giulieri, N. Thirion-Moreau and P.-Y. Arquès, "Blind sources separation based on quadratic time-frequency representations: a method without pre-whitening", In Proc. ICASSP'2003, Honk-Kong, pp. 289-292, April 2003.

[5] A. Bousbia-Salah, A. Belouchrani and H. Bousbia-Salah, "A one step time-frequency blind identification", in Proc. ISSPA'2003, Paris, France, July 2003.

[6] L. Giulieri, H. Ghennioui, N. Thirion-Moreau and E. Moreau, "Non-Orthogonal Joint-Diagonalization of Spatial Quadratic Time Frequency Matrices for Source Separation", *IEEE Signal Processing Letters*, Vol. 12, n°5, pp. 415-418, May 2005.

[7] C. Févotte and C. Doncarli, "Two contributions to blind source separation using time-frequency distributions", *IEEE Signal Processing Letters*, Vol. 11, No. 3, pp. 386-389, March 2004.

[8] N. Linh-Trung, A. Belouchrani, K. Abed-Meraim and B. Boashash, "Separating more sources than sensors using time-frequency distributions", in Proc. ISSPA'2001, Kuala-Lumpur, Malaysia, August 2001.

[9] Pham, D.-T. and Cardoso, J.-F, "Blind separation of instantaneous mixtures of non-stationary sources", In Proc. ICA'2000, Helsinki, Finland, pp. 187-193, 2000.

[10] E. M. Fadaili, N. Thirion-Moreau and E. Moreau, "Non-Orthogonal Zero-Diagonalization for Source Separation Based on Time-Frequency Representation", in Proc. ICASSP'2005, Philadelphia PA, USA, vol. IV, pp. 297-300, March 2005.

[11] P. Flandrin, *Time-Frequency/Time-Scale Analysis*, Academic Press, 1999.

[12] A.J. Van Der Veen, "Joint diagonalization via subspace fitting techniques", In Proc. ICASSP'2001, pp. 2773-2776, Salt Lake City, USA, May 2001.

[13] Y. Zhang and M. Amin, "Subspace analysis of spatial time-frequency distribution matrices", in *IEEE Trans. on Signal Processing*, Vol.49, No. 4, pp.747-759, April 2001.

Suivi de Trajectoires Temps-Fréquence par Filtrage Particulaire

Corentin DUBOIS[1] et Manuel DAVY[2]

[1]IRCCyN UMR 6597
1 rue de la Noë, BP 92101, 44321 Nantes Cedex 03, France

[2]LAGIS UMR 8146
Cité scientifique, BP 48, 59651 Villeneuve d'Ascq Cedex, France
corentin.dubois@irccyn.ec-nantes.fr

Résumé – Cet article reprend notre travail précédent sur la détection et le suivi de composantes mono-fréquence. Il propose de placer ce problème d'estimation séquentiel dans le domaine temporel en utilisant un modèle élargi contenant l'information de phase. Pour réaliser cette estimation, un algorithme de Monte Carlo performant, connu sous le nom de filtrage particulaire Rao-Blackwellisé, est implémenté.

Abstract – This paper builts on our previous work on detection and tracking of individual frequency components. It proposes to cast this sequential estimation problem into the time domain using an enlarged model featuring phase information. To perform estimation, an efficient Monte Carlo algorithm, known as Rao-Blackwellised particle filter, is implemented.

1 Introduction

Un problème important en traitement du signal est l'extraction séquentielle d'information fréquentielle dépendante du temps, à partir d'un signal non stationnaire. Les domaines d'application sont divers avec, entre autres, le traitement de la musique ou de la parole, la surveillance vibratoire des systèmes mécaniques ou, plus généralement, l'analyse informatique de l'environnement sonore (Computational Auditory Scene Analysis, CASA). Dans notre précédent travail [1], nous avons traité ce problème par une méthode permettant la détection et le suivi de plusieurs fréquences instantanées afin de décomposer le signal, en ligne, en une somme de composantes fréquentielles. Dans cette méthode, le signal est étudié via son spectrogramme et la détection ainsi que le suivi des composantes fréquentielles sont faits dans le plan Temps-Fréquence (TF), ce qui donne de bons résultats. Cependant, le choix du plan TF comme espace de représentation du signal pour effectuer l'estimation de la fréquence et de l'amplitude des composantes détectées, a pour conséquence la perte de l'information sur la phase initiale. Cet article propose de modifier le modèle utilisé afin de mener la détection et le suivi des composantes fréquentielles directement dans le domaine temporel, sans perte d'information.

Dans [1], chaque composante est caractérisée, à l'instant t, par une fréquence instantanée et une amplitude. Dans le modèle proposé ici, un terme de phase est ajouté. Pour un souci de simplicité, ou plus précisement de linéarité comme nous le verrons plus loin, le modèle retenu n'est pas de la forme $a\cos(2\pi ft+\varphi)$ avec φ la phase initiale mais plutôt de la forme $a\cos(2\pi ft) + b\sin(2\pi ft)$. Ainsi, le problème se résume en l'estimation, à chaque instant t, du nombre k_t de composantes fréquentielles, des fréquences $\mathbf{f}_t = [f_{t,1}, \ldots, f_{t,k_t}]$ et des amplitudes $\mathbf{a}_t = [a_{t,1}, \ldots, a_{t,k_t}]$ et $\mathbf{b}_t = [b_{t,1}, \ldots, b_{t,k_t}]$.

A l'instar du problème de suivi de cibles mobiles, la détection et le suivi des composantes fréquentielles sont réalisés dans le cadre bayésien séquentiel. Le problème de filtrage statistique consiste donc en l'estimation séquentielle du paramètre discret k_t et du vecteur continu $\boldsymbol{\theta}_t = [\mathbf{f}_t\,,\,\mathbf{a}_t\,,\,\mathbf{b}_t]$ dont la dimension dépend de k_t. Ceci nécessite la définition d'une vraisemblance (équation d'observation) et d'une densité *a priori* (équation de transition). Pour rendre l'étude locale, on considère une fenêtre $\mathbf{w}$ de longueur $L_{\mathbf{w}}$ centrée sur l'instant t. Soit $\mathbf{x} = [x_1, x_2, \ldots]$ un signal à temps discret, le vecteur d'observation $\mathbf{y}_t$, à l'instant t, est le signal fenêtré. Pour $\tau = 1, \ldots, L_{\mathbf{w}}$:

$$\mathbf{y}_t = \mathbf{x}[t + \tau - L_{\mathbf{w}}/2] \cdot \mathbf{w}[\tau] \qquad (1)$$

La vraisemblance est notée $p\big(\mathbf{y}_t|k_t, \boldsymbol{\theta}_t\big)$ et l'*a priori* est noté $p\big(k_t, \boldsymbol{\theta}_t|k_{t-1}, \boldsymbol{\theta}_{t-1}\big)$. Associés à $p\big(k_0, \boldsymbol{\theta}_0\big)$, ils forment un système de Markov à sauts (Jump Markov System, JMS) [2].

Pour réaliser la détection et le suivi, nous utilisons les méthodes de Monte Carlo Séquentielles (MCS). Ces méthodes permettent l'estimation en ligne de paramètres en combinant la puissance des méthodes d'échantillonnage de Monte Carlo avec l'inférence bayésienne [3]. Dans ce cas, l'algorithme est souvent appelé Filtre Particulaire (FP). Dans [1], le modèle d'observation est non linéaire et le Filtre de Kalman (FK) ne peut pas être implémenté alors que le FP prend en compte les changements de dimension de l'état et les non linéarités du modèle d'observation [4]. Le modèle que nous proposons ici permet de nuancer ce choix. En effet, conditionnellement à k_t et à $\mathbf{f}_t$, le modèle est linéaire gaussien. Ainsi, le nombre de composantes ainsi que leurs fréquences respectives étant estimés, il est possible de calculer analytiquement les amplitudes, en combinant le FP pour calculer la distribution de k_t et de $\mathbf{f}_t$ avec une banque de FK pour calculer la distribution de $\mathbf{a}_t$ et de $\mathbf{b}_t$. Cette stratégie est connue sous le nom de "Rao-Blackwellisation" [5].

Ce papier est organisé de la manière suivante : dans la partie 2, on présente le modèle bayésien séquentiel. Puis, dans la partie 3, on explicite l'algorithme mis en œuvre pour cette méthode. En particulier, l'étape de Rao-Blackwellisation, dans l'algorithme de FP, est développée. Quelques résultats sont donnés dans la partie 4 avant de conclure et de présenter des perspectives dans la partie 5.

2 Modèle bayésien séquentiel

2.1 Equation d'observation

Les fréquences et les amplitudes sont supposées stationaires par rapport à la longueur $L_{\mathbf{w}}$ de $\mathbf{w}$. Soit g_t la fonction définie par :

$$\begin{array}{rccc} g_t: & \mathbb{N}\times\mathbb{R}^{3k_t} & \longrightarrow & \mathbb{R}^{L_{\mathbf{w}}} \\ & (k_t,\boldsymbol{\theta}_t) & \longmapsto & \mathbf{s}_t\cdot\mathbf{w} \end{array} \tag{2}$$

avec, pour $\tau = 1,\ldots,L_{\mathbf{w}}$,

$$\mathbf{s}_{t,\tau} = \sum_{j=1}^{k_t} a_{t,j}\cos(2\pi f_{t,j}\tau) + b_{t,j}\sin(2\pi f_{t,j}\tau) \tag{3}$$

La fonction g_t, équation (2), a pour but de rendre le modèle génératif de l'observation, équation (3), local autour de l'instant t, en le multipliant par la fenêtre $\mathbf{w}$, sans perte d'information. L'équation d'observation est la suivante, pour $\xi = 1,\ldots,L_{\mathbf{w}}$:

$$\mathbf{y}_t[\xi] = g_t(k_t,\boldsymbol{\theta}_t) + v_t^{\mathbf{y}}[\xi] \tag{4}$$

où $v_t^{\mathbf{y}}$ est un bruit blanc gaussien de moyenne nulle et de variance $r_t^{\mathbf{y}}$. Comme dans [1], on autorise la variance à évoluer au cours du temps, selon l'équation

$$\log\left(r_t^{\mathbf{y}}\right) = \log\left(r_{t-1}^{\mathbf{y}}\right) + \varepsilon_{t-1} \tag{5}$$

avec

$$\varepsilon_{t-1} \sim \mathcal{N}(0,\sigma_\varepsilon^2) \tag{6}$$

Le but est de rendre la variance dépendante du temps afin de permettre au modèle de s'adapter aux différentes situations comme la stationarité ou l'évolution rapide [1] de la fréquence ou de l'amplitude. Pour cela, un choix simple et performant est encore une marche aléatoire comme équation dynamique [2]. Ce modèle d'évolution porte sur le logarithme de la variance afin de garantir la positivité de $r_t^{\mathbf{y}}$. Enfin, la valeur de σ_ε est fixée, sans nécessiter de réglage fin.

La vraisemblance est obtenue à partir de l'équation (4) et est donnée par :

$$p\big(\mathbf{y}_t|(k_t,\boldsymbol{\theta}_t)\big) = \left(2\pi r_t^{\mathbf{y}}\right)^{-\frac{L_{\mathbf{w}}}{2}} \exp\left(-\frac{\left\|\mathbf{y}_t - g_t(k_t,\boldsymbol{\theta}_t)\right\|^2}{2r_t^{\mathbf{y}}}\right) \tag{7}$$

2.2 Equation de transition

Pour des raisons de généralité, on n'impose pas de modèle dynamique particulier pour gérer l'évolution des paramètres inconnus de l'instant $t-1$ à l'instant t. Par exemple, de tels modèles sont difficiles à définir quand on traite des ambiances sonores car ces signaux sont souvent composites (parole, bruit environnemental, drone urbain, etc ...). De plus, notre modèle suppose la stationarité des paramètres par rapport à la longueur de la fenêtre. C'est pourquoi, à l'instant t, les équations de transition sont définies par des marches aléatoires centrées sur l'estimation du paramètre à l'instant précédent. Pour $j = 1,\ldots,k_t$:

$$f_{t,j} = f_{t-1,j} + v_{t-1,j}^{\mathbf{f}} \tag{8}$$

$$a_{t,j} = a_{t-1,j} + v_{t-1,j}^{\mathbf{a}} \tag{9}$$

$$b_{t,j} = b_{t-1,j} + v_{t-1,j}^{\mathbf{b}} \tag{10}$$

1. La notion de rapidité d'évolution des paramètres du signal est à comparer à la longueur de la fenêtre.

	$k_{t-1}=k_{min}$	$k_{min}<k_{t-1}<k_{max}$	$k_{t-1}=k_{max}$
b_t	1/10	1/10	0
e_t	9/10	8/10	9/10
d_t	0	1/10	1/10

TAB. 1: *Probabilités d'évolution du nombre k_t de composantes. Les probabilités b_t, e_t et d_t sont définies par l'équation* (13)*.*

où $v_{t-1,j}^{(.)}$ est un bruit blanc gaussien de moyenne nulle et de variance $r_{t-1,j}^{(.)}$. De la même manière que pour les observations, on autorise les variances $r_{t,j}^{\mathbf{f}}$, $r_{t,j}^{\mathbf{a}}$ et $r_{t,j}^{\mathbf{b}}$ à évoluer au cours du temps. Par exemple, pour $r_{t,j}^{\mathbf{f}}$, $j = 1,\ldots,k_t$:

$$\log\left(r_{t,j}^{\mathbf{f}}\right) = \log\left(r_{t-1,j}^{\mathbf{f}}\right) + \varphi_{t-1,j} \tag{11}$$

avec

$$\varphi_{t-1,j} \sim \mathcal{N}(0,\sigma_\varphi^2) \tag{12}$$

Le modèle contient aussi le nombre k_t de composantes. Ce paramètre étant discret, on adopte une marche aléatoire discrete et l'équation de transition est donnée par :

$$k_t = k_{t-1} + \begin{cases} 1 & \text{avec une probabilité } b_t \\ 0 & \text{avec une probabilité } e_t \\ \text{-1} & \text{avec une probabilité } d_t \end{cases} \tag{13}$$

où les probabilités b_t, e_t et d_t sont données dans le tableau 1. La valeur exacte des probabilités a peu d'importance. En effet, le but est d'autoriser k_t à augmenter ou à diminuer tout en gardant prépondérante la possibilité de le laisser constant. Nous posons ainsi l'heuristique que le nombre de composantes évolue à une échelle temporelle supérieure à l'intervalle de temps entre $t-1$ et t. Enfin, on impose à k_t d'être compris entre k_{min} et k_{max}.

3 Algorithme d'estimation séquentielle

L'idée de base des méthodes MCS est l'utilisation d'une grille aléatoire et adaptative pour calculer une approximation de la densité *a posteriori* de l'état [4]. Plus précisement, la densité *a posteriori* $p(k_{0:t},\boldsymbol{\theta}_{0:t}|\mathbf{y}_{1:t})$ est définie, à une constante près, comme le produit de la vraisemblance et de la densité *a priori*. Elle peut se factoriser comme suit :

$$p\big(k_{0:t},\mathbf{f}_{0:t},\mathbf{a}_{0:t},\mathbf{b}_{0:t}|\mathbf{y}_{1:t}\big) = p\big(\mathbf{a}_{0:t},\mathbf{b}_{0:t}|\mathbf{y}_{1:t},k_{0:t},\mathbf{f}_{0:t}\big)p\big(k_{0:t},\mathbf{f}_{0:t}|\mathbf{y}_{1:t}\big) \tag{14}$$

Le système de Markov à sauts adopté peut s'écrire :

$$k_t \sim P(k_t|k_{t-1}) \tag{15}$$

$$\mathbf{f}_t \sim p(\mathbf{f}_t|\mathbf{f}_{t-1}) \tag{16}$$

$$\begin{bmatrix}\mathbf{a}_t\\ \mathbf{b}_t\end{bmatrix} = A(k_t,\mathbf{f}_t)\begin{bmatrix}\mathbf{a}_{t-1}\\ \mathbf{b}_{t-1}\end{bmatrix} + B(k_t,\mathbf{f}_t)\mathbf{v}'_{t-1} \tag{17}$$

$$\mathbf{y}_t = C(k_t,\mathbf{f}_t)\begin{bmatrix}\mathbf{a}_t\\ \mathbf{b}_t\end{bmatrix} + D(k_t,\mathbf{f}_t)\mathbf{v}''_{t-1} \tag{18}$$

avec $\mathbf{v}'_{t-1}$ et $\mathbf{v}''_{t-1}$ des bruits blancs gaussiens de moyenne nulle et de variance unité. Les densités $P(k_t|k_{t-1})$ et $p(\mathbf{f}_t|\mathbf{f}_{t-1})$ sont données par les équations (13) et (8) respectivement et les matrices A, B, C et D sont connues (voir annexe A). Dans notre cas, il apparait que la densité $p\big(\mathbf{a}_{0:t},\mathbf{b}_{0:t}|\mathbf{y}_{1:t},k_{0:t},\mathbf{f}_{0:t}\big)$, dans l'équation (14), est une gaussienne. Ses deux premiers moments peuvent donc être calculés analytiquement, une fois la

A l'instant $t = 0$
  Pour $i = 1 \dots N$
    Initialisation
      Générer $\left(k_0^{(i)}, \boldsymbol{\theta}_0^{(i)}\right)$ selon $p(k_0, \boldsymbol{\theta}_0)$
      Poser $\bar{\boldsymbol{\theta}}_0^{(i)} = \mathbb{E}[\boldsymbol{\theta}_0^{(i)}]$
A l'instant $t \geq 1$
  Pour $i = 1 \dots N$
    Mise à jour des particules
      Générer $k_t^{(i)} \sim P(k_t|k_{t-1})$
      Générer $\mathbf{f}_t^{(i)} \sim q_t(\mathbf{f}_t^{(i)}|\mathbf{f}_{t-1}^{(i)})$
      Calculer $\boldsymbol{\mu}_t^{(i)}$ et $\boldsymbol{\Sigma}_t^{(i)}$ à partir de $\mathbf{y}_t$, $k_t^{(i)}$, $\mathbf{f}_t^{(i)}$, $\boldsymbol{\mu}_{t-1}^{(i)}$ et $\boldsymbol{\Sigma}_{t-1}^{(i)}$
    Mise à jour des hyperparamètres c'est-à-dire des variances qui peuvent évoluer comme les paramètres.
    Calcul des poids
  Normalisation : les poids sont normalisés afin d'avoir une somme unitaire.
  Etape de sélection : multiplier ou supprimer les particules en fonction de leur poids.
  Sortie : à partir de l'approximation particulaire de la densité *a posteriori* ainsi obtenue, calculer la quantité qui nous intéresse, c'est-à-dire, l'espérance de la marginale de cette densité : $\mathbb{E}[h_t(\boldsymbol{\theta}_{0:t})] \approx \frac{1}{N}\sum_{i=1}^{N} h_t(\boldsymbol{\theta}_{0:t}^{(i)})$ avec $h_t(\boldsymbol{\theta}_{0:t}) = \boldsymbol{\theta}_t$.

TAB. 2: *Grandes lignes de l'algorithme de FPRB implémenté pour notre méthode, à l'instant t.*

marginale $p\left(k_{0:t}, \mathbf{f}_{0:t}|\mathbf{y}_{1:t}\right)$ estimée. Ainsi, notre problème d'estimation séquentielle est traité en combinant le FP pour estimer la distribution de k_t et de $\mathbf{f}_t$ avec une banque de FK pour calculer les moments des distributions des amplitudes correspondants aux fréquences [5]. A chaque instant t, le Filtre Particulaire Rao-Blackwellisé (FPRB) produit un ensemble de N particules pondérées $\left\{\left(k_{0:t}^{(i)}, \mathbf{f}_{0:t}^{(i)}\right), \omega_t^{(i)}\right\}$ où $\omega_t^{(i)}$ est le poids, donnant une approximation particulaire de la marginale :

$$\hat{p}(k_{0:t}, \mathbf{f}_{0:t}|\mathbf{y}_{1:t}) = \sum_{i=1}^{N} \omega_t^{(i)} \delta_{\left(k_{0:t}^{(i)}, \mathbf{f}_{0:t}^{(i)}\right)}(k_{0:t}, \mathbf{f}_{0:t})$$

En supposant qu'un tel ensemble est disponible à l'instant $t-1$, l'algorithme de FPRB met à jour cet ensemble de l'instant $t-1$ à l'instant t, comme suit. Dans un premier temps, chaque particule $k_{0:t-1}^{(i)}, \mathbf{f}_{0:t-1}^{(i)}$ est étendue avec un nouvel état $k_t^{(i)}, \mathbf{f}_t^{(i)}$ généré par $q_t\left(k_t, \mathbf{f}_t|k_{t-1}^{(i)}, \mathbf{f}_{t-1}^{(i)}\right)$, la densité d'importance. Ensuite, la moyenne $\boldsymbol{\mu}_t^{(i)}$ et la covariance $\boldsymbol{\Sigma}_t^{(i)}$ de $[\mathbf{a}_{0:t}\ \mathbf{b}_{0:t}]^{\mathrm{T}}$ sont propagées par les équations du FK :

$$\begin{aligned}
\boldsymbol{\mu}_{t|t-1}^{(i)} &= A\boldsymbol{\mu}_{t-1}^{(i)} \\
\boldsymbol{\Sigma}_{t|t-1}^{(i)} &= A\boldsymbol{\Sigma}_{t-1}^{(i)}A^{\mathrm{T}} + BB^{\mathrm{T}} \\
S_t^{(i)} &= C\boldsymbol{\Sigma}_{t|t-1}^{(i)}C^{\mathrm{T}} + DD^{\mathrm{T}} \\
y_{t|t-1}^{(i)} &= C\boldsymbol{\mu}_{t|t-1}^{(i)} \\
\boldsymbol{\mu}_t^{(i)} &= \boldsymbol{\mu}_{t|t-1}^{(i)} + \boldsymbol{\Sigma}_{t|t-1}^{(i)}C^{\mathrm{T}}S_t^{-1(i)}\left(y_t - y_{t|t-1}^{(i)}\right) \\
\boldsymbol{\Sigma}_t^{(i)} &= \boldsymbol{\Sigma}_{t|t-1}^{(i)} - \boldsymbol{\Sigma}_{t|t-1}^{(i)}C^{\mathrm{T}}S_t^{-1(i)}C\boldsymbol{\Sigma}_{t|t-1}^{(i)}
\end{aligned}$$

Enfin, les poids sont mis à jour :

$$\omega_t^{(i)} \propto p\left(\mathbf{y}_t|\mathbf{y}_{1:t-1}, k_{0:t-1}^{(i)}, \mathbf{f}_{0:t-1}^{(i)}\right) \frac{p\left(k_t^{(i)}, \mathbf{f}_t^{(i)}|k_{t-1}^{(i)}, \mathbf{f}_{t-1}^{(i)}\right)}{q_t\left(k_t^{(i)}, \mathbf{f}_t^{(i)}|k_{t-1}^{(i)}, \mathbf{f}_{t-1}^{(i)}\right)}$$

où $\propto$ signifie "proportionnel à". On peut noter que le terme $p\left(\mathbf{y}_t|\mathbf{y}_{1:t-1}, k_{0:t-1}^{(i)}, \mathbf{f}_{0:t-1}^{(i)}\right)$ ne peut pas se mettre sous la forme

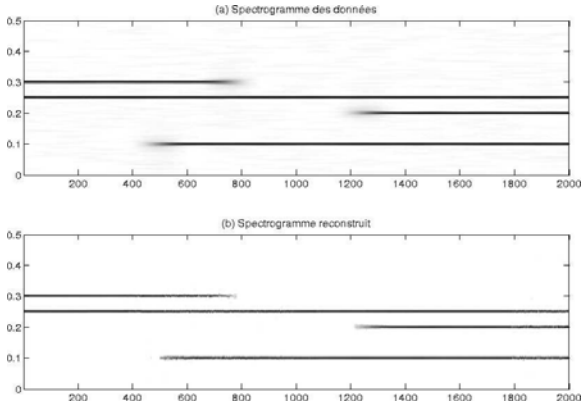


FIG. 1: *Signaux à composantes stationnaires. (a) Spectrogramme des données. La colonne de ce spectrogramme, à l'instant t, est le vector d'observation $\mathbf{y}_t$. (b) Spectrogramme reconstruit. La colonne de ce spectrogramme, à l'instant t, est donnée par $\sum_{i=1}^{N} g_t(k_t^{(i)}, \boldsymbol{\theta}_t^{(i)})$.*

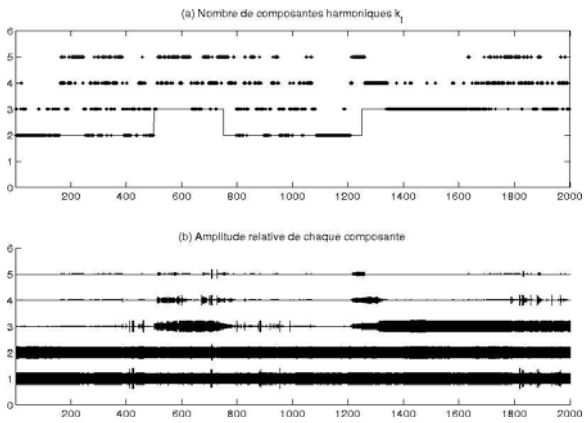


FIG. 2: *Signaux à composantes stationnaires. (a) Nombre de composantes harmoniques. Le trait plein correspond à la valeur simulée, les pointillés à la valeur estimée. (b) Amplitude relative de chaque composante.*

$p\left(\mathbf{y}_t|k_t^{(i)}, \mathbf{f}_t^{(i)}\right)$ à cause de la dépendance aux valeurs passées via $\mathbf{a}_{0:t}$ et $\mathbf{b}_{0:t}$. Il se calcule néanmoins par l'équation (7).

Les grandes lignes de l'algorithme de filtrage particulaire Rao-Blackwellisé (FPRB) mis en œuvre pour notre méthode sont données dans le tableau 2.

4 Résultats

Nous avions obtenu de bons résultats avec notre précédent modèle [1]. Il s'agit donc maintenant de faire une étude comparative afin de mettre en évidence les apports de ce nouveau modèle. Pour cela, nous avons appliqué l'algorithme décrit dans le tableau 2 à des signaux synthétiques pour pouvoir comparer les paramètres estimés aux paramètres simulés. L'étude porte sur deux catégories de signaux. La première, dont les résultats sont présentés dans les figures 1 et 2, concerne des signaux dont les composantes sont stationnaires c'est-à-dire que la fréquence et l'amplitude de chaque composante est constante au cours du temps, seul le nombre de composantes k_t change. Pour la deuxième catégorie, figures 3 et 4, une des composantes

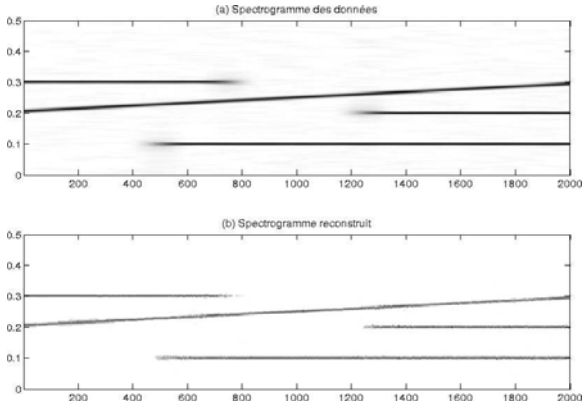


FIG. 3: *Signaux avec une composante non stationnaire. (a) Spectrogramme des données. La colonne de ce spectrogramme, à l'instant t, est le vector d'observation $\mathbf{y}_t$. (b) Spectrogramme reconstruit. La colonne de ce spectrogramme, à l'instant t, est donnée par $\sum_{i=1}^{N} g_t(k_t^{(i)}, \boldsymbol{\theta}_t^{(i)})$.*

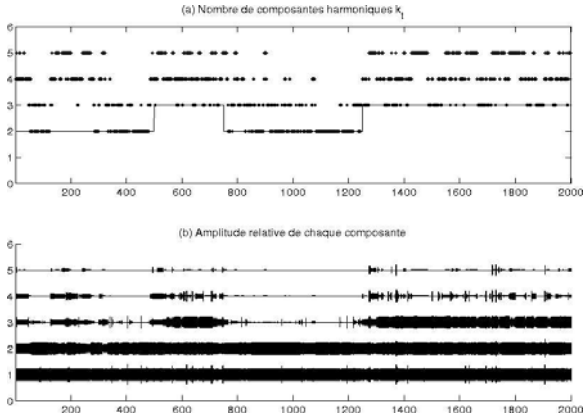


FIG. 4: *Signaux avec une composante non stationnaire. (a) Nombre de composantes harmoniques. Le trait plein correspond à la valeur simulée, les pointillés à la valeur estimée. (b) Amplitude relative de chaque composante.*

est non stationnaire : sa fréquence varie au cours du temps. Dans les deux cas, les signaux, de longueur 2000, sont noyés dans un bruit blanc additif avec un rapport signal sur bruit de $RSB = 21\text{dB}$ et les paramètres de l'algorithme sont les mêmes : $N = 500$, $k_{min} = 1$, $k_{max} = 5$ et la fenêtre est une fenêtre de Hamming de longueur $L_{\mathbf{w}} = 255$.

Les figures 1 et 3 montrent que les fréquences et les amplitudes de chaque composante sont estimées avec une bonne précision et que le bruit présent dans les données ne perturbe pas leur estimation. Sur ce point, les résultats de [1] sont améliorés. Par contre, les figures 2(a) et 4(a) mettent en évidence que l'estimation du nombre de composantes est moins bonne et les figures 2(b) et 4(b) montrent que ceci est d'autant plus vrai quand le signal contient une composante non stationnaire. Ceci peut s'expliquer par le fait que quand une fréquence évolue au cours du temps, cela se ressent aussi sur la phase, phénomène qui était transparent dans [1]. Elles montrent aussi qu'une étape de post-traitement, basée par exemple sur un seuil pour les amplitudes, permettrait une meilleure estimation du nombre de composantes.

5 Conclusion et perspectives

La détection et le suivi de plusieurs fréquences instantanées présentes simultanément dans un signal peut être effectuée directement dans le domaine temporel. Ceci nécessite la prise en compte dans le modèle de la phase initiale sans pour autant augmenter le nombre de paramètres à estimer. En effet, le nouveau modèle étant linéaire gaussien conditionnellement à une partie de l'état, on peut ajouter une étape de Rao-Blackwellisation à l'algorithme de filtrage particulaire utilisé pour réaliser l'estimation. Cette méthode présente des avantages par rapport à notre précédent travail, notamment pour l'estimation des fréquences et des amplitudes. L'estimation du nombre de composante reste, elle, moins précise, notamment dans le cas de composantes non stationnaires. Une étape de post-traitement peut néanmoins être envisagée afin d'améliorer cette estimation. Une autre solution pourrait être de favoriser la diminution du nombre de composantes en augmentant la probabilité correspondante.

A Définition des matrices du modèle de Markov à sauts

Les matrices A et B, de taille $2k_t \times 2k_t$, sont définies par l'équation (17), qui reprend les équations (9) et (10) :

$$A = \mathbf{I}_{2k_t \times 2k_t} \quad B = \text{diag}\left(\left[\sqrt{r_{t-1,j}^{\mathbf{a}}} \cdots \sqrt{r_{t-1,j}^{\mathbf{b}}}\right]\right)$$

L'équation (18) reprend l'équation d'observation (4) et permet de définir la matrice C, de taille $L_{\mathbf{w}} \times 2k_t$, et la matrice D :

$$C = \left[\left(C_j^{(c)}\right) \cdots \left(C_j^{(s)}\right)\right] \quad D = \sqrt{r_{t-1}^{\mathbf{y}}}\mathbf{I}_{L_{\mathbf{w}} \times L_{\mathbf{w}}}$$

avec

$$C_j^{(c)} = \begin{bmatrix} \cos(2\pi f_{t,j} 1)\mathbf{w}[1] \\ \vdots \\ \cos(2\pi f_{t,j} \tau)\mathbf{w}[\tau] \\ \vdots \\ \cos(2\pi f_{t,j} L_{\mathbf{w}})\mathbf{w}[L_{\mathbf{w}}] \end{bmatrix} C_j^{(s)} = \begin{bmatrix} \sin(2\pi f_{t,j} 1)\mathbf{w}[1] \\ \vdots \\ \sin(2\pi f_{t,j} \tau)\mathbf{w}[\tau] \\ \vdots \\ \sin(2\pi f_{t,j} L_{\mathbf{w}})\mathbf{w}[L_{\mathbf{w}}] \end{bmatrix}$$

Références

[1] C. Dubois, M. Davy, and J. Idier, "Tracking of Time-Frequency Components using Particle Filtering," in *Proc. IEEE ICASSP*, Philadelphia, PA, USA, March 2005.

[2] C. Andrieu, M. Davy, and A. Doucet, "Efficient Particle Filtering for Jump Markov Systems. Application to Time-Varying Autoregressions," *IEEE trans. on Signal Processing*, July 2003.

[3] Z. Chen, "Bayesian Filtering: From Kalman Filters to Particle Filters, and Beyond," *On line: http://soma.crl.mcmaster.ca/∼zhechen/homepage.htm*, 2003.

[4] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Springer-Verlag, 2001.

[5] N. de Freitas, "Rao-Blackwellised Particle Filtering for Fault Diagnosis," in *Proc. IEEE Aerospace Conference*, Big Sky, MT, USA, March 2002, vol. 4, pp. 1767–1772.

Applications musicales du traitement de signal : synthèse et prospective

H. VINET[1]

[1]STMS, Ircam-CNRS, 1, place Igor Stravinsky 75004 Paris

vinet@ircam.fr

Résumé – L'objet de cette communication est de proposer une synthèse des applications musicales du traitement de signal, des problématiques de recherche qui leur sont liées et des directions prospectives qui se dégagent sur la base de travaux récents dans ce domaine. Après l'exposé de notions préliminaires, relatives au système technique musical et à l'analyse des différentes représentations numériques des informations musicales, cette synthèse se concentre sur trois types de fonctions principales : la synthèse et le traitement des sons musicaux, la spatialisation sonore et les technologies d'indexation et d'accès.

Abstract *–This communication aims at providing a synthesis of the musical applications of digital signal processing, of related research issues, and of future directions that emerge from recent works in that field. After introducing preliminary notions related to the music technical system and to the analysis of different types of digital representations of music information, it focuses on three main function types : audio synthesis and processing, sound spatialization and audio indexing and access technologies.*

1. Introduction

Les applications musicales du traitement de signal ont longtemps été le domaine réservé de centres spécialisés, notamment en lien avec la création contemporaine (synthèse, transformation des sons) et les télécommunications (codage). Elles ont connu au cours des 20 dernières années deux importantes phases successives d'extension, la première liée à la généralisation des techniques de production et de diffusion audionumérique, la seconde, en cours, à la diffusion d'enregistrements par les réseaux et aux méthodes d'indexation et de recherche dans les bases de données musicales. L'objet de cette communication est de proposer une synthèse de l'ensemble des méthodes et problématiques existantes et de présenter les directions prospectives qui se dégagent sur la base de recherches en cours, notamment en lien avec les travaux de l'Ircam. Après l'exposé de notions préliminaires, elle se concentre sur trois types de fonctions principales : la synthèse et le traitement sonores, la spatialisation et les technologies d'indexation et d'accès.

2. Notions préliminaires

2.1 Le système technique musical

Toute activité musicale peut se décomposer en trois phases de production, transmission et réception des artefacts sonores (Figure 1), la transmission étant effectuée directement dans le cas du concert, et dans les autres cas par l'intermédiaire de supports codant les informations musicales et nécessitant à la réception une étape d'accès (sélection et obtention des éléments) et de reproduction sonore. Les supports ainsi produits peuvent être réinvestis pour de nouvelles étapes de production, liées par exemple à la diffusion des contenus selon de nouveaux formats (*re-mastering*), moyennant éventuellement certaines étapes de traitement (restauration d'archives sonores par exemple).

La phase de production réside dans l'utilisation d'instruments et appareils issus d'une activité de lutherie ou

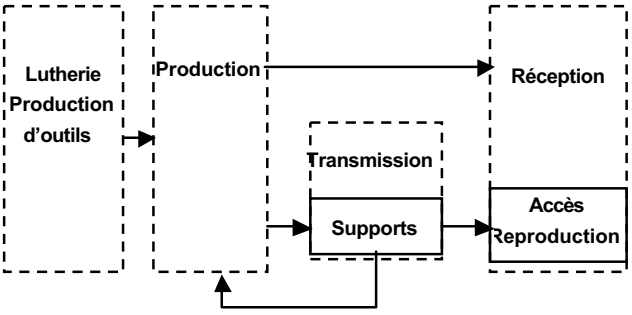


FIG. 1 : Phases du système technique musical

production d'outils, qui est ici mentionnée car les environnements informatiques de production musicale donnent désormais accès, comme nous le verrons plus loin, à de telles possibilités à travers des fonctions de programmation. Elle se décompose elle-même en trois phases distinctes (Figure 2), relatives respectivement à la production des matériaux musicaux isolés (signaux issus de sons instrumentaux ou synthétiques, éventuellement traités, éléments de partitions, etc.), à leur composition, à la fois en temps (montage) et en superposition (mixage, effets) en un contenu composite, puis au codage de ce produit musical au format de diffusion (*mastering* CD et DVD par exemple).

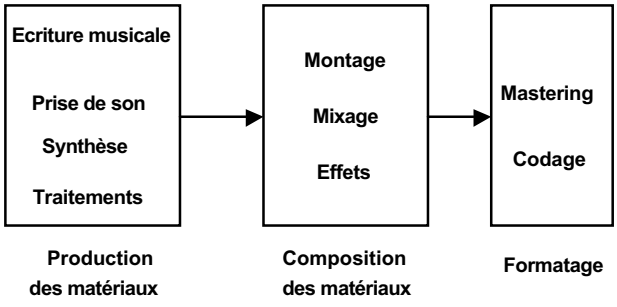


FIG. 2 : Phases de production musicale

Notre étude des différentes fonctions techniques musicales comme applications du traitement de signal se concentrera dans la suite sur les étapes de lutherie, de production de matériaux et de dispositifs d'accès et de reproduction, et ne traitera pas des aspects de codage des informations à travers les différents formats de compression et de diffusion.

2.2 Représentations numériques des informations musicales

Même si les signaux audionumériques constituent l'une des principales formes de diffusion des informations musicales, il convient de les situer du point de vue des différentes représentations existantes des phénomènes musicaux et dans leurs interrelations avec celles-ci. J'ai proposé récemment une typologie des représentations numériques de la musique telles qu'elles apparaissent dans les différentes applications, en dégageant quatre principaux types et en montrant qu'ils s'organisent, dans un sens d'abstraction croissant, selon les niveaux « Physique », « Signal », « Symbolique » et « Cognitif » (cf Figure 3)[26]. Le niveau symbolique décrit les contenus musicaux sous forme de structures à un niveau de discrétisation relevant de la théorie musicale : hauteurs (gamme tempérée), intensités (nuances), rythmes (structures temporelles définies comme multiples et sous-multiples d'une pulsation de base). Le niveau physique rend compte de modélisations acoustiques des phénomènes sonores, en particulier pour la synthèse et le rendu spatial. Le niveau cognitif code des informations de plus haut niveau que le niveau symbolique et inclut tous types de connaissances et descripteurs, notamment qualitatifs, des contenus musicaux. Il intervient notamment sous la forme de métadonnées dans les applications de bases de données musicales dotées de fonctions de recherche par contenu. Quant aux représentations de type signal, elles concernent tout autant les signaux audionumériques sous différentes formes (signaux multicanaux, compressés, etc.) que les signaux de contrôle, issus par exemple de captations gestuelles, selon des variations plus lentes. Je reviens ici sur mes définitions initiales dans lesquelles les signaux de contrôle étaient liés au niveau symbolique pour des raisons d'homogénéité quantitative, car les méthodes de traitement de ces informations s'avèrent davantage relever de techniques de signal à des taux d'échantillonnages faibles (typiquement 30 Hz à 1 kHz) que de méthodes symboliques.

Il est aisé de montrer que ces niveaux sont ordonnés par quantité d'information décroissante, les informations contenues dans les représentations de plus haut niveau étant associées à des connaissances implicites (culturelles par exemple) extérieures aux données de représentation. Les processus d'analyse et de synthèse constituent les opérateurs génériques de conversion entre niveaux, et prennent de formes spécifiques en ce qui concerne les représentations musicales. Par exemple, la réduction d'information opérée par l'analyse concerne notamment, dans le passage du physique au signal, une réduction spatiale, équivalent de la *prise de son*, consistant à passer d'une pression acoustique fonction de l'espace et du temps à des signaux d'une variable temporelle. Le passage du signal au symbolique concerne une double discrétisation, à la fois des occurrences temporelles et des valeurs prises par les grandeurs concernées (passage du continuum des fréquences et énergies à des échelles discrètes de hauteurs et de nuances).

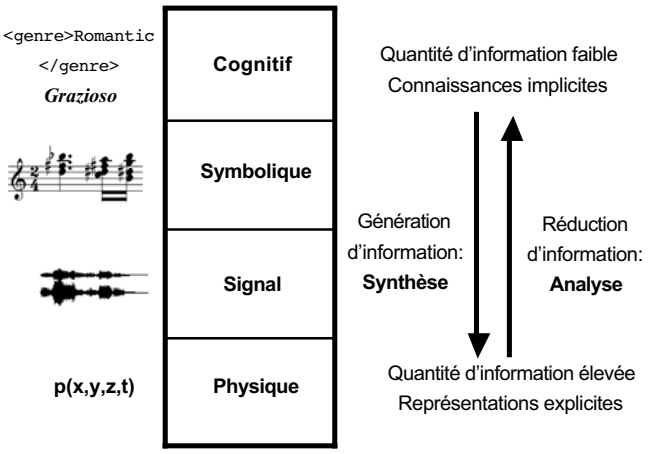


FIG. 3 : Les niveaux de représentations musicales

3. Synthèse et traitement sonores

Les fonctions de synthèse et traitement sonore, relevant de la phase « Production des matériaux » de la Figure 2, constituent les applications musicales les plus importantes du traitement de signal. Après une revue des modèles les plus utilisés, j'aborderai les questions d'architecture du point de vue des possibilités de programmation et des structures de traitement et de contrôle, en montrant comment les évolutions de celles-ci sont susceptibles de contribuer à renouveler les fonctions des applications musicales.

3.1 Modèles de synthèse et de traitement

3.1.1 Modèles de synthèse

Les premières synthèses audionumériques ont été réalisées par l'équipe de Max Mathews aux Bell Labs dans les années 50 et reposaient sur la combinaison d'oscillateurs périodiques, produisant des sons très typés. Dans les années 80, la mise au point de techniques de synthèse par modulation de fréquence à l'Université de Stanford, permettant la production de structures spectrales riches et variées avec des ressources de calcul limitées, a permis l'essor des synthétiseurs de la famille DX7 de Yamaha. Les techniques utilisées dans les synthétiseurs se sont par ailleurs concentrées sur l'échantillonnage de sons enregistrés dotés de fonctions simples de post-traitement (variation de taux d'échantillonnage pour la transposition, enveloppes en amplitude, etc.), avec comme métaphore de contrôle celle du clavier à travers la norme MIDI. Dans un contexte plus expérimental, la mise au point par Xavier Rodet à l'Ircam des FOF (fonctions d'ondes formantiques), modèle temporel de synthèse source-filtre adapté aux formants vocaux a rendu possibles les premières synthèses de voix chantée [16].

Les méthodes de *synthèse par modélisation physique*, fondées sur la modélisation acoustique des sources sonores, se sont développées plus récemment. Bien que nécessitant des ressources de calcul plus importantes, elles présentent de nombreux avantages par rapport aux modèles de signaux :

production de sons plus riches et proches des sons réels, paramètres physiques plus signifiants, reproduction de comportements non-linéaires caractéristiques de certains instruments (transitoires, bifurcations et régimes chaotiques). Il existe plusieurs approches adaptées à la modélisation de différentes classes d'instruments :

- les *guides d'ondes*, principalement développés à l'Université de Stanford et intégrées au synthétiseur VL1 de Yamaha [24], peu coûteuses en calcul mais limitées à certaines structures algorithmiques,

- la *synthèse modale*, formalisme utilisé dans le logiciel Modalys de l'Ircam [10], permettant la modélisation d'un grand nombre de classes d'objets de base (tubes, plaques, cordes, membranes,…) et leur assemblage, par l'intermédiaire d'interactions non-linéaires, pour « construire » des instruments virtuels sans limite de complexité. L'intérêt de la synthèse modale est aussi que les données modales d'un objet peuvent soit résulter d'une résolution analytique dans le cas de structures simples, soit de mesures d'analyse modale pour des corps réels plus complexes.

Au-delà de problèmes de modélisation, l'une des principales difficultés inhérentes aux modèles physiques est celle de leur contrôle : certains modèles se présentent comme des systèmes dynamiques non-linéaires et leur contrôle pour obtenir les sons voulus nécessite l'adjonction d'une interface supplémentaire, relevant de l'expertise de l'interprète, effectuant la variation conjointe des différents paramètres à partir de données d'entrée plus musicales : hauteur à produire, mode de jeu choisi, etc. Plusieurs approches peuvent être menées à cet effet à partir de signaux produits par des instruments réels : problèmes inverses dans le cas de modèles analytiques [8], apprentissage. Cette problématique, encore relativement peu abordée dans les applications existantes, a récemment débouché, dans le cadre du projet RIAM Windset, sur un ensemble de modules commerciaux de synthèse d'instruments à vent dotés de contrôles musicaux[1]

3.1.2 Modèles pour le traitement musical

Il existe une grande variété de fonctions de traitement sonore dans les applications existantes, allant du filtrage au débruitage en passant par différents types d'effets. Nous nous concentrerons ici sur les traitements plus spécifiquement musicaux : transposition, changement de la durée sans changement de hauteur (time-stretching), morphing et hybridation entre sons, etc. Plusieurs critères doivent être pris en compte pour l'évaluation des différents modèles : les classes de sons reproduits, l'espace de variations possibles, la présence ou non d'artefacts à l'écoute, la fidélité de simulation de fonctions de traitements par rapport à leurs référents acoustiques, le coût de calcul. A titre d'illustration des principaux modèles utilisés, on peut citer :

- le *vocodeur de phase*, consistant en une répartition du contenu spectral par bandes de fréquences, pouvant être traitées indépendamment. Cette approche, adaptée à de larges classes de sons, est aujourd'hui généralement modélisée par une analyse/synthèse par TFCT et permet des traitements tels que filtrage, time-stretching et transposition avec une bonne qualité sonore. Les avancées récentes sur ce modèle concernent notamment la reconstitution des phases pour le time stretching [7], ainsi que la prise en compte des transitoires, un rendu plus naturel du time stretching étant obtenu par un traitement limité aux portions quasi-stationnaires des sons [19]. Le logiciel Audiosculpt de l'Ircam (Figure 4), reposant sur le noyau d'analyse/ synthèse SuperVP, propose ainsi des fonctions d'édition temps-fréquence des sons à partir d'une analyse de type TFCT, ainsi que la visualisation de différentes analyses superposées au sonagramme : fréquence fondamentale, analyses de partiels et de formants, détection de transitoires, etc.

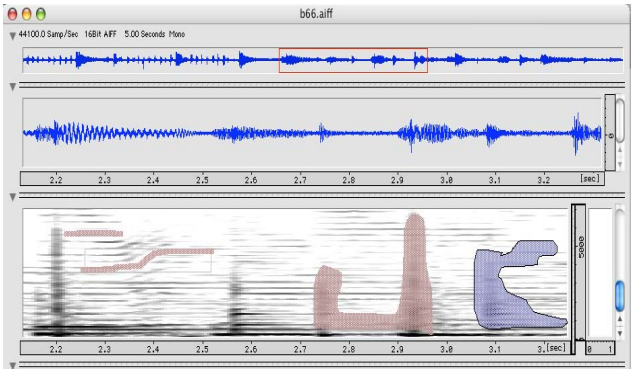


FIG. 4 : Logiciel Audiosculpt de l'Ircam

- le *modèle additif ou sinusoïdal*, consistant à décomposer un son en une somme de sinusoïdes de fréquences et d'amplitudes à variations lentes et d'un signal résiduel de bruit [17]. Ce modèle, plus compact en données d'analyse que le précédent, convient à des classes de sons plus limitées (spectre discret). Des méthodes efficaces de synthèse par TFR inverse existent et les problèmes d'analyse varient selon les types de sons : calcul de fréquence fondamentale pour les sons harmoniques, méthodes de détection de pics et de suivi de partiels pour les sons non-harmoniques [20]. Au-delà de fonctions de transposition et de time-stretching, ou de variation du contenu spectral par modification des amplitudes des partiels, une fonction musicale intéressante de ce modèle est celle de « morphing sonore », par interpolation des fréquences et amplitudes des partiels appariés entre le son de départ et celui de destination.
- les méthodes de type PSOLA (*pitch synchronous overlap add*) sont également utilisées, en particulier associées avec le modèle additif [14], et sont notamment intéressantes du fait de leur faible coût de calcul pour la synthèse par des méthodes temporelles OLA,
- même si leurs applications musicales restent encore à notre connaissance expérimentales, les méthodes de décomposition sur des familles d'atomes de type *Matching Pursuit* présentent un intérêt pour l'analyse des signaux audio, de même que, pour la synthèse, les méthodes de *concaténation d'unités* appliquées aux signaux musicaux [22].

Le format SDIF (*Sound Description Interchange Format*), est devenu un standard de fichiers dans la communauté d'informatique musicale pour représenter sous forme binaire tous types de données d'analyse issues des modèles présentés ci-dessus, ainsi que de nombreuses autres (enveloppes spectrales, diphones, etc.) [23].

[1] www.arturia.com/en/brass/brass.php

3.2 Architectures pour la synthèse et le traitement

3.2.1 Langages musicaux

Si les applications industrielles du traitement de signal font appel à des circuits spécialisés pour la réalisation de fonctions spécifiques, les modes d'implantation des algorithmes dans le domaine de l'informatique musicale suivent généralement une autre logique, consistant à exploiter les ressources de calcul existantes pour une gamme aussi large que possible de fonctions de synthèse et de traitement. Dès les débuts de la synthèse audionumérique, la conception de fonctions de calcul élémentaires a été associée à la réalisation de *langages musicaux*, langages informatiques spécialisés pour la musique, dotés notamment de modèles du temps et permettant de programmer des algorithmes de calcul complexes à partir de ces opérateurs élémentaires. De nombreux langages musicaux ont ainsi été développés dans la communauté d'informatique musicale, notamment MusicV et plus récemment Csound[2], dont les concepts reposent sur l'articulation de deux métaphores musicales, la notion d' « instrument » lié à l'algorithme de synthèse ou de traitement, et celle de « partition », définissant la variation des paramètres d'entrée de l'instrument au cours du temps.

Le lien entre langages musicaux et traitement temps réel a été récemment rendu possible par la diffusion du logiciel Max/MSP[3], qui, à raison de plusieurs milliers de licences vendues par an, est devenu un environnement de référence pour toute la communauté d'informatique musicale Max est un langage de programmation visuelle permettant de réaliser un algorithme de traitement en temps réel sous forme de « patch » à partir de fonctions de base, ou « objets » (cf Figure 5). Une interface de programmation permet à des tierces parties de développer leurs propres objets, et l'un des atouts de Max est qu'il est accompagné de plusieurs milliers d'objets et patchs assurant un ensemble très large de fonctions (synthèse et traitement audio, vidéo, interfaces à des dispositifs d'entrée, à des protocoles réseau, etc.).

Les travaux récents liés à Max visent notamment à doter celui-ci de types de données complexes, permettant, comme dans le cas de la bibliothèque « Gabor » [21], de réaliser des familles d'objets d'analyse/ synthèse de signal en temps réel s'échangeant des structures de données d'analyse indexées sur le temps (listes de partiels, enveloppes spectrales, etc.) et les resynchronisant automatiquement pour la synthèse à travers un mécanisme OLA généralisé. Cette approche repose sur des travaux récents dans cet environnement sur la gestion de la mémoire visant à concilier efficacité pour le traitement en temps réel et allocation dynamique des données Une autre direction d'investigation actuelle consiste à mettre en place des structures d'implantation adaptées pour la synthèse par modélisation physique, permettant notamment d'étendre les fonctions de « lutherie virtuelle » du logiciel Modalys (cf §3.1.1) à la réalisation d'instruments fonctionnant en temps réel.

[2] www.csounds.com
[3] http://www.cycling74.com/products/maxmsp.html

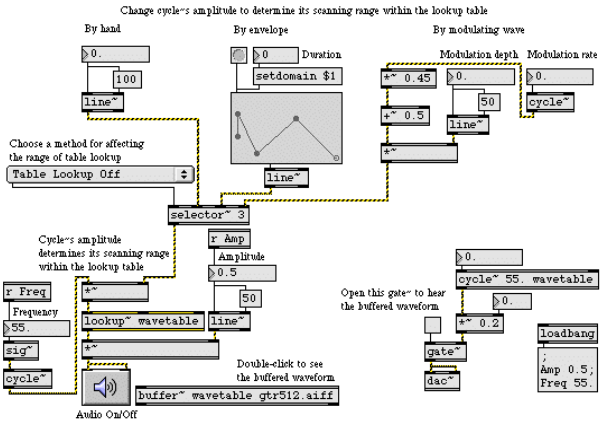


FIG. 5 : Exemple de patch du logiciel Max

3.2.2 Structures de traitement et de contrôle : vers un traitement par le contenu

L'objet de cette section est de reprendre les notions abordées au §3.1 dans une perspective plus large en étudiant les différentes structures de traitement du point de vue de la prise en compte qu'elles opèrent des contenus musicaux, et en particulier des modes de contrôle qu'elles autorisent. Cette analyse se fondera sur les notions de niveaux de représentation des informations musicales définies au §2.2. Le cas de la Figure 6-a, le plus courant, est celui de traitements n'opérant aucune différenciation en fonction des signaux d'entrée. La Figure 6–b illustre la production de certains paramètres de traitement par analyse du signal d'entrée, par exemple le calcul d'une enveloppe spectrale avant transposition de signaux produits par une structure de type source-filtre. La figure 6-c rend compte d'une classe assez large de modèles effectuant une décomposition des signaux d'entrée et opérant des traitements différenciés selon leurs contenus (par exemple time-stretching sur parties stationnaires uniquement et parties transitoires inchangées). La figure 6-d concerne en particulier les structures de traitement abordées au §3.1.2 sous la forme de modèles paramétriques d'analyse/synthèse, les traitements étant effectués par transformations dans l'espace des paramètres. Un cas particulier est celui où les traitements sont contrôlés par des paramètres de niveau symbolique, dont la signification musicale est directement pertinente pour l'utilisateur. Certaines applications commerciales intègrent de telles approches. Ainsi, le logiciel Melodyne permet l'édition musicale de sons monodiques après une pré-analyse dans le domaine symbolique, à travers des fonctions musicales telles que transposition ou calage sur la grille temporelle d'un rythme donné[4]. Le logiciel GarageBand[5] d'Apple permet à des utilisateurs non-musiciens de constituer facilement des séquences musicales à partir de boucles instrumentales préenregistrées, le calage des hauteurs et des instants associés

[4] www.celemony.com/cms/index.php?id=melodyne
[5] www.apple.com/ilife/garageband

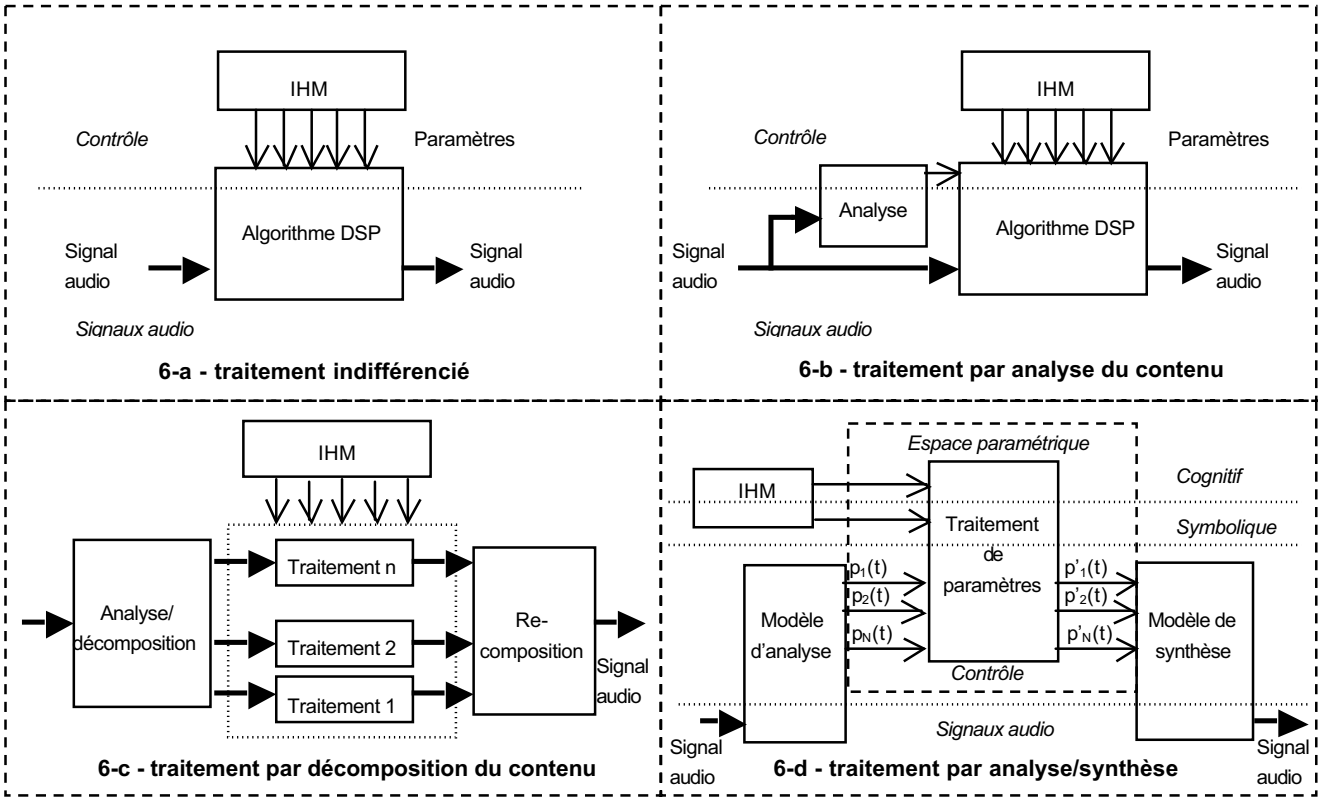


FIG. 6 : Structures de traitement et de contrôle

à chaque note étant effectué automatiquement en fonction du contexte musical où est placé le nouveau son (harmonie, tempo) à partir de formats (AppleLoops) combinant représentations de types signaux et symboliques. La figure 6–d concerne une acception plus générale de la notion de traitement par le contenu, celui-ci étant spécifié non seulement selon ses paramètres symboliques, mais aussi à partir de descripteurs relevant du niveau cognitif, s'attachant à spécifier par exemple des caractéristiques qualitatives du résultat sonore attendu : son « plus métallique », « moins sourd », etc.. Cette problématique relève de celle de la description des contenus sonores abordée au §5.1. Une approche complémentaire consiste à élaborer des *modèles de contrôle* pour l'analyse et la synthèse, décrivant les évolutions des paramètres de contrôle selon le contexte (symbolique et cognitif) de plus haut niveau. Cette visée, encore prospective, relève notamment de recherches sur la modélisation de l'interprétation, comme fonction équivalant à instancier des signaux de contrôle à variation lente, pilotant un modèle de synthèse, à partir d'une prescription de type symbolique+cognitif (partition). Le cas le plus général du traitement par le contenu est celui qui combine les structures des figures 6–c et 6–d, les traitements spécifiques à chaque partie du son pouvant faire l'objet d'un modèle d'analyse/ synthèse adapté à chacune d'elle.

3.2.3 Extensions de la notion d'instrument

La notion traditionnelle d'instrument de musique se trouve bouleversée par l'évolution des techniques de synthèse en temps réel. Une première extension de cette notion d'instruments conduit à les définir comme des systèmes reliant interfaces gestuelles et dispositifs de synthèse sonore en temps réel, selon un mode d'interaction immédiate. Une différence notable entre les instruments électroniques et acoustiques tient en effet dans le découplage que les premies permettent entre l'énergie du geste et celle du son, autorisant tous types de mises en correspondance (*mapping*) entre paramètres issus du geste et paramètres de synthèse. Au-delà de fonctions de correspondance statique reliant p paramètres gestuels à n paramètres de synthèse, des recherches récentes visent également à extraire, de signaux gestuels multidimensionnels issus d'ensembles de capteurs, des formes temporelles de plus haut niveau (modes de jeu, geste dansé) pour piloter des processus musicaux. Cette approche relève de problèmes de modélisation et de traitement de signaux de contrôle. abordés au paragraphe précédent.

4. Spatialisation sonore

Cette acception désigne toutes techniques s'attachant au caractère spatial des sons selon des fonctions de représentation et manipulation, de captation et de reproduction/ simulation/ synthèse des scènes sonores. Je me concentrerai ici sur les modèles de spatialisation et sur les dispositifs de reproduction et de synthèse spatiale, en laissant de côté les problématiques de captation, qui reposent sur des méthodes symétriques à certains égards de celles de liés à la diffusion.

4.1 Modèles pour la spatialisation

Sans parler à ce stade de modèles, les principales formes de codage des informations sonores spatiales dans les applications usuelles reposent sur des systèmes à plusieurs canaux de signaux audionumériques (stéréo, 5.1), captés, transmis et reproduits séparément. Ces procédés, de mise en œuvre relativement aisée, ne permettent cependant la

reproduction et simulation de scènes sonores spatiales que de manière très approximative, en particulier du fait du manque de prise en compte des conditions de restitution (réponse de la salle, directivité des haut-parleurs, etc.). Le format Ambisonic [13], qui propose la modélisation de champs acoustiques sous la forme d'une décomposition en harmoniques sphériques, offre un mode de codage plus précis de scènes tridimensionnelles à partir de signaux multiples. Les modèles permettant de décrire les informations spatiales selon un point de vue plus abstrait s'attachent à caractériser d'une part des sources sonores isolées (position, rayonnement), d'autre part la réponse de l'espace dans lequel elles sont placées. Il existe deux classes principales de modèles de spatialisation, toutes deux normalisées dans MPEG4 : l'approche géométrique et l'approche perceptive. Reposant sur un modèle géométrique de la salle (discrétisation des parois selon des portions de surface de caractéristiques acoustiques homogènes), les modèles géométriques permettent de calculer avec précision la réponse impulsionnelle caractérisant la réponse de la salle sur un couple (source, récepteur). Ils conviennent davantage à des simulations statiques, le cas de sources en déplacement nécessitant de recalculer la réponse impulsionnelle à chaque position. Le modèle perceptif, issu d'études expérimentales sur la qualité acoustique des salles [12], caractérise celles-ci selon un ensemble de paramètres perceptifs liés à la source, à la salle et à leurs interactions. A partir d'un modèle simplifié de réponse impulsionnelle en zones spectro-temporelles, il est possible de produire automatiquement une telle fonction à partir des paramètres perceptifs et donc de simuler la qualité acoustique correspondante [11]. L'avantage de cette approche, qui combine simulation de la localisation des sources et de l'effet de salle, est que la spatialisation peut être spécifiée selon des paramètres de haut niveau, indépendamment du dispositif de reproduction utilisé en s'adaptant automatiquement à celui-ci : cette approche relève des méthodes de traitement par le contenu décrites plus haut.

4.2 Reproduction et synthèse spatiales

Les techniques les plus courantes pour simuler un espace sonore et notamment des sources de position donnée, éventuellement mobiles, reposent sur des systèmes à base de plusieurs haut-parleurs, par pondération des amplitudes des signaux envoyés à chaque source. Cette méthode est notamment à l'œuvre dans les potentiomètres panoramiques des consoles. Cependant, elle est approximative, car les indices d'amplitude ne sont pas les seuls pertinents dans la perception auditive de la localisation, et parce que le rendu de tels effets est fortement dépendant de la position des auditeurs dans l'espace de restitution. La synthèse binaurale (par casque) résulte quant à elle d'études systématiques sur la perception spatiale des sons [2] et se fonde sur des ensembles de réponses impulsionnelles mesurées aux niveaux des deux tympans pour chaque direction d'incidence de son : les HRTF (*Head Related Trasnfer Functions*). Dans le rendu par casque, la convolution du signal monophonique à spatialiser par une paire de HRTF donnée provoque l'impression d'un son venant de la direction d'incidence correspondante. Cependant, chaque individu dispose d'un jeu de HRTF qui lui sont propres et la conception de modèles minimisant les différences interindividuelles constitue une problématique actuelle de recherche, de même que le calcul automatique des HRTF à partir des caractéristiques morphologiques de la tête. Ces disparités peuvent être réduites lorsque le système de restitution par casque est couplé, comme dans le cas du projet européen LISTEN[6] à un suivi qui permet de compenser en temps réel la variation de la position de la tête : l'adaptation du sujet fait qu'il est alors possible de synthétiser des sources dont la localisation reste relativement stable, même s'il a « revêtu » des HRTF qui ne sont pas les siennes.

Un autre système de restitution spatiale prometteur est l'holophonie (analogie acoustique de l'holographie), en particulier selon la méthode de Wavefield synthesis (WFS). Celle-ci, initiée par l'Université de Delft, se fonde sur le théorème de Huygens pour simuler le rayonnement acoustique de sources par l'intermédiaire de transducteurs, jouant le rôle de sources secondaires, entourant l'auditoire, en particulier sous la forme de panneaux multi-actuateurs [1,4]. L'intérêt de cette méthode, qui fait l'objet de recherches actuelles et est vouée à supplanter les systèmes à haut-parleurs pour certaines applications, réside notamment dans le rendu homogène qu'elle produit dans tout l'espace de restitution et dans les possibilités nouvelles de simulation qu'elle ouvre, notamment de sources virtuelles à l'infini (ondes planes) voire même situées dans l'espace de diffusion.

Une approche en quelque sorte duale consiste à réaliser des sources multi-haut-parleurs à rayonnement contrôlé par filtrage des signaux envoyés à chaque haut-parleur. Il est notamment possible d'approximer une figure de directivité donné (celui d'un instrument particulier par exemple) avec une configuration donnée de haut-parleurs. Un tel système a fait l'objet d'expérimentations dans le contexte de la création musicale, mettant en scène ce dispositif excitant la salle de concert de manières différentes selon les signaux envoyés et le contexte musical : diffusion de signaux polyphoniques selon des directivités différentes, variation de la directivité au cours du déroulement du son [29].

5. Indexation audio et technologies d'accès

La généralisation conjuguée de l'Internet et de codecs de compression de signaux audio provoque un bouleversement du système de diffusion de musiques enregistrées. Les problèmes de protection des contenus conjugués aux nouvelles possibilités ouvertes par la diffusion des enregistrements par les réseaux (accès potentiel à des millions de morceaux, systèmes de recommandation personnalisée) ont suscité une forte demande qui a contribué à la structuration, depuis 2000, d'une importante communauté pluridisciplinaire dans le domaine de la recherche d'informations musicales[7]. La problématique centrale liée à ce domaine est celle de la caractérisation des contenus musicaux, à travers la mise en œuvre de métadonnées de description, selon la terminologie de MPEG-7, adaptées à chaque classe d'applications. Après une introduction à cette problématique en lien à celle de l'extraction automatisée des informations, les avancées

[6] http://listen.imk.fraunhofer.de

[7] MIR ou Music information retrieval. Voir : www.ismir.net

récentes dans ce domaines seront illustrées à partir d'exemples issus des projets européens CUIDADO[8] et SemanticHIFI qui se consacrent respectivement à concevoir des bases de données audio reposant sur l'utilisation systématique de descriptions automatiquement extraites des signaux [27], et à préfigurer les chaînes Hi-fi de demain fondées sur des fonctions de manipulation par le contenu des informations musicales [28]

5.1 Description et extraction des contenus musicaux

Il n'existe pas de manière unique de décrire les contenus musicaux, chaque application particulière nécessitant la mise en œuvre de structures de données adaptées. Une grille d'analyse utile des différents types de descriptions consiste à différencier d'une part celles qui sont d'ordre subjectif (et donc variables en fonctions des contextes d'utilisation) et celles d'ordre objectif, ces dernières se divisant en celles qui peuvent être automatiquement calculées à partir des informations musicales (signaux, symboliques) et celles qui nécessitent une intervention humaine. Ainsi, les informations éditoriales (artistes, titre) sont de type objectif-manuel, et les codages d'« empreintes digitales » (identifiants compacts caractérisant un enregistrement), la durée ou l'énergie moyenne du signal sont de type automatique. La constitution de descripteurs subjectifs adaptés, qui relèvent du niveau cognitif dans la terminologie du §2.2, nécessite la mise en œuvre de méthodes d'ingénierie des connaissances (processus *top-down*) et leur obtention automatisée passe par leur mise en relation, généralement par apprentissage, avec des paramètres de bas niveau automatiquement extraits des signaux (processus *bottom-up*). Ainsi, les réalisations opérationnelles dans ce domaine se fondent généralement sur des compromis entre structuration des connaissances relevant de la cognition musicale et état de l'art en matière d'analyse automatisée. Il reste encore beaucoup à faire en analyse de signal pour arriver à des descriptions relativement triviales du point de vue cognitif, si l'on considère par exemple l'état de l'art en analyse de fréquence fondamentale, qui progresse encore dans le cas de sons monodiques [5] et produit de premiers résultats probants dans des cas polyphoniques [30].

5.2 Navigation inter-documents

Ce type de fonctions, qui vise la *navigation par le contenu musical* dans les bases de données d'enregistrements, fait appel à la constitution de descripteurs caractérisant globalement le contenu d'un morceau donné selon différents critères. Un certain nombre de descripteurs pertinents peuvent être extraits des signaux : tempo, intensité, « timbre » orchestral, présence de voix, etc. Certaines approches, relevant notamment de la classification automatique, visent également à apprendre les caractéristiques de catégories définies par les utilisateurs (« rock », « calme », « ambiance»,…) à partir d'ensembles d'exemples [9,31]. Une autre heuristique de navigation, transversale à cette notion de catégories, est celle de recherche par l'exemple, sur la base de mesures de similarités entre sons. A partir d'un son de départ, le système calcule une mesure de similarité avec tous les sons de la base et rend une liste d'items classés par similarité décroissante. De telles mesures peuvent combiner des critères très variés à partir de descripteurs extraits des signaux, selon des poids relatifs préétablis ou configurables par l'utilisateur. Leur intérêt réside notamment dans le caractère souvent imprévu des appariements obtenus, tout en restant cohérent du point de vue des critères retenus pour la similarité.

5.3 Navigation intra-documents

Ce type de fonction, expérimenté dans le cadre du projet SemanticHIFI, vise à fournir des interfaces de représentation et de navigation à l'intérieur du contenu musical d'un morceau, qui dépasse les interfaces traditionnelles des chaînes hi-fi. Les approches poursuivies distinguent d'une part l'analyse de la structure temporelle des morceaux, soit en alignant les enregistrements audio à des représentations symboliques de type MIDI [18] et en autorisant ainsi la navigation dans le fichier son à partir d'une interface symbolique, soit en analysant directement une structure temporelle de haut niveau comme succession et appariement d'état stables du point de vue de critères d'évolution spectrale (introduction, refrain, couplets) [15]. D'autre part, des expérimentations sont menées sur la navigation à l'intérieur de la polyphonie d'un morceau à partir d'interfaces permettant de définir les positions relatives des différentes sources instrumentales ou polyphoniques et de l'auditeur et de produire le rendu spatial correspondant à la situation où l'auditeur se trouverait au milieu de l'orchestre et se rapprocherait de tel ou tel instrument [6]. Une telle fonction impose de disposer, à partir des formats stéréophoniques, de méthodes assurant la séparation automatique des sources, objet de recherches en cours [25], ou directement des enregistrements sous forme multipiste. Ce dernier cas montre l'intérêt d'une extension des formats actuels de distribution musicale à des contenus plus riches (métadonnées, canaux multiples), dont la réalisation technique serait d'ores et déjà envisageable à travers l'ensemble de la chaîne numérique de production, de diffusion et d'accès des enregistrements.

6. Conclusion

Ce tour d'horizon des applications musicales du traitement de signal et des problématiques de recherche qui leur sont liées met en évidence les tendances globales suivantes :

- la généralisation des techniques de traitement de signal, précédemment concentrées sur les outils de production et méthodes de codage, à l'ensemble de la chaîne production/ diffusion/ interfaces d'accès/ reproduction, ces deux derniers maillons étant actuellement l'objet d'importants développements ;
- l'évolution des fonctions de traitement des signaux audionumériques vers des techniques de manipulation par le contenu, en lien avec d'autres types de représentations musicales (physique, symbolique, cognitif), faisant appel à des méthodes d'analyse et de synthèse des signaux en tant qu'opérations de conversion vers ces autre niveaux de représentations, à travers notamment l'utilisation de techniques d'apprentissage et d'analyse statistique ;

[8] Content-based Unified Interfaces and Descriptors for digital Audio Databases available Online

- une fois les modèles de signaux stabilisés, les perspectives ouvertes par la modélisation du contrôle, en tant que signaux à variation lente, résultant notamment d'une analyse de l'interprétation, soit à partir des enregistrements ou de données issues de dispositifs de captation gestuelle ;
- le potentiel d'évolution des modes de production, de diffusion et d'accès de la musique vers des formats étendus (représentations symboliques, métadonnées, enregistrements multipistes) et des fonctions de manipulation configurables et personnalisables (programmation, lutherie virtuelle, interfaces d'accès), fournissant des modes d'interaction plus riches avec les artefacts musicaux.

Remerciements

Les développements qui précèdent ont été inspirés notamment par les recherches, auxquelles ils font largement référence, dirigées à l'Ircam par Xavier Rodet (Analyse/ synthèse des sons), Olivier Warusfel (Spatialisation) et René Caussé (Modélisation physique).

Références

[1] Berkhout A. J., *A Holographic Approach to Acoustic Control.* Journal of the Audio Engineering Society, Vol. 36, No. 12, 1988.

[2] Blauert, J., *Spatial Earing. The psychophysics of Human Sound Localization*, MIT Press, 1997.

[3] Bogaards, N., Röbel, A. et Rodet, X. *Sound Analysis and Processing with AudioSculpt 2.* Proc. Int. Computer Music Conf., 2004.

[4] Corteel, E. *Adaptations de la Wave Field Synthesis aux conditions réelles*. Thèse de doctorat, Université Paris 6, 2004.

[5] de Cheveigné, A. et Kawahara, H. *YIN, a fundamental frequency estimator for speech and music.* Journal of the Acoustical Society of America, Vol. 111, 2002.

[6] Delerue, O. *Spatialisation du son et programmation par contraintes : le système MusicSpace*. Thèse de doctorat, Université Paris 6, 2004.

[7] Dolson, M. et Laroche, J. *Improved phase vocoder time-scale modification of audio* IEEE Transactions on Speech and Audio Processing, vol.7, no.3, 1999.

[8] Hélie, T. *Modélisation physique d'instruments de musique en systèmes dynamiques et inversion.* Thèse de doctoratn Université Paris 11, 2002.

[9] Herrera, P., Peeters, G. et Dubnov, S. *Automatic Classification of Musical Sounds*. Journal of New Musical Research, 2003.

[10] Iovino, F., Caussé, R. et Dudas; R. *Recent work around Modalys and Modal Synthesis*. Proc. of the Int. Computer Music Conf., 1997.

[11] Jot, J.M., *Efficient Models for Distance and Reverberation Rendering in Computer Music and Virtual Audio Reality.* Proc. Int. Computer Music Conf., 1997.

[12] Jullien, J.P. 1995. *Structured Model for the Representation and the Control of Room Acoustical Quality.* Proc. Int. Congress on Acoustics. 1995.

[13] Malham, D.G. et Myatt, A. *3-D Sound Spatialization using Ambisonic Techniques.* Computer Music Journal, Vol 19 N°4, MIT Press, 1995.

[14] Peeters, G. et Rodet, X., *Non-stationary Analysis/Synthesis using Spectrum Peak Shape Distortion, Phase and Reassignment*, Proc. ICSPAT, 1999.

[15] Peeters, G. *Deriving Musical Structures from Signal Analysis for Music Audio Summary Generation : "Sequence" and "State" approach.* LNCS, vol. 2771, Springer Verlag, 2003.

[16] Rodet, X. *Time-domain formant-wave-function synthesis*. Computer Music Journal, Vol.8, N°.3, MIT Press, 1984.

[17] Rodet, X., *Sinusoidal + Residual Models for Musical Sound Signals Analysis/Synthesis*. Applied Signal Processing, Vol. 4 , n° 3, 1998.

[18] Rodet, X., Escribe, J. et Durigon, S. *Improving score to audio alignment: Percussion alignment and Precise Onset Estimation,* Proc. Int. Computer Music Conf., 2004.

[19] Roebel, A. *A new approach to transient processing in the phase vocoder*. Proc. Int. Conf. on Digital Audio Effects (DAFx'03), London, 2003.

[20] Roebel, A. *Adaptive additive modeling with continuous parameter trajectories*. accepté pour publication à IEEE Transactions on Speech and Audio Processing.

[21] Schnell, N. et Schwarz, D., *Gabor, Multi-Representation Real-Time Analysis/Synthesis*. Proc. Int. Conf. on Digital Audio Effects 2005.

[22] Schwarz, D. *Data-Driven Concatenative Sound Synthesis.* Thèse de doctorat, Université Paris 6, 2004.

[23] Schwarz, D. et Wright, M. *Extensions and Applications of the SDIF Sound Description Interchange Format.* Proc. Int. Computer Music Conf., 2000.

[24] Smith, J. *Physical Modeling using Digital Waveguides*, Computer Music Journal, Vol. 16, N° 4, MIT Press, 1992.

[25] Vincent, E. *Modèles d'instruments pour la séparation de sources et la transcription d'enregistrements musicaux.* Thèse de doctorat, Université Paris 6, 2004.

[26] Vinet, H. *The Representation Levels of Music Information.* LNCS, vol. 2771, Springer Verlag, 2003.

[27] Vinet, H, Herrera, P. et Pachet, F. *The CUIDADO Project*. Proc. Int. Conf. on Music Information Retrieval, Ircam, Paris, 2002.

[28] Vinet, H. *The SemanticHIFI Project*. accepté pour publication à Proc. Int. Computer Music Conf., 2005.

[29] Warusfel, O. et Misdariis, N, *Directivity synthesis with 3D array of loudspeakers : application for stage performance.,* Proc. Int. Conf. on Digital Audio Effects 2001.

[30] Yeh C. et Roebel, A. *Physical principles driven joint evaluation of multiple F0 hypotheses* ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing. 2004.

[31] Zils, A. et Pachet, *F. Automatic Extraction of Music Descriptors from Acoustic Signals using EDS*. Proc. of the 116th AES Convention, 2004.

Estimation de mouvement par maillages et lignes de rupture pour le codage de séquences vidéo

Nathalie CAMMAS[1], Stéphane PATEUX[1], Luce MORIN[2], Benjamin LE GUEN[1]

[1]France Télécom R&D, 4 Rue du Clos Courtel, 35510 Cesson-Sévigné, France

[2]IRISA-INRIA, Campus de Beaulieu, 35042 Rennes, France

nathalie.cammas@francetelecom.com,stephane.pateux@francetelecom.com,
luce.morin@irisa.fr,benjamin.leguen@francetelecom.com

Résumé – Dans le cadre de l'analyse et du codage de séquences vidéo, nous présentons ici une nouvelle structure de maillage pour la représentation du mouvement dans les zones d'occlutation. Nous proposons l'utilisation de la notion de lignes de rupture qui permet de représenter une discontinuité dans le champ de mouvement and nous créons un maillage non-manifold permettant d'estimer le mouvement dans les zones d'occultations sans entraîner de dégénérations des mailles. Le maillage non-manifold représente de façon satisfaisante les discontinuités du mouvement introduites par les objets en mouvement et améliore l'estimation du mouvement dans les zones d'occultation. Les résultats montrent une meilleure prédiction de la texture dans ces zones, améliorant ainsi les performances de codage.

Abstract – In this paper, we present a novel mesh structure for representing the motion in occlusion areas for video analysis and video coding purposes. We use the notion of crackline which represents a rupture in a motion field and we create a non-manifold mesh that allows to estimate the motion in the occlusion areas without degeneration of the triangles. The non-manifold mesh can adequately represent the discontinuities in the motion introduced by moving objets and it improves the motion estimation in the occlusion areas. Results show the better prediction of the texture in those areas, thus improving the coding efficiency.

1 Introduction

La représentation du mouvement apparent dans une séquence vidéo reste un problème majeur dans le cadre du codage vidéo. En effet, les performances globales d'un système de codage vidéo dépendent fortement de la qualité de l'estimation du mouvement dans cette séquence. Une mauvaise estimation du mouvement implique une mauvaise prédiction des informations de texture et donc un coût de codage important.

Pour estimer le mouvement dans une séquence vidéo, des techniques basées blocs (type MPEG, H26x) et des techniques basées maillages [1, 2, 5] ont été proposées.

Les maillages 2D déformables offrent des avantages intéressants par rapport à la représentation par blocs. Ils permettent une représentation continue du champ de mouvement, en évitant les effets de blocs qui peuvent apparaître à bas débits avec d'autres techniques, et offrent une description compacte. Les maillages permettent une représentation de mouvements plus complexes et plus précise.

Cependant, tout comme les méthodes par blocs, les méthodes par maillages échouent dans l'estimation du mouvement dans les zones d'occultation. Une zone d'occultaltion est une zone de la séquence vidéo dans laquelle des objets apparaissent, disparaissent, se croisent, créant des phénomènes de découvrement et/ou recouvrement de texture. Dans ces zones, l'estimation du mouvement est difficile car la mise en correspondance devient presque impossible d'une image à l'autre.

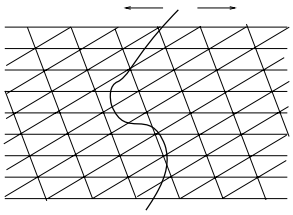
Lors de l'estimation du mouvement par maillages, les découvrements et recouvrements créent des triangles dégénérés: allongés, contractés voire retournés qui entraînent une mauvaise prédiction de la texture.

Pour limiter la déformation des triangles, des techniques proposent de restreindre le déplacement des nœuds du maillage dans des zones de contraintes [3, 4], d'autres techniques proposent de corriger les déplacements des nœuds ayant créés des dégénérations [5, 6, 7]. Toutes ces techniques permettent de limiter la déformation des triangles mais introduisent de faux mouvements. Dans [6], les auteurs proposent d'adapter la structure du maillage aux frontières des objets créant des occlusions, cependant ce traitement doit être ré-appliqué pour chaque image et la structure du maillage n'est pas cohérente au cours du temps.

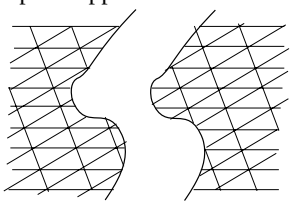
Dans cet article, nous proposons une nouvelle structure de maillage localement adaptatif se basant sur la notion de lignes de rupture et offrant une représentation du mouvement dans les zones d'occultation.

2 Lignes de rupture

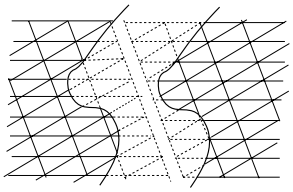
La notion de lignes de rupture a été introduite dans [9]. Une ligne de rupture représente une discontinuité dans le champ de mouvement, cette notion est illustrée sur la figure 1. L'apparition d'une ligne de rupture dans le champ de mouvement est due au mouvement des différents objets. Le maillage est découpé le long de la ligne de rupture et chaque côté du maillage est remaillé indépendamment de l'autre côté, sans les rejoindre. Les deux parties de maillage peuvent alors bouger librement l'une par rapport à l'autre. Le remaillage est appliqué de chaque côté de la ligne de rupture en créant des recouvrements de mailles qui permettront de prendre en compte de futurs découvrements de texture dans cette zone. Cette technique permet d'éviter



(a) Une ligne de rupture apparait dans le champ de mouvement



(b) Le maillage est découpé le long de la ligne de rupture



(c) Chaque côté du maillage est remaillé séparément

FIG. 1: La notion de lignes de rupture sur un maillage

l'apparition de triangles allongés ou tassés et de mieux estimer le mouvement dans les zones d'occultation.
La ligne de rupture est positionnée sur le contour de l'objet créant la discontinuité. Dans le cas d'une ligne de rupture fermée, la découpe du maillage est équivalente à une représentation par objets de la séquence vidéo. Dans le cas d'une ligne de rupture ouverte, les points extrémités de la ligne de rupture nécessite un traitement particulier. En effet, la ligne de rupture est détectée et positionnée à l'intérieur d'une zone d'occultation. En dehors de la zone d'occultation, le maillage reste un maillage 2D classique. Autour de la ligne de rupture, la représentation à l'aide des deux portions de maillage est similaire à une représentation objet, cependant les deux portions de maillage appartiennent au même maillage global. Les deux portions de maillage se rejoignent au niveau des points extrémités de la ligne de rupture. A cet endroit, le maillage devient non-manifold.

3 Estimation du mouvement à l'aide d'un maillage non-manifold

Cette section présente le processus d'estimation du mouvement et détaille la construction du maillage non-manifold ainsi que la reconstruction des images à partir d'un tel maillage.

3.1 Estimation du mouvement

La figure 2 présente le processus d'estimation du mouvement à l'aide du maillage non-manifold. Une première estimation du mouvement est effectuée à l'aide d'un maillage régulier sur un groupe d'images. Le mouvement estimé permet la détection des zones d'occultation dans la séquence par la mesure de la déformation des mailles. Nous considérons connue la segmen-

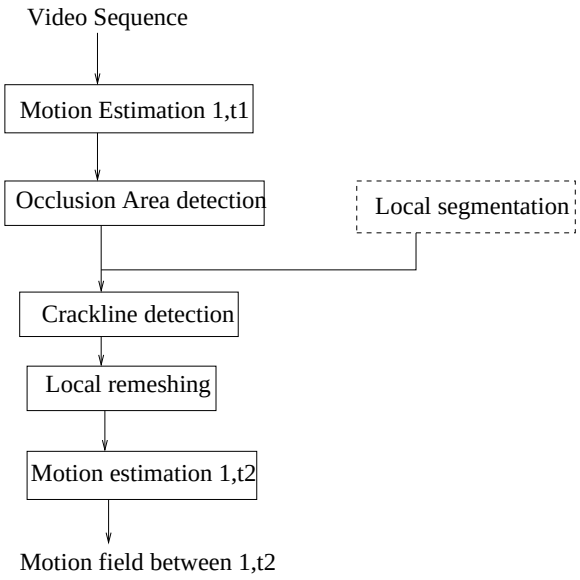


FIG. 2: Estimation du mouvement

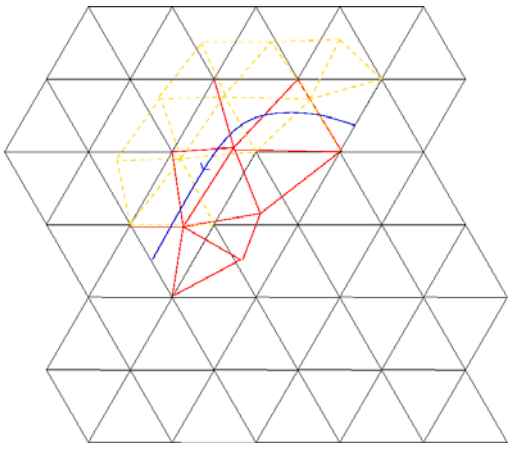


FIG. 3: Maillage non-manifold: le maillage a été coupé le long de la ligne de rupture (en bleu) Puis, chaque partie a été remaillée séparément.

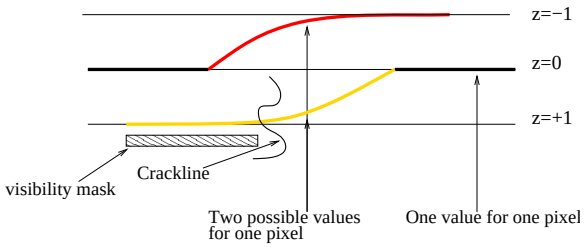


FIG. 4: Technique du z-order et masques de visibilité. La figure représente une coupe du maillage de la figure 3

tation objet locale dans ces zones d'occultation.
Dans chaque zone d'occultation, la ligne de rupture est alors placée sur le contour de l'objet. Le maillage non-manifold est ensuite construit à partir du maillage régulier et de la ligne de rupture. Le mouvement est enfin réestimé à l'aide du maillage non-manifold.

3.2 Construction du maillage non-manifold

Le maillage régulier est découpé le long de la ligne de rupture. Chaque côté est alors remaillé en utilisant les arêtes des triangles du maillage régulier pour créer de nouveaux triangles. Le remaillage est propagé en dehors de la zone d'occultation pour prendre en compte de futurs découvrements de texture. Le maillage régulier utilisé est un maillage hiérarchique, la ligne de rupture et le remaillage sont alors propagés dans la hiérarchie. Les niveaux de maillage plus grossiers sont également découpés et remaillés. La ligne de rupture disparait lorsqu'elle est entièrement contenue dans une maille d'un niveau, ce niveau n'est alors pas découpé.
La figure 3 montre le maillage non-manifold obtenu à un niveau de maillage fin. La ligne bleue représente la ligne de rupture. Le maillage régulier est représenté par les mailles en noires. Les mailles rouges et jaunes pointillées sont les mailles issues du remaillage de chaque côté de la ligne de rupture: les mailles rouges sont en avant, les jaunes pointillées vers l'arrière. On peut voir sur la figure que le maillage est non-manifold aux points extrémités de la ligne de rupture. En ces points, des arcs partagent plus de deux mailles.

3.3 Reconstruction des images

La technique de remaillage a créée des recouvrements de mailles pour prendre en compte de futurs découvrements de texture. Cependant, ces recouvrements de mailles génèrent une ambiguïté lors de la reconstruction d'une image. En effet, plusieurs triangles peuvent alors reconstruire un même pixel. Pour gérer ce problème, nous utilisons une technique de z-order couplée à des masques de visibilité.
La technique du z-order est similaire à celle du z-buffer en 3D. Une valeur de z est affectée aux nœuds du maillage et permet d'indiquer si le triangle est en avant ou en arrière de la scène. Comme la ligne de rupture peut être définie n'importe où sur un triangle, des masques de visibilité sont calculés pour chaque triangle indiquant quelle partie du triangle par rapport à la position de la ligne de rupture est visible. Pour éviter l'apparition de pixels de luminance inconnue qui pourraient survenir dans le cas du codage avec pertes de la ligne de rupture, les masques de visibilité ne sont définis que pour les triangles en avant de la scène. La figure 4 illustre la technique du z-order et des masques de visibilité.
Cette technique permet de reconstruire une image sans ambiguïté sur les valeurs de luminance utilisées.

	MM	Non-MM	Ligne de rupture
Mobile And Calendar	8.1kb/s	7.9kb/s	1kb/s
Erik	18.68kb/s	18.20kb/s	3kb/s
Flower Garden	20.70kb/s	17kb/s	2kb/s

TAB. 1: Coût de codage du mouvement et de la ligne de rupture pour un maillage manifold (MM) et un maillage non-manifold (non-MM)

4 Résultats

Le schéma d'estimation du mouvement a été présenté en section 3. Une présentation plus détaillée de l'estimateur de mouvement est disponible dans [8].
Dans nos expérimentations, le mouvement a été estimé sur un groupe de huit images. Nous avons utilisé une segmentation connue a priori pour détecter la ligne de rupture dans la zone d'occultation. Les tests ont été effectués sur les séquences Mobile And Calendar (CIF, 30Hz) et Erik (CIF, 30Hz). Pour la séquence Mobile And Calendar, la zone d'occultation est la zone frontière entre le ballon rouge et l'arrière-plan, la ligne de rutpure a été positionnée sur le contour du ballon rouge. Pour la séquence Erik, la ligne de rupture est positionnée sur le contour de la personne en avant-plan. La figure 6 montre les résultats d'estimation du mouvement pour la séquence Mobile And Calendar à l'aide d'un maillage manifold et du maillage non-manifold créé avec la ligne de rupture. La figure montre également la prédiction d'images de la séquence par la première image compensée en mouvement. Sans les lignes de rupture, la prédiction dans la zone d'occultation montre des textures étirées. Ce phénomène est très nettement attenué avec le maillage non-manifold. Le même constat peut être fait pour la séquence Erik. Le maillage non-manifold permet d'améliorer l'estimation du mouvement dans les zones d'occultation ainsi que la prédiction de la texture dans ces zones.
Le maillage non-manifold a été intégré dans un système de codage vidéo [10]. Dans ce schéma, il est nécessaire de prendre en compte le codage de la ligne de rupture, cependant son coût est faible (de 1 à 3kb/s). De plus, l'utilisation du maillage non-manifold permet d'améliorer le coût de codage du mouvement, ce qui permet de compenser alors le surcoût de la ligne de rupture. Le tableau 1 montre le coût de codage du mouvement et de la ligne de rupture pour un maillage manifold et un maillage non-manifold pour les séquences Erik, Mobile And Calendar et Flower Garden.

5 Conclusion

Nous avons présenté une nouvelle structure de maillage non-manifold utilisant la notion de lignes de rupture. Le maillage non-manifold permet de représenter les discontinuités du mou-

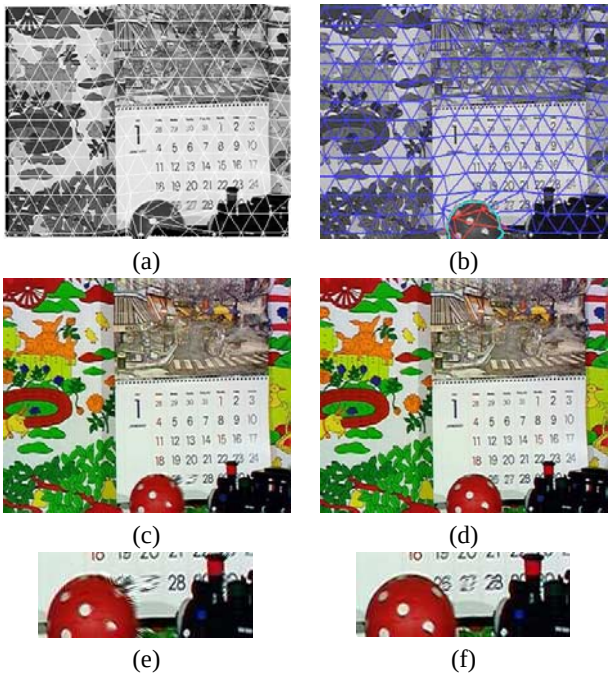


FIG. 5: Estimation du mouvement: comparaison entre un maillage manifold (a) et notre maillage non-manifold (b). (c) Prédiction de la texture par un maillage manifold et (e) zoom sur la zone d'occultation. (d) Prédiction de la texture par notre maillage non-manifold et (f) zoom sur la zone d'occultation.

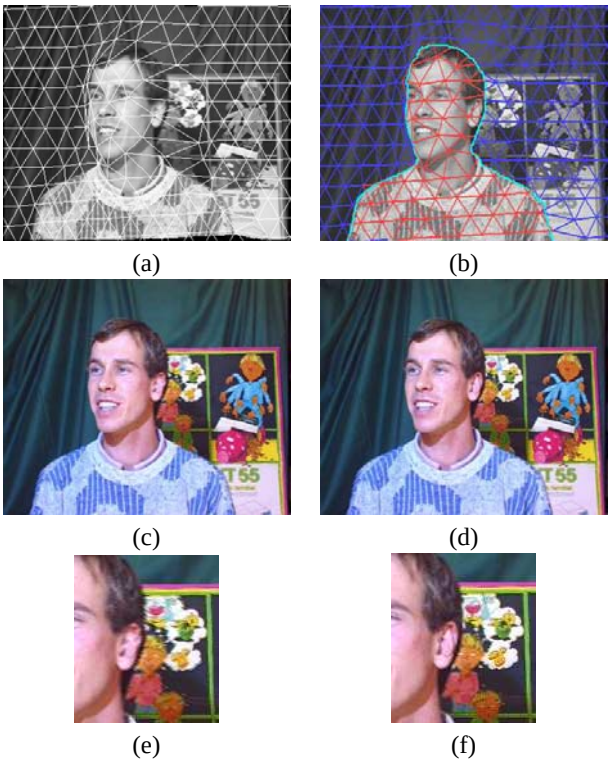


FIG. 6: Estimation du mouvement: comparaison entre un maillage manifold (a) et notre maillage non-manifold (b). (c) Prédiction de la texture par un maillage manifold et (e) zoom sur la zone d'occultation. (d) Prédiction de la texture par notre maillage non-manifold et (f) zoom sur la zone d'occultation.

vement introduites par les objets en mouvement. Ce maillage a été utilisé dans un schéma d'estimation du mouvement d'une séquence vidéo. Nous avons présenté des résultats visuels et quantitatifs sur des séquences réelles qui montrent que le maillage non-manifold permet d'améliorer l'estimation du mouvement dans les zones d'occultation et la prédiction de la texture dans ces zones. Les travaux futurs sur ce schéma auront pour but de proposer une représentation de la texture sous la forme de surface adaptée au maillage non-manifold afin de l'intégrer dans un schéma de codage vidéo.

References

[1] Y. Nakaya and H. Harashima, "Iterative motion estimation method using triangular patches for motion compensation" *Proceedings of SPIE Conference on Visual Communication and Image Processing*, Boston, USA, Nov 1991, vol 1605-2, pp. 500-510.

[2] Y. Wang and O. Lee, "Active mesh-A feature seeking and tracking image sequence representation scheme" *IEEE Transactions on image processing*, vol 3:5, pp. 610-622, Sept 1994.

[3] Y. Wang and O. Lee and A. Vetro, "Use of two-dimensional deformable mesh structures for video coding, part II- The analysis problem and a region-based coder employing an active mesh representation" *IEEE Transactions on circuits and systems for video technology*, vol 6:6, pp. 647-659, Dec 1996.

[4] N. Laurent and P. Lechat and H. Sanson, "Limitation of triangles overlapping in mesh-based motion estimation using augmented lagrangian" *5th IEEE International Conference on Image Processing, ICIP'98*, Chicago, USA, Oct 1998, vol 2, pp. 223-227.

[5] C. Toklu and A. M. Tekalp and T. Erdem, "2-D Triangular mesh-based mosaicking for object tracking in the presence of occlusion" *Proc. of SPIE Visual Communication and Image Processing, VCIP'97*, pp. 328-337, 1997.

[6] Y. Altunbasak and A.M. Tekalp, "Occlusion-adaptive, content-based mesh design and forward tracking", *IEEE Transactions on image processing*, vol 6:9, pp. 1270-1280, Sept 1997.

[7] N. Laurent, "Hierarchical mesh-based global motion estimation, including occlusion area detection" *7th IEEE International Conference on Image Processing, ICIP'00*, Sept 2000, Vancouver, Canada.

[8] S. Pateux and G. Marquant and D. Chavira-Martinez, "Object mosaicking via meshes and cracklines technique. Application to low bit-rate video coding" *Proceedings of Picture Coding Symposium 2001*, Seoul, Korea, 2001.

[9] G. Marquant and S. Pateux and C. Labit, "Mesh and "Crack Lines": Application to object-based motion estimation and higher scalability" *7th IEEE International Conference on Image Processing, ICIP'00*, Sept 2000, Vancouver, Canada.

[10] N. Cammas and S. Pateux, "Fine grain scalable video coding using 3D wavelets and active meshes" *SPIE Visual Communications and Image Processing, VCIP 2003* Santa Clara, USA, Jan 2003.

Codage vidéo par schéma lifting avec gestion des occlusions

Thomas ANDRÉ, Marc ANTONINI, Michel BARLAUD

Laboratoire I3S, CNRS - UMR 6070, Université de Nice-Sophia Antipolis
Bât. Algorithmes / Euclide B, BP 121, 2000 route des Lucioles - 06903 Sophia Antipolis Cedex, France
{andret,am,barlaud}@i3s.unice.fr

Résumé – Le schéma lifting compensé en mouvement est utilisé dans la plupart des codeurs vidéo basés ondelettes. Cependant, l'estimation et la compensation de mouvement à l'aide de blocs entraîne l'apparition d'artéfacts visibles autour des objets en mouvement et du bord des images. Dans ce papier, nous proposons une nouvelle méthode de filtrage temporel qui fait appel à une segmentation et une estimation de mouvement conjointes. Le principe consiste à attribuer un mouvement à des régions de forme adaptable au lieu d'utiliser des blocs. Nous présentons d'une part l'algorithme de filtrage «Puzzle» et étudions les conditions de son inversibilité. D'autre part, nous proposons une méthode d'extraction des régions d'occlusion à partir des informations de segmentation et de mouvement ; ces régions sont ensuite utilisées pour gérer les occlusions. Les premiers résultats expérimentaux confirme la diminution des effets de blocs ; la bonne gestion des occlusions permet une baisse significative de l'entropie des sous-bandes temporelles.

Abstract – Motion-compensated lifting schemes have become a reference for the temporal filtering of video data. However, block-based motion estimation and compensation produce annoying blocking artifacts around the moving objects and near the borders of the images. In this paper, we propose a new lifted temporal filtering method, based on joint segmentation and motion estimation. This method consists in attributing locally the motion information to content-adapted regions instead of blocks. We first present our "Puzzle filtering" algorithm and we state the conditions for its invertibility. Then, we propose a method to extract regions of occlusion from the motion and segmentation information. The obtained regions are finally exploited within the proposed Puzzle filtering. First experimental results show that the blocking artifacts are completely removed and that the occlusions are successfully managed, which results in an important subband entropy decrease.

1 Introduction

La compression vidéo connaît un essor important depuis quelques années. Des normes très efficaces ont été établies, aboutissant par exemple aux codeurs hybrides MPEG-4 et H.264/AVC. Les codeurs basés ondelettes [1, 2], qui font intervenir un schéma lifting $t + 2D$ compensé en mouvement [3], permettent un meilleur support de la scalabilité [4] et atteignent presque les performances des codeurs hybrides [5]. Toutefois, de nombreux travaux de recherches visent encore à en améliorer les performances, et notamment celles du filtrage temporel en ondelettes. En particulier, la plupart des codeurs vidéo cités ci-dessus font appel à une estimation et une compensation du mouvement à base de «block-matching», qui ne gère pas les occlusions ni le chevauchement d'objets. En conséquence, systématiquement, certains blocs sont situés à cheval sur deux régions de mouvements différents, ce qui provoque l'apparition d'effets de blocs. Une meilleure gestion de ce problème devrait permettre d'améliorer grandement les performances de codage.

Nous proposons une méthode de filtrage temporel fondée sur le schéma lifting compensé en mouvement [5], qui utilise les informations d'estimation de mouvement et de segmentation, afin de mieux prendre en compte les objets en mouvement. Nous supposons qu'il n'y a dans l'image que deux régions différentes, un «objet» et un «fond» ; cette hypothèse, réaliste si locale, permet de simplifier le problème, d'autant plus que la segmentation ne doit pas être nécessairement sémantique pour que l'algorithme fonctionne.

Dans ce papier, nous décrivons tout d'abord une méthode de filtrage temporel compensé en mouvement fondée sur le schéma lifting, que nous appelons «Puzzle», capable de filtrer différentes régions avec un mouvement différent (section 2). Ensuite, nous décrivons un algorithme capable d'extraire les régions utiles à la gestion des occlusions (section 3), à partir des informations fournies par un algorithme conjoint de segmentation et d'estimation de mouvement. Enfin, nous présentons les premiers résultats expérimentaux, qui montrent l'efficacité de la méthode pour gérer les occlusions et donc diminuer fortement l'entropie des sous-bandes temporelles.

2 Schéma lifting orienté régions

2.1 Diviser un macrobloc en régions

La compensation de mouvement a permis une amélioration significative des performances des codeurs vidéos, même avec un modèle très simple de mouvement de translation uniforme par blocs. Cependant, cette technique entraîne l'apparition d'artéfacts (effet de blocs) dans les images reconstruites. Plusieurs méthodes ont été proposées pour réduire ces effets de blocs. Par exemple, les derniers standards MPEG et H.264, ainsi que de nombreux codeurs basés ondelettes, font appel à des blocs de taille variable. D'autres alternatives, comme les maillages déformables [6, 7], ont également été envisagées.

Nous proposons une méthode flexible qui utilise des régions de forme arbitraire. La figure 1 montre en effet qu'il est souvent plus précis de diviser un macrobloc en deux régions plutôt qu'en blocs plus petits. Nous avons choisi de modéliser la limite des régions par quatre points de contrôle interpolés par

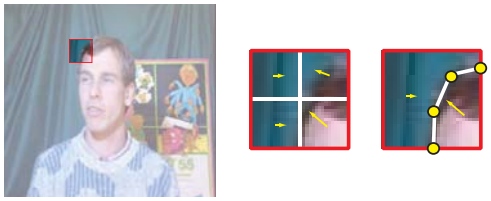


FIG. 1 – Division d'un macrobloc en deux régions adaptatives au lieu de quatre blocs.

une spline, qui permet une représentation précise de la courbe tout en limitant son coût de codage. Il est à noter que le surcoût d'information à transmettre est comparable dans les deux cas. En effet, 4 vecteurs se codent sur $4 \times 2 \times N_v$ bits, alors que 2 vecteurs et 4 points se codent sur $2 \times 2 \times N_v + 4 \times 2 \times N_p$ bits, où $N_v \in [5;8]$ est la précision (en bits) des coordonnées des vecteurs, et $N_p \in [4;5]$ celle des coordonnées des points, fonctions des paramètres d'estimation du mouvement et de segmentation. Par ailleurs, nous montrons en section 3 que cette représentation en régions permet une gestion précise des occlusions.

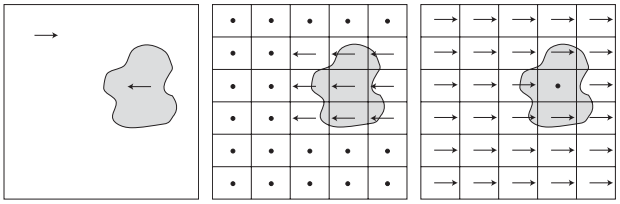
La méthode proposée nécessite au préalable de connaître la segmentation de chaque image en régions de mouvements différents (Fig. 2). Cette information peut être obtenue manuellement (par exemple en post-production cinématographique), ou bien à l'aide d'un algorithme conjoint de segmentation et d'estimation du mouvement [8, 9].

2.2 Filtrage orienté régions

Dans cette partie, nous supposons connue la segmentation de chaque image de la séquence en deux régions : un «objet» arbitraire, et un «fond» arbitraire. Cette segmentation n'est pas nécessairement sémantique, et le rôle de l'objet et du fond est totalement symétrique. Nous supposons également que nous connaissons le mouvement respectif de ces deux régions. Ces informations sont données par un algorithme conjoint d'estimation de mouvement et de segmentation, ou bien par une segmentation manuelle suivie d'une estimation de mouvement classique, telle que celle utilisée dans MPEG.

L'objectif est de filtrer temporellement l'objet et le fond de façon indépendante, en utilisant leur mouvement respectif, à l'aide par exemple de la transformée en ondelettes (2,2) sous sa forme de schéma lifting. La figure 2a montre l'exemple d'une séquence simple dans laquelle un objet rigide se déplace sur un fond également en mouvement. L'estimateur de mouvement découpe chaque image en macroblocs, puis détermine dans chaque macrobloc un «objet» et un «fond», d'après les informations de segmentation. Ensuite, il détermine séparément le mouvement de ces deux régions, à l'aide d'un algorithme classique de mise en correspondance. Finalement, on obtient deux régions par macrobloc, et un vecteur mouvement par région et par bloc, comme indiqué sur les figures 2b et 2c.

Notons (I_i) les images de la séquence originale. Pour chaque image i, notons M_i^{obj} le masque de l'objet : pour chaque pixel p de l'image, on aura $M_i^{obj}(p) = 1$ si p est situé sur l'objet, et $M_i^{obj}(p) = 0$ sinon. Le masque du fond, noté M_i^{fond}, est



(a) Segmentation (b) Mouvement : objet (c) Mouvement : fond

FIG. 2 – Exemple d'une séquence simple où un objet rigide (gris) se déplace sur un fond en mouvement. Il y a un vecteur mouvement par bloc pour l'objet, et un par bloc pour le fond, représentés par des flèches. Un point représente un vecteur nul.

défini de la même façon, et on a $M_i^{fond} = \overline{M_i^{obj}}$.

Notons également $v_{i\to j}^{obj}(b)$ le vecteur représentant le mouvement de l'objet du bloc b, de l'image i vers l'image j. Le mouvement du fond, $v_{i\to j}^{fond}(b)$, est défini de la même façon. Ces vecteurs permettent de compenser en mouvement des images ou des masques ; par exemple, on notera $I_{2i-1}\left(v_{2i-1\to 2i}^{fond}\right)$ l'image I_{2i-1} compensée en mouvement à l'aide des vecteurs mouvement du fond $v_{2i-1\to 2i}^{fond}(b)$, donc recalée temporellement sur l'image I_{2i}.

2.2.1 Analyse

La séquence originale (I_i) peut être filtrée temporellement avec les filtres de la transformée (2,2) compensée en mouvement. Selon si le mouvement du fond ou celui de l'objet est utilisé, on obtient la sous-bande haute-fréquence H_i^{fond} ou H_i^{obj}, calculées comme suit :

$$H_i^{fond} = I_{2i} - \frac{1}{2}\left[I_{2i-1}\left(v_{2i-1\to 2i}^{fond}\right) + I_{2i+1}\left(v_{2i+1\to 2i}^{fond}\right)\right]$$

$$H_i^{obj} = I_{2i} - \frac{1}{2}\left[I_{2i-1}\left(v_{2i-1\to 2i}^{obj}\right) + I_{2i+1}\left(v_{2i+1\to 2i}^{obj}\right)\right]$$

Puisque l'on souhaite filtrer l'objet avec le mouvement de l'objet, et le fond avec le mouvement du fond, il suffit d'exprimer la sous-bande haute-fréquence finale H_i sous la forme :

$$H_i = H_i^{fond} * M_{2i}^{fond} + H_i^{obj} * M_{2i}^{obj}$$

où l'opérateur $*$ représente la multiplication terme à terme de deux matrices. Finalement, H_i est localement égale à H_i^{fond} ou bien à H_i^{obj}, les deux masques M_i^{obj} et M_i^{fond} étant complémentaires dans l'image I_i.

La sous-bande basse-fréquence est calculée de façon similaire.

2.2.2 Synthèse

La réversibilité du schéma proposé peut être établie rapidement à partir des équations précédentes. En inversant le schéma lifting de façon classique, on obtient :

$$I_{2i-1} * \left[M_{2i-1}^{fond} + M_{2i-1}^{obj}\right] = L_i + f_L(H_{i-1}, H_i)$$

$$I_{2i} * \left[M_{2i}^{fond} + M_{2i}^{obj}\right] = H_i + f_H(I_{2i-1}, I_{2i+1})$$

où $f_L(H_{i-1}, H_i)$ et $f_H(I_{2i-1}, I_{2i+1})$, obtenues après un calcul simple, sont des expressions indépendantes de l'image que l'on cherche à reconstruire. Puisque la réunion des masques M_i^{obj} et M_i^{fond} couvre la totalité de la surface de l'image, ceci établit la réversibilité de la transformée.

2.2.3 Généralisation à N régions

Par la suite, nous verrons que la gestion des occlusions nécessite de distinguer plus de 2 régions, qui seront traitées différemment, avec des mouvements ou des filtres différents. Supposons que chaque macrobloc est maintenant composé de N régions. Notons M_i^n le masque qui caractérise la région n de l'image i, et $v_{i\to j}^n$ le mouvement de cette région n entre l'image i et l'image j.

En nous appuyant sur les résultats précédents, nous pouvons écrire par exemple la sous-bande HF comme suit :

$$H_i = \sum_{n=1..N} H_i^n * M_{2i}^n$$

avec, dans le cas d'un filtrage (2,2) :

$$H_i^n = I_{2i} - \frac{1}{2}\left[I_{2i-1}(v_{2i-1\to 2i}^n) + I_{2i+1}(v_{2i+1\to 2i}^n)\right]$$

La réversibilité est alors établie de la même façon que précedemment, en inversant le schéma lifting. Par exemple, en inversant le pas de prédiction, on obtient l'équation de reconstruction suivante :

$$I_{2i} * \left[\sum_{n=1..N} M_{2i}^n\right] \quad = \quad H_i + f_H(I_{2i-1}, I_{2i+1})$$

En d'autres termes, le schéma proposé est réversible si les masques $(M_i^n)_{n=1..N}$ sont complémentaires dans l'image i.

Les calculs qui concernent le pas de mise à jour sont similaires.

3 Gestion des occlusions

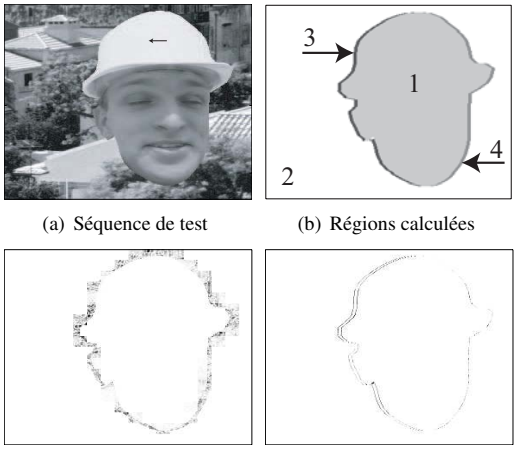
Le schéma de filtrage proposé en section 2, dans sa version généralisée, peut être utilisé pour gérer les occlusions lors du filtrage temporel compensé en mouvement. Il suffit, en effet, de subdiviser les deux régions «objet» et «fond» en sous-régions occultées ou non, et d'appliquer le bon filtre à chaque sous-région.

3.1 Régions et filtrage

Ainsi, par exemple, le fond de l'image I_i peut être subdivisé en trois régions : le fond qui est apparu par rapport à l'image précédente I_{i-1}, le fond qui est apparu par rapport à l'image suivante I_{i+1}, et le fond qui est visible partout sur les trois images successives. Cette dernière région est filtrée à l'aide d'une transformée symétrique, comme la transformée (2,2). Les deux autres régions le sont par une transformée causale ou anti-causale selon la direction souhaitée, comme l'ondelette de Haar. De cette façon, les régions occultées n'interviennent pas dans le calcul des sous-bandes.

L'objet est traité de la même manière que le fond ; au total, six sous-régions doivent être déterminées afin de traiter les occlusions.

La figure 3b montre un exemple simple d'une telle séparation en sous-régions. Dans cet exemple, l'objet n'est pas occulté : il est donc entièrement prédictible en bidirectionnel et sera filtré par transformée (2,2). En revanche, le fond doit être divisé en 3 sous-régions afin de tenir compte des occlusions ; les sous-régions occultées seront filtrées à l'aide de l'ondelette de Haar dans la direction appropriée. Les sous-bandes HF et BF finales seront obtenues à l'aide des équations généralisées données en section 2.2.3.



(a) Séquence de test (b) Régions calculées

(c) Filtrage classique (d) Filtrage proposé

FIG. 3 – Séquence de test utilisée (a) ; régions extraites par l'algorithme proposé (1 = objet prédictible dans les deux directions, 2 = fond prédictible dans les deux directions, 3 = fond prédictible à partir de l'image suivante, 4 = fond prédictible à partir de l'image précédente) ; exemple de sous-bandes haute-fréquence (HF) obtenues avec un filtrage classique (c) et avec l'algorithme proposé (d).

3.2 Extraction des régions

Il est maintenant nécessaire de déterminer ces 6 sous-régions, en utilisant uniquement des informations disponibles au codage et au décodage. Nous proposons donc un algorithme qui permet de les extraire à partir des informations sur la segmentation simple en «objet» et «fond», et sur le mouvement de ces deux régions. Par exemple, déterminons la région du fond de I_i qui est prédictible à partir de l'image suivante I_{i+1} seulement ; cette région est définie par son masque que nous noterons $M_i^{fond,suiv}$. Il s'agit de la zone du fond qui est visible dans l'image I_i mais pas dans l'image I_{i-1} ; en d'autres termes, elle correspond au fond de l'image I_i, privé du fond de l'image I_{i-1} que l'on a recalé au préalable sur l'image I_i :

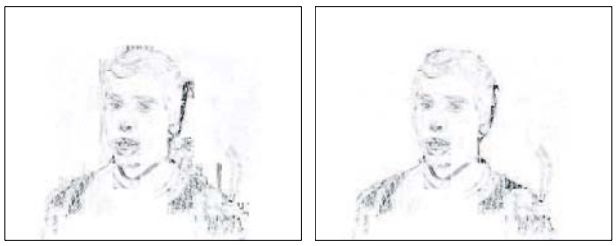
$$M_i^{fond,suiv} = M_i^{fond} - \left[M_i^{fond} \bigcap M_{i-1}^{fond}\left(v_{i-1\to i}^{fond}\right)\right]$$

Le masque du fond prédictible à partir de l'image précédente I_{i-1} seulement, noté $M_i^{fond,prc}$, est calculé de la même manière. On s'assure alors que les deux masques sont complémentaires. Dans le cas contraire, on leur retranche leur partie commune, qui correspond à la zone du fond qui n'est prédictible dans aucune direction. Enfin, la zone du fond prédictible dans les deux directions est calculée en retranchant à la zone du fond les deux zones calculées précédemment.

Le traitement du «fond» et de «l'objet» est réalisé de manière symétrique.

4 Premiers résultats

La figure 3 montre les premiers résultats expérimentaux obtenus avec une séquence de test synthétique, dans laquelle un



(a) Filtrage classique (b) Filtrage par Puzzle

FIG. 4 – Filtrage HF de la séquence «Erik» par les deux méthodes : image 2. Les pixels les plus foncés représentent les plus grandes valeurs absolues.

objet rigide texturé (la tête du personnage) bouge sur un fond texturé. La segmentation de l'objet a été ici réalisée à la main, et contient une erreur d'un pixel sur la position du contour. Le mouvement a été estimé à l'aide d'un algorithme de «block-matching» classique, pour une taille de blocs de 16×16 pixels.

L'algorithme d'extraction de régions présenté en section 3 fournit 4 régions (b), et non 6, puisque l'objet est rigide et ne disparaît pas d'une image à l'autre. Un filtrage par schéma lifting classique compensé en mouvement fournit des sous-bandes haute-fréquence (HF) telles que celle montrée en (c). L'algorithme proposé fournit la sous-bande (d), qui ne contient plus d'effets de blocs mais seulement des «contours» d'erreur dus à l'imprécision de la segmentation. L'étude de l'énergie et de l'entropie des sous-bandes (tableau 1) confirme que ces dernières peuvent être codées bien plus efficacement par la méthode proposée que par la méthode classique.

Filtres		$(2,2)$		$(2,0)$	
Filtrage		B	P	B	P
Puissance	HF $(.10^{-3})$	9.4	1.0	9.4	1.0
	BF $(.10^{4})$	1.54	1.53	1.53	
Entropie (bpp)	HF	1.36	0.17	1.36	0.17
	BF	8,52	7.75	7.60	

TAB. 1 – Comparaison de la méthode classique de filtrage par blocs (B) et de la méthode proposée (P), pour deux ensembles de filtres temporels [10] : puissance et entropie de la sous-bande haute-fréquence (HF) et basse-fréquence (BF), moyenne sur les 8 premières images.

Nous avons également testé la méthode proposée sur la séquence réelle «Erik» ; un exemple des sous-bandes HF obtenues est présenté sur la figure 4. Nous avons comparé les sous-bandes obtenues en subdivisant des macroblocs de 32×32 en blocs de 16×16 (méthode classique) ou en régions (méthode proposée) là où c'était nécessaire, c'est-à-dire dans les zones d'occlusion. Si les sous-bandes produites par les deux méthodes avaient une énergie et une entropie comparables, il apparaît que la méthode proposée permet néanmoins d'obtenir des sous-bandes plus propres, avec beaucoup moins d'effets de blocs.

5 Conclusion

Nous avons présenté une nouvelle méthode de filtrage temporel par schéma lifting qui utilise des informations de segmentation et d'estimation de mouvement conjointes afin de gérer les occlusions. Cette méthode consiste à appliquer un schéma lifting temporel adapté à des régions, de forme quelconque, qui prennent en compte les occlusions. Ces régions sont calculées à partir des seules informations de segmentation et de mouvement. Les premiers résultats montrent une amélioration de la qualité des sous-bandes et une diminution des effets de blocs autour des objets en mouvement. Les travaux futurs concerneront l'intégration de cette méthode dans un codeur vidéo complet.

Références

[1] S.J. Choi and J.W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. on Image Processing*, vol. 8, no. 2, pp. 155–167, Feb. 1999.

[2] S. Cho and W.A. Pearlman, "A full-featured, error-resilient, scalable wavelet video codec based on the Set Partitioning in Hierarchical Trees (SPIHT) algorithm," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 12, pp. 171–170, Mar. 2002.

[3] J. Viéron, C. Guillemot, and S. Pateux, "Motion compensated 2D+t wavelet analysis for low rate fgs video compression," in *Proc. of Tyrrhenian Intern. Workshop on Digital Comm.*, Capri, Italy, Sept. 2002.

[4] G. Pau, C. Tillier, B. Pesquet-Popescu, and H. Heijmans, "Motion compensation and scalability in lifting-based video coding," *EURASIP Signal Processing : Image Communication, special issue on Wavelet Video Coding*, pp. 577–600, Aug. 2004.

[5] M. Cagnazzo, T. André, M. Antonini, and M. Barlaud, "A model-based motion compensated video coder with JPEG2000 compatibility," in *IEEE Intern. Conf. on Image Processing*, Singapore, Oct. 2004, pp. 2255–2258.

[6] A. Secker and D. Taubman, "Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression," *IEEE Transaction on Image Processing*, vol. 12, no. 12, pp. 1530–1542, Dec. 2003.

[7] N. Cammas and S. Pateux, "Fine grain scalable video coding using 3d wavelets and active meshes," in *SPIE Visual Communications and Image Processing, VCIP 2003*, January 2003.

[8] E. Debreuve, M. Gastaud, M. Barlaud, and G. Aubert, "A region-based joint motion computation and segmentation on a set of frames," in *Proceedings of European Workshop on Image Analysis for Multimedia International Services (WIAMIS)*, Montreux, Swiss, Apr. 2005.

[9] S. Boltz, E. Debreuve, and M. Barlaud, "A joint motion segmentation algorithm for video coding," in *Proceedings of EUSIPCO*, Antalya, Turkey, 2005 (to appear).

[10] T. André, M. Cagnazzo, M. Antonini, M. Barlaud, N. Božinović, and J. Konrad, "(N,0) motion-compensated lifting-based wavelet transform," in *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004.

Réduction du débit et de la complexité de décodage par l'analyse de scène pour la télé surveillance basée sur Motion JPEG 2000

J. MEESSEN, C. PARISOT, X. DESURMONT et J.-F. DELAIGLE

Multitel A.S.B.L., Avenue Copernic 1, B-7000 Mons, Belgique.

jerome.meessen@multitel.be

Résumé – *Dans ce papier, nous proposons un système de codage/transmission vidéo orienté objet utilisant le standard Motion JPEG 2000 [1] pour le stockage et la diffusion efficaces de vidéo surveillance sur les canaux bas débit. La méthode présente les mêmes avantages que les techniques actuelles de codage de régions d'intérêt tout en améliorant considérablement le rapport débit/qualité moyen de la vidéo transmise, lorsque les caméras sont statiques. Nous proposons ici de transmettre le flux vidéo en transmettant d'une part un flux Motion JPEG 2000 contenant une estimation du fond rafraîchie régulièrement et, d'autre part, un second flux Motion JPEG 2000 ne contenant que les objets mobiles de la scène. Cette méthode fournit une meilleure qualité de la vidéo décodée, qui se rapproche de celle obtenue avec un codeur MPEG-4, mais aussi une complexité réduite pour le client par rapport à un encodage Motion JPEG2000 standard, avec un surcoût de stockage négligeable. Les flux de surveillance stockés sur le serveur sont completement conformes au standard Motion JPEG 2000.*

Abstract – *In this paper, we propose a new object-based video coding/transmission system using the emerging Motion JPEG 2000 standard for the efficient storage and delivery of video surveillance over low bandwidth channels. The method provides the advantages of state-of-the art techniques for Regions-Of-Interest (ROI) coding while significantly improving the average bitrate/quality ratio of delivered video when cameras are static. We transmit only ROIs of each frame as well as an automatic estimation of the background at a lower frame rate in two separate Motion JPEG 2000 streams. This method provides both better video quality and reduced client CPU usage with negligible storage overhead when compared to standard Motion JPEG 2000 encoding. Video surveillance streams stored on the server are fully compliant with existing Motion JPEG 2000 decoders.*

1. INTRODUCTION

Dans ce papier, nous proposons un système de codage/transmission vidéo orienté objet utilisant le standard Motion JPEG 2000 [1] pour le stockage et la diffusion efficaces de vidéo surveillance sur les canaux bas débit. Les techniques de compression JPEG 2000 basées sur les Régions d'Intérêt (ROI) [2][3] permettent d'assurer une qualité supérieure pour les régions d'intérêt par rapport au reste lorsque le débit disponible est trop faible pour transmettre chaque image avec une qualité suffisante. Dans les systèmes à caméra fixe, des parties du fond sont cependant transmises de façon redondante à chaque image de la séquence Motion JPEG 2000. Nous proposons ici de transmettre le flux vidéo en transmettant d'une part un flux Motion JPEG 2000 contenant une estimation du fond rafraîchie régulièrement et, d'autre part, un second flux Motion JPEG 2000 ne contenant que les objets mobiles de la scène.

La figure suivante présente l'architecture générale du système proposé :

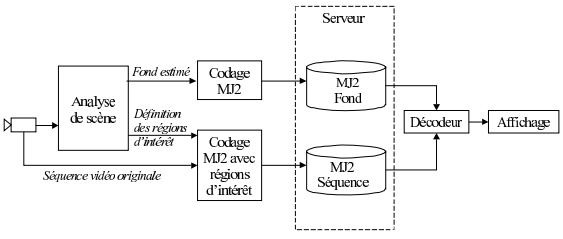


Figure 1: Architecture générale du système

Le flux vidéo d'entrée est analysé de façon automatique par un module réalisant une estimation du fond de la séquence mis à jour à chaque image. La segmentation des objets mobiles de la scène (ROI) est aussi réalisée par ce module. La séquence originale est codée en Motion JPEG 2000, les données relatives aux objets mobiles étant placées dans une première couche de qualité du flux compressé, le reste des images dans une deuxième couche. Le décodage de cette séquence permet la reconstruction de chacune des images originales. L'estimation du fond est codée, elle aussi,

dans un flux Motion JPEG 2000 à intervalles réguliers,par exemple une fois toutes les quatre secondes. Le serveur envoie au décodeur l'estimation du fond et une partie des données de la première couche de la séquence compressée. Le client reconstruit alors la séquence à partir des deux flux sans nécessiter une description explicite des zones d'intérêt.

Le paragraphe *2* traite de la segmentation automatique de la scène. Le paragraphe 3 décrit l'approche proposée pour la diffusion optimisée des flux vidéo segmentés. Avant de conclure, le paragraphe 4 présente les résultats obtenus.

2. SÉPARATION AUTOMATIQUE DU FOND ET DES OBJETS MOBILES

Pour tenir compte de fonds variables (bruit d'acquisition, mouvements périodiques…) nous utilisons un algorithme de segmentation statistique temps réel modélisant le fond par une mixture de Gaussiennes en chaque pixel [4][5][6].

A chaque image, la valeur du pixel est comparée à l'état courant de la mixture de Gaussiennes correspondante. Si cette valeur appartient à l'une des Gaussiennes les plus souvent observées, le pixel est classé comme faisant partie du fond. Sinon il est segmenté comme faisant partie d'un objet mobile. Les paramètres de la Gaussienne (moyenne, fréquence…) contenant la valeur du pixel sont mis à jour. Si le pixel n'appartient à aucune Gaussienne, c'est la Gaussienne la moins souvent utilisée qui est mise à jour.

Des opérateurs morphologiques de dilatation et d'érosion ainsi que des critères prenant en compte les caractéristiques des zones segmentées permettent d'extraire les objets d'intérêt de la scène. A tout instant une image du fond estimé peut être obtenue en prenant pour chaque pixel la moyenne de la Gaussienne la plus probable.

3. CODAGE ET DÉCODAGE DIFFÉRENCIANT FOND ET OBJETS MOBILES

Les résultats de l'analyse sont utilisés pour encoder d'une part un flux Motion JPEG 2000 du fond estimé à intervalles réguliers (par exemple une image toutes les quatre secondes) et d'autre part un flux Motion JPEG 2000 contenant la séquence originale. Dans ce deuxième flux, les codeblocks relatifs aux régions d'intérêt sont isolés dans une première couche de qualité tandis que les autres codeblocks sont placés dans une deuxième couche de qualité. Les deux flux sont par exemple codés sans perte pour une utilisation éventuelle en justice. A qualité identique pour les deux flux, le surcoût d'encodage et de stockage est alors seulement de 1% pour une séquence à 25 images par seconde.

Lorsqu'un client avec peu de ressources,ex. un PDA, accède à la vidéo, nous lui envoyons le flux du fond et les données de la première couche du flux de la séquence en tronquant ces flux, par exemple, au même débit par image. Davantage de débit peut être alloué à la séquence comprenant deux couches. Les régions d'intérêt sont donc transmises avec une qualité supérieure (dans le cas de petites régions d'intérêt) ou égale (si les régions d'intérêt couvrent la totalité de l'image) au fond de la séquence.

Pour chaque image du fond reçue, le client décode l'image sans effectuer la transformée en ondelettes inverse. A chaque image du flux de la séquence reçue, le client décode le flux binaire. Les codeblocks contenant de l'information sont des codeblocks relatifs aux régions d'intérêt. Les codeblocks vides de toute information sont supposés ne pas avoir été transmis car faisant partie du fond de la séquence (deuxième couche de qualité). Pour les coefficients d'ondelettes de chacun de ces codeblocks vides, nous utilisons alors ceux des codeblocks de même position précédemment décodés depuis le flux Motion JPEG 2000 du fond (cf. Figure 2).

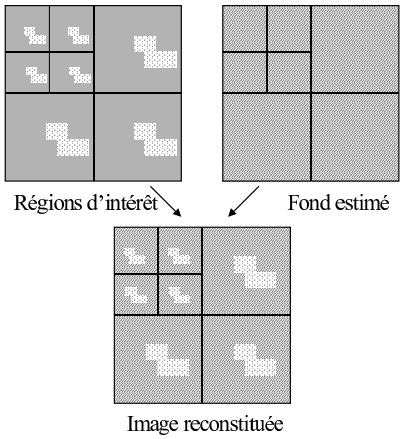


Figure 2: Reconstruction d'une image dans le domaine JPEG 2000 au décodeur

Le surcoût maximum de débit et de complexité de décodage est de 1% tandis que le gain par rapport à une séquence Motion JPEG 2000 standard à qualité identique peut atteindre, lorsqu'il n'y a pas de régions d'intérêt, un facteur 100 à la fois en débit et en complexité de décodage. Le gain est directement lié à la taille et au nombre de régions d'intérêt détectées dans la séquence.

4. RÉSULTATS

Nous avons évalué notre système sur la séquence *speedway* en CIF à 25 images par seconde du consortium Motion JPEG 2000. Le module d'analyse ne requiert qu'environ 15% du temps CPU sur un Dell Precision 3 GHz, ce qui est faible par rapport à la complexité d'encodage de la vidéo (huit fois plus). Notons que les résultats issus de cette analyse peuvent également servir à la détection automatique d'évènements anormaux en vidéo surveillance.

Un débit de 634 kbps a été alloué au flux de la séquence et 26 kps pour le flux du fond (0,25 image par seconde). La Figure 3 montre les résultats de la reconstruction de l'image 197. Le pic du rapport signal sur bruit (PRSB) sur la

séquence de 200 images est de 34,7 dB contre 30,6 dB pour Motion JPEG 2000 standard, soit un gain supérieur à 4 dB. La qualité subjective de la séquence est aussi améliorée. Le fond restitué est de meilleure qualité et est beaucoup plus robuste aux artefacts générés par le bruit d'acquisition contenu dans la séquence. En effet, les fluctuations des erreurs de compression au cours du temps n'existent plus pour le fond alors qu'elles peuvent être importantes d'une image à l'autre dans le cas de Motion JPEG 2000 standard. Même dégradé, le fond restitué fluctue donc peu au cours du temps, ce qui rend la visualisation plus naturelle et moins fatiguante pour l'opérateur.

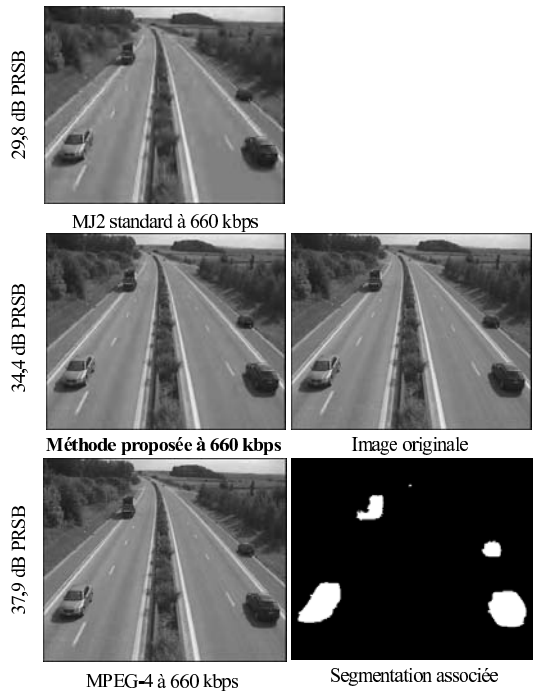


Figure 3: Comparaison de l'approche proposée à Motion JPEG 2000 (MJ2) à 660 kbps.

5. CONCLUSION

L'approche Motion JPEG 2000 basée sur des régions d'intérêt produit un fond dégradé par rapport au Motion JPEG 2000 standard. Ici, nous utilisons deux flux conformes au standard Motion JPEG 2000. L'un des deux contient la séquence d'origine encodée avec les régions d'intérêt dans sa première couche de qualité tandis que le second flux contient une estimation statistique du fond des images. Le surcoût en terme de stockage et d'encodage est négligeable tandis que le gain en terme de débit et de complexité de décodage peut être très important lorsqu'il y a peu d'activité dans la scène. Le gain en PRSB atteint 4 dB pour la séquence *speedway* à 660 kbps. La qualité visuelle de restitution de la séquence est proche celle obtenue avec un encodeur MPEG-4, tout en conservant les propriétés de scalabilité de JPEG 2000.

6. REMERCIEMENTS

Ce travail a été finance par la commission européenne dans le cadre du projet WCAM (FP6 IST-2003-507204)[7].

7. RÉFÉRENCES

[1] ISO/IEC 15444-3 "JPEG 2000 part 3: Motion JPEG 2000"

[2] V. Sanchez, A. Basu, M. Mandal, "Prioritized Region Of Interest Coding in JPEG2000", *IEEE trans. on CSVT*, vol. 14(9), pp. 1149-1155, Septembre 2004.

[3] J. Meessen, C. Parisot, C. Le Barz, D. Nicholson and J.-F. Delaigle, "WCAM: Smart Encoding for Wireless Surveillance", *Proc. of SPIE Image and Video Communications and Processing (IVCP 05)*, San Jose, USA, Janvier 2005.

[4] C. Stauffer, W.E.L. Grimson, "Adaptive Background mixture models for real-time tracking", Proc. of IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 246-252, Juin 1999.

[5] K. Kim, T. Horprasert, D. Harwood, L. Davis, "Codebook-based Background Subtraction and Performance Evaluation Methodology", 2003.

[6] X. Desurmont, C. Chaudy, A. Bastide, C. Parisot, J.F. Delaigle, B. Macq, "Image analysis architectures and techniques for intelligent systems", *IEE proc. on Vision, Image and Signal Processing, Special issue on Intelligent Distributed Surveillance Systems*, vol 152(2), pp. 224-231, Avril 2005.

[7] FP6 IST-2003-507204 WCAM, "Wireless Cameras and Audio-Visual Seamless Networking,", *http://www.ist-wcam.org*.

Codage par descriptions multiples pour la transmission vidéo

Christophe TILLIER[1], Teodora PETRISOR[1], Béatrice PESQUET-POPESCU[1], Jean-Christophe PESQUET[2]

[1]ENST, Département du Traitement du Signal et des Images
46, rue Barrault, 75634 Paris Cédex 13, FRANCE

[2]IGM et UMR-CNRS 8049, Université de Marne-la-Vallée,
5, Bd Descartes, Champs sur Marne, 77454 Marne la Vallée Cédex 2, FRANCE

{tillier,petrisor,pesquet}@tsi.enst.fr, pesquet@univ-mlv.fr

Résumé – Dans cet article, nous nous intéressons à l'utilisation du codage par descriptions multiples pour la transmission de séquences vidéo scalables obtenues avec un codeur ondelettes $t+2D$. Ces techniques sont utiles pour des réseaux commutés, à perte de paquets. Contrairement aux approches existantes, la redondance n'est pas introduite dans le domaine spatial mais dans le domaine temporel, grâce à des bancs de filtres dyadiques sur-échantillonnés, compensés en mouvement. Nous fournissons ici un cadre de travail pour construire et étudier des schémas de codage par descriptions multiples temporels, basés sur des décompositions en ondelettes redondantes avec un facteur de redondance réduit, et nous validons cette approche par une application au codage vidéo.

Abstract – In this paper, we address the problem of multiple description coding for the transmission over best-effort networks of scalable video sequences obtained with a $t+2D$ wavelet codec. Unlike previous approaches, which consider redundancy insertion in the spatial domain, in this paper we introduce redundancy in the temporal domain of the video sequence, using motion compensated oversampled dyadic filter banks. We are giving here the framework for building such temporal multiple description schemes based on redundant wavelet decompositions with a reduced redundancy feature. The general approach is applied to video coding.

1 Introduction

En codage vidéo scalable, un intérêt croissant s'est manifesté, au cours de ces dernières années, pour des structures compensées en mouvement utilisant la décomposition de Haar ou des bancs de filtres biorthogonaux 5/3 opérant temporellement [1], [2]. Dans les applications de communication, les représentations scalables sont utiles pour l'adaptation du flux binaire aux variations de bande passante ou aux caractéristiques du récepteur, mais dans le cas de transmission sur des réseaux « best effort », une difficulté nouvelle dans la reconstruction est due aux pertes de paquets.

Le codage par descriptions multiples (MDC) [3], qui est étroitement lié au codage conjoint source-canal, permet de répondre à ce besoin en introduisant de la redondance au niveau de la source pour gérer les pertes de paquets. Cette technique vise à construire des descriptions corrélées qui sont transmises sur des canaux indépendants tout-ou-rien. Dans le cas d'échecs de transmission sur certains canaux, des décodeurs dits « latéraux » doivent être capables de reconstruire la source avec une qualité acceptable. En revanche, la réception de toutes les descriptions doit permettre une reconstruction de qualité supérieure, qui est obtenue en sortie du décodeur dit « central ».

Dans ce travail, nous explorons des schémas de descriptions multiples basés sur une décomposition temporelle en ondelettes redondantes opérant dans un codeur $t+2D$. Notons que la plupart des travaux précédents effectuait la construction des descriptions multiples d'une vidéo à l'aide de plusieurs boucles de prédiction temporelle dans un schéma de codage hybride, [4], [5]. Pour les codeurs hybrides un état de l'art récent des techniques de MDC en vidéo est présenté dans [6].

Il existe aussi des schémas MDC qui emploient des codeurs en ondelettes 3D. Néanmoins, contrairement à d'autres approches, où le banc de filtres en ondelettes était appliqué indépendamment dans le domaine spatial pour chaque image [7], [8], ou sur les trames résultant d'une décomposition $t+2D$ [9], notre approche concerne l'utilisation de bancs de filtres sur-échantillonnés dans la décomposition temporelle. Des travaux précédents [10], [11] ont introduit de la redondance temporelle dans un banc de filtres 3-bandes. Ici nous proposons plusieurs possibilités de construction des descriptions corrélées à partir d'un banc de filtres dyadique sur-échantillonné, appliqué dans la direction temporelle d'une séquence vidéo traitée avec un codeur $t+2D$. Les propriétés de tels banc de filtres agissant comme des codeurs de canal sur des canaux à effacements ont récemment été explorées dans [12], [13], [14].

Nous nous différencions de ces méthodes en imposant une particularité supplémentaire à nos schémas, qui est un facteur de redondance réduit. Nous obtenons cette propriété en effectuant un sous-échantillonnage d'un facteur 2 des trames de détails temporels après la décomposition de base. Cela est équivalent à dire que nos schémas ont une redondance (en nombre de coefficients d'ondelettes) égale à la taille d'une sous-bande d'approximation temporelle. Pourtant, dans ce cas la reconstruction parfaite n'étant plus garantie, nous sommes amenés à étudier l'inversibilité du décodeur central. De plus, ceci permet aussi d'analyser les effets du bruit de quantification sur la reconstruction.

Dans le paragraphe suivant, nous présentons le banc de filtres considéré ainsi que les schémas MDC proposés. Dans le paragraphe 3 nous considérerons l'application de ces schémas au codage vidéo, puis nous fournissons des résultats de simulation dans le paragraphe 4. Enfin, nous concluons dans le para-

graphe 5.

2 Représentations par bancs de filtres redondants

En considérant un banc de filtres dyadique associé à une décomposition en ondelettes nous allons générer une structure redondante, à partir de laquelle plusieurs configurations de descriptions multiples sont possibles. Nous allons analyser et comparer ces configurations et proposer une solution qui conserve la propriété de reconstruction parfaite, tout en réduisant la redondance.

Soit $(x_n)_{n\in\mathbb{Z}}$, un signal d'entrée temporel et notons $(h_n)_{n\in\mathbb{Z}}$ (resp. $(g_n)_{n\in\mathbb{Z}}$) le filtre passe-bas (resp. passe-haut) d'un banc de filtres d'analyse dyadique à reconstruction parfaite. Les expressions des coefficients d'approximation et de détails sont alors

$$a_n^{\mathrm{I}} = \sum_k h_{2n-k}\, x_k \tag{1}$$

$$d_n^{\mathrm{I}} = \sum_k g_{2n-k}\, x_k. \tag{2}$$

En sous-échantillonnant aux instants impairs plutôt qu'aux instants pairs, les coefficients d'approximation et de détails deviennent

$$a_n^{\mathrm{II}} = \sum_k h_{2n-1-k}\, x_k \tag{3}$$

$$d_n^{\mathrm{II}} = \sum_k g_{2n-1-k}\, x_k. \tag{4}$$

En considérant ces 4 ensembles de coefficients, nous avons naturellement une décomposition redondante, le nombre des coefficients étant multiplié par 2.

Il est toutefois possible de construire des représentations plus « économiques » tout en gardant la propriété souhaitée de reconstruction parfaite. Pour ce faire, nous décimons les séquences précédentes d'un facteur 2, et nous introduisons alors les notations $\hat{a}_n^{\mathrm{I}} = a_{2n}^{\mathrm{I}}$ et $\check{a}_n^{\mathrm{I}} = a_{2n-1}^{\mathrm{I}}$, des notations similaires étant utilisées pour les séquences de coefficients de détails vues précédemment.

Ainsi, le vecteur $\overline{\mathbf{c}}_n$ qui contient toutes les séquences sous-échantillonnées possibles est :

$$\overline{\mathbf{c}}_n = \left(\hat{a}_n^{\mathrm{I}}\ \check{a}_n^{\mathrm{I}}\ \hat{a}_n^{\mathrm{II}}\ \check{a}_n^{\mathrm{II}}\ \hat{d}_n^{\mathrm{I}}\ \check{d}_n^{\mathrm{I}}\ \hat{d}_n^{\mathrm{II}}\ \check{d}_n^{\mathrm{II}}\right)^{\top}.$$

Le système proposé est facilement décrit en utilisant une décomposition polyphase. Introduisons les composantes polyphases 4-bandes des filtres d'analyse :

$$h_i(n) = h_{4n-i}$$
$$g_i(n) = g_{4n-i}$$

où $i \in \{0,\ldots,3\}$, ainsi que les transformées en z correspondantes $H_i(z)$ et $G_i(z)$.

De la même manière, nous définissons les quatre composantes polyphases du signal d'entrée $x_n^{(i)} = x_{4n+i}$, $i \in \{0,\ldots,3\}$, et le vecteur des composantes polyphases correspondant :

$$\mathbf{x}_n = \left(x_n^{(0)}\ x_n^{(1)}\ x_n^{(2)}\ x_n^{(3)}\right)^{\top}$$

Les équations (1) - (4) peuvent se ré-écrire à l'aide de la représentation polyphase comme suit :

$$\overline{\mathbf{C}}(z) = \overline{\mathbf{M}}(z)\ \mathbf{X}(z) \tag{5}$$

où $\overline{\mathbf{C}}(z)$ et $\mathbf{X}(z)$ sont les transformées en z de $(\overline{\mathbf{c}}_n)_{n\in\mathbb{Z}}$ et de $(\mathbf{x}_n)_{n\in\mathbb{Z}}$. $\overline{\mathbf{M}}(z)$ est la matrice de transfert polyphase globale.

$$\overline{\mathbf{M}}(z) = \begin{bmatrix} H_0(z) & H_1(z) & H_2(z) & H_3(z) \\ H_2(z) & H_3(z) & H_0(z)z^{-1} & H_1(z)z^{-1} \\ H_1(z) & H_2(z) & H_3(z) & H_0(z)z^{-1} \\ H_3(z) & H_0(z)z^{-1} & H_1(z)z^{-1} & H_2(z)z^{-1} \\ G_0(z) & G_1(z) & G_2(z) & G_3(z) \\ G_2(z) & G_3(z) & G_0(z)z^{-1} & G_1(z)z^{-1} \\ G_1(z) & G_2(z) & G_3(z) & G_0(z)z^{-1} \\ G_3(z) & G_0(z)z^{-1} & G_1(z)z^{-1} & G_2(z)z^{-1} \end{bmatrix}.$$

Des schémas de faible redondance s'obtiennent, par exemple, en éliminant 2 des 8 composantes de $\overline{\mathbf{c}}_n$. Néanmoins, un choix judicieux des séquences omises est nécessaire, pour garder une distorsion latérale acceptable. Pour ce faire, tous les schémas que nous construisons conservent les sous-bandes d'approximations issues des deux décompositions, et éliminent seulement la moitié des composantes polyphases des sous-bandes de détails.

Le vecteur $\mathbf{c}_n$ résultant correspond à un facteur de redondance égal à $3/2$. Le banc de filtres sur-échantillonné résultant est tel que $\mathbf{C}(z) = \mathbf{M}(z)\ \mathbf{X}(z)$, où $\mathbf{C}(z)$ est la transformée en z de $(\mathbf{c}_n)_{n\in\mathbb{Z}}$ et $\mathbf{M}(z)$ est la matrice de transfert polyphase du schéma considéré.

Plusieurs décompositions possibles peuvent malgré tout être obtenues suivant le choix de la sous-matrice $\mathbf{M}(z)$. Nous nous sommes intéressés plus particulièrement aux 4 solutions suivantes, pour lesquelles nous spécifions également la manière de construire les 2 descriptions.

- **Schéma R :** ce schéma consiste à séparer les coefficients de détails issus d'une analyse classique monodimensionnelle (à sous-échantillonnage critique) en deux groupes : les coefficients d'indice pair et les coefficients d'indice impair, chacun des groupes appartenant à l'une des descriptions. Les coefficients d'approximation sont simplement dupliqués, d'où

$$\mathbf{c}_n = (\underbrace{\hat{a}_n^{\mathrm{I}}\ \check{a}_n^{\mathrm{I}}\ \hat{d}_n^{\mathrm{I}}}_{1^e \text{ description}}\ \underbrace{\hat{a}_n^{\mathrm{I}}\ \check{a}_n^{\mathrm{I}}\ \check{d}_n^{\mathrm{I}}}_{2^e \text{ description}})^{\top}.$$

- **Schéma D1 :** on distribue les coefficients de détail de la même manière que dans le schéma précédent, mais dans la seconde description, au lieu de répéter les coefficients d'approximation de la première description, on utilise les coefficients d'approximation de la seconde base. Cela correspond à

$$\mathbf{c}_n = (\underbrace{\hat{a}_n^{\mathrm{I}}\ \check{a}_n^{\mathrm{I}}\ \hat{d}_n^{\mathrm{I}}}_{1^e \text{ description}}\ \underbrace{\hat{a}_n^{\mathrm{II}}\ \check{a}_n^{\mathrm{II}}\ \check{d}_n^{\mathrm{I}}}_{2^e \text{ description}})^{\top}.$$

- **Schéma D2 :** on ajoute plus de « diversité » dans le choix des coefficients de détail en prenant en compte les coefficients d'indices pairs de la seconde base :

$$\mathbf{c}_n = (\underbrace{\hat{a}_n^{\mathrm{I}}\ \check{a}_n^{\mathrm{I}}\ \hat{d}_n^{\mathrm{I}}}_{1^e \text{ description}}\ \underbrace{\hat{a}_n^{\mathrm{II}}\ \check{a}_n^{\mathrm{II}}\ \hat{d}_n^{\mathrm{II}}}_{2^e \text{ description}})^{\top}.$$

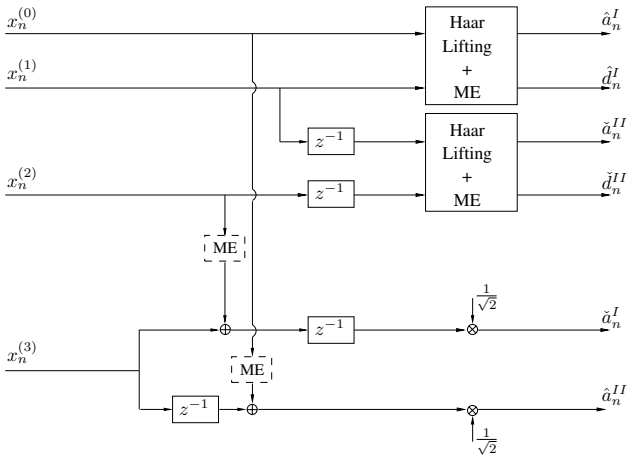


FIG. 1 – Schéma en lifting 4-bandes redondant (l'estimation de mouvement optionnelle est indiquée en pointillés).

– **Schéma D3 :** en sélectionnant dans la seconde base les coefficients de détails d'indices impairs plutôt que ceux d'indices pairs, on obtient alors

$$\mathbf{c}_n = (\underbrace{\hat{a}_n^{\mathrm{I}}\ \check{a}_n^{\mathrm{I}}\ \hat{d}_n^{\mathrm{I}}}_{1^e \text{ description}}\ \underbrace{\hat{a}_n^{\mathrm{II}}\ \check{a}_n^{\mathrm{II}}\ \check{d}_n^{\mathrm{II}}}_{2^e \text{ description}})^{\mathsf{T}}.$$

En l'absence de quantification, les deux premiers schémas sont évidemment inversibles puisqu'ils contiennent tous les coefficients obtenus à partir d'une analyse à sous-échantillonnage critique classique. Pour les 2 autres schémas, l'inversibilité n'est pas garantie *a priori*.

Reconstruire le signal à partir des représentations redondantes proposées, revient en termes de MDC à définir le schéma de synthèse utilisé par le décodeur central. Avec les notations précédentes, le problème est de trouver une matrice de transfert $\mathbf{W}(z)$ polynomiale de dimension $N \times K$ telle que :

$$\mathbf{W}(z)\mathbf{M}(z) = \mathbf{I}_{N\times N} \tag{6}$$

où

$$\mathbf{W}(z) = [W_{i,j}(z)]_{1\le i\le N, 1\le j\le K}$$

et

$$\mathbf{M}(z) = [M_{i,j}(z)]_{1\le i\le K, 1\le j\le N}$$

avec $K = 6$ et $N = 4 < K$. Ce problème n'ayant pas de solution unique, nous avons mis en évidence une solution optimale (minimisant les effets dûs aux bruits de quantification) par une approche de type pseudo-inverse [15].

3 Application au codage vidéo

Une difficulté supplémentaire inhérente au codage vidéo est la prise en compte de l'estimation/compensation de mouvement.

Dans chaque description une paire de coefficients d'approximation et de détail est obtenue par un étage lifting classique. L'application d'une transformée compensée en mouvement pour cette paire se fait naturellement [16]. En revanche, le coefficient d'approximation d'indice complémentaire apparaît seul, suite à l'élimination du coefficient de détail correspondant. Ceci est illustré dans la figure 1.

Pour le calcul de ce coefficient, deux stratégies différentes peuvent être envisagées, suivant que l'on fasse intervenir ou non une compensation de mouvement. En introduisant cette compensation de mouvement (en pointillés sur la figure 1), on augmente un peu la redondance car on a à coder un champ de vecteurs de mouvement supplémentaire. Pourtant, ce champ permet d'obtenir une sous-bande d'approximation très proche d'une image naturelle, ce qui facilite son codage. En plus, lors du décodage latéral l'efficacité de reconstruction est aussi augmentée.

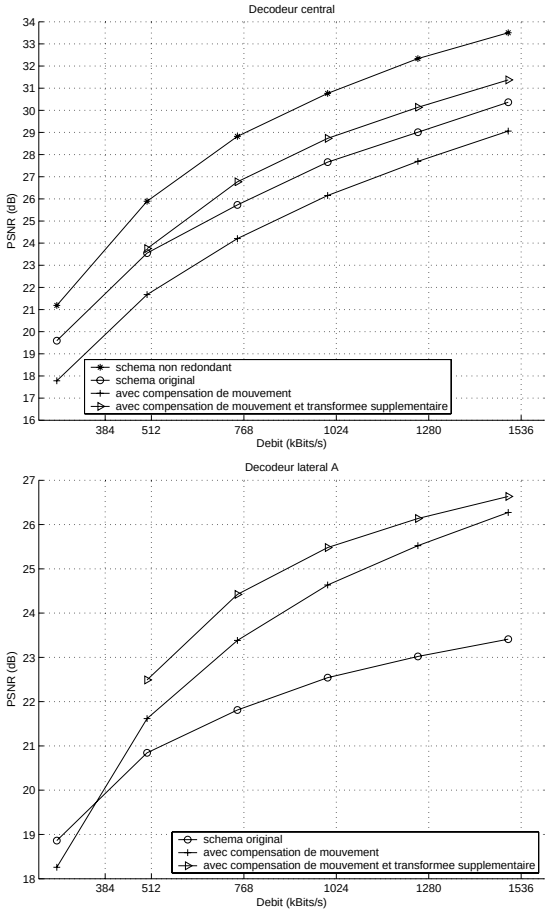


FIG. 2 – Courbes débit-distorsion comparant les différentes stratégies d'introduction de compensation de mouvement dans le schéma D3 pour le décodeur central (gauche) et un décodeur latéral (droite) pour la séquence « MOBILE » (CIF à 30fps)

Après une étude sur la corrélation temporelle des deux sous-bandes d'approximation d'une description, une troisième stratégie s'est avèrée intéressante. Elle consiste à appliquer une nouvelle analyse temporelle de Haar sur ces approximations pour augmenter l'efficacité du codage.

4 Simulations et résultats

Parmi les 4 schémas étudiés, R et D2 donnent de moins bonnes performances que D1 et D3 en terme de puissance du bruit de quantification dans la séquence reconstruite, et nous ne nous attarderons donc pas à présenter leurs résultats. Dans les simulations, nous effectuons $J = 3$ niveaux de décomposition temporelle avec la transformée de Haar compensée en mouvement [17]. Au dernier niveau, nous utilisons l'un des 2 schémas

analysés D1 ou D3. La redondance globale est $1 + 2^{-J}$, les trames de détails pour les résolutions $j < J$ étant distribuées de manière équitable entre les 2 descriptions. Enfin, pour les décodeurs latéraux, une approche par pseudo-inverse a aussi été développée pour la reconstruction optimale des descriptions.

Les schémas proposés ont été testés sur plusieurs séquences CIF à 30fps. Dans un premier temps, on compare les schémas D1 et D3 en faisant une décomposition temporelle sur trois niveaux, compensée en pel entier, seulement sur les deux premiers et, pour l'instant, sans aucune estimation de mouvement sur le niveau ou nous avons introduit la redondance. Les sous-bandes obtenues sont décomposées spatialement avec des ondelettes biorthogonales 9/7 puis codées en utilisant l'algorithme MC-EZBC [18].

Cette comparaison en termes de performances débit-distorsion est présentée dans le tableau 1 pour les décodeurs centraux et latéraux. Comme préconisé par la théorie, le décodeur central du schéma D1 donne de meilleurs résultats que celui de D3. Le décodeur latéral noté A (correspondant à la première description) dans le tableau 1 est identique pour les 2 schémas, alors que pour le décodeur noté B (description 2), une dissymétrie de construction dans le schéma D1 explique l'écart de performances.

TAB. 1 – Comparaison débit-distorsion : PSNR (dB) à différents débits (Kbs), pour les séquences « FOREMAN » et « MOBILE » (CIF à 30fps) sur 3 niveaux de décomposition en ondelettes.

« FOREMAN » schéma D1						
débit	250	500	750	1000	1500	3000
central	29.48	32.19	33.85	34.98	36.85	40.53
latéral A	26.05	27.20	27.78	28.13	28.66	29.51
latéral B	24.32	24.84	25.06	25.16	25.29	25.43
« FOREMAN » schéma D3						
débit	250	500	750	1000	1500	3000
central	29.27	32.01	33.68	34.79	36.68	40.39
latéral A	26.05	27.20	27.78	28.13	28.66	29.51
latéral B	25.26	26.16	26.62	26.88	27.28	27.96
« MOBILE » schéma D1						
débit	250	500	750	1000	1500	3000
central	19.89	22.18	23.54	24.88	26.55	30.61
latéral A	18.96	20.15	20.78	21.24	21.91	23.07
latéral B	18.36	19.24	19.70	19.93	20.37	20.90
« MOBILE » schéma D3						
débit	250	500	750	1000	1500	3000
central	19.70	21.95	23.33	24.61	26.33	30.43
latéral A	18.96	20.15	20.78	21.24	21.91	23.07
latéral B	18.81	19.83	20.35	20.72	21.28	22.23

Dans la suite on augmente l'efficacité du codage en passant à une compensation au $8^{ème}$ de pel sur tous les niveaux temporels et nous comparons les résultats des trois stratégies décrites dans la section 3 sur le schéma D3, qui est plus équilibré. Ces résultats se trouvent dans la figure 2. Nous pouvons remarquer que l'introduction de compensation de mouvement pour le $2^{ème}$ coefficient d'approximation améliore considérablement les performances des décodeurs latéraux, mais diminue les performances du décodeur central. Cet inconvénient est résolu avec la nouvelle transformée de Haar appliquée sur les coefficients d'approximation.

5 Conclusions

Dans cet article, nous avons présenté plusieurs schémas de codage par descriptions multiples temporelles basés sur des décompositions en ondelettes redondantes avec un facteur de redondance réduit. Le facteur de redondance est réglable à l'aide du nombre de niveaux de décomposition temporelle effectué. Nous avons étudié la reconstruction de ces schémas, optimale en termes d'effet du bruit de quantification. Les schémas étudiés ont été appliqués au codage vidéo scalable et robuste et nous avons constaté que, dans ce cas, la compensation de mouvement joue un rôle essentiel.

Références

[1] J. Xu, Z. Xiong, S. Li, and Y. Zhang, "Three-Dimensional Embedded Subband Coding with Optimized Truncation (3D-ESCOT)," *Applied and Computational Harmonic Analysis*, vol. 10, pp. 290–315, 2001.

[2] B.-J. Kim, Z. Xiong, and W.A. Pearlman, "Very low bit-rate embedded video coding with 3-D set partitioning in hierarchical trees (3D-SPIHT)," *IEEE Trans on Circ. and Syst. for Video Tech.*, vol. 8, pp. 1365–1374, 2000.

[3] V.K. Goyal, *Beyond Traditional Transform Coding*, Ph.D. thesis, Univ. California, Berkeley, 1998.

[4] A. R. Reibman, H. Jafarkhani, Y. Wang, M. T. Orchard, and R. Puri, "Multiple-description video coding using motion-compensated temporal prediction," *IEEE Trans. on Circ. and Syst. for Video Technology*, vol. 12(3), pp. 193–204, March 2002.

[5] Y. Wang and S. Lin, "Error-resilient video coding using multiple description motion compensation," *IEEE Trans. on Circ. and Syst. for Video Technology*, vol. 12(6), pp. 438–452, June 2002.

[6] Y. Wang, A.R. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proceedings of the IEEE*, vol. 93(1), pp. 57–70, January 2005.

[7] I. V. Bajic and J. W. Woods, "Domain-based multiple description coding of images and video," *IEEE Trans. on Image Proc.*, vol. 12(18), pp. 1211–1225, October 2003.

[8] M. Fumagalli, R. Lancini, and A. Stanzione, "Video transmission over ip by using polyphase downsampling multiple description coding," *Proc. IEEE Int. Conf. on Multimedia and Expo*, August 2001, pp. 1095–1098.

[9] M. Pereira, M. Antonini, and M. Barlaud, "Multiple description coding for internet video streaming," *Proc. IEEE Int. Conf. Image Processing*, 2003.

[10] B. Pesquet-Popescu C. Tillier and M. van der Schaar, "Multiple descriptions scalable video coding," *Proc. of EUSIPCO*, Sept. 2004.

[11] M. van der Schaar and D.S. Turaga, "Multiple description scalable coding using wavelet-based motion compensated temporal filtering," *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Sept. 2003.

[12] V.K. Goyal, J. Kovačević, and Martin Vetterli, "Quantized frame expansions as source-channel codes for erasure channels," *Proc. IEEE Data Compression Conf.*, pp. 326–335, March 1999.

[13] J. Kovačević, P.L. Dragotti, and V. Goyal, "Filter bank frame expansions with erasures," *IEEE Trans. on Inf. Theory*, vol. 48, no. 6, pp. 1439–1450, June 2002.

[14] R. Motwani and C. Guillemot, "Tree-structured oversampled filterbanks as joint source-channel codes : application to image transmission over erasure channels," *IEEE Transactions on Signal Processing*, vol. 52 (9), pp. 2584 – 2599, Sept. 2004.

[15] T. Petrisor, C. Tillier, B. Pesquet-Popescu, and J.-C. Pesquet, "Temporal multiple description schemes for scalable video using wavelet frames," en préparation.

[16] T. Petrisor, C. Tillier, B. Pesquet-Popescu, and J.-C. Pesquet, "Redundant multiresolution analysis for multiple description video coding," *International Workshop on Multimedia Signal Processing, Sienne, Italie*, 2004.

[17] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, Salt Lake City, UT, May 2001.

[18] "3D MC-EZBC software package," disponible sur MPEG CVS.

Compression de maillages 3D de grande résolution par transformée en ondelettes au « fil de l'eau »

Akram Elkéfi, Sami Abbadi, Marc Antonini, Chokri Ben Amar

Laboratoire I3S, CNRS – UMR 6070, Université de Nice-Sophia Antipolis

Bât. Algorithmes/Euclide B,BP 121, 2000 route des lucioles, 06903 Sophia Antipolis Cedex, France

Tél. : 33(0)4.92.94.27.21 – Fax : 33(0)4.92.94.28.98

{elkefi,abbadi,am}@i3s.unice.fr

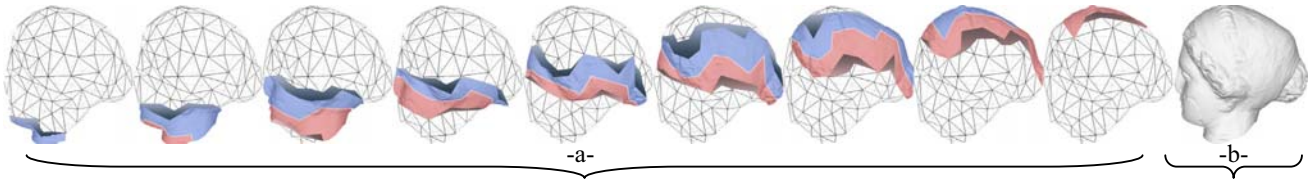


Fig. 1. *Visualisation du processus de compression au fil de l'eau d'un maillage 3D multirésolution : (a) L'acquisition spirale envoie un triangle avec son voisinage pour initialiser la transformée puis le codage. Les étapes suivantes correspondent à l'évolution suivant une lecture spirale. (b) Reconstruction avec tous les niveaux de résolution (détails).*

Résumé – Dans ce papier, nous nous intéressons à la compression dans le domaine des ondelettes de très grands maillages au fur et à mesure de leur acquisition. Nous proposons une méthode de compression progressive utilisant une transformée en ondelettes au "fil de l'eau" pour les maillages 3D semi-réguliers. Cette méthode consiste à faire le traitement au fur et à mesure de l'acquisition des données tout en réduisant considérablement l'espace mémoire utilisé. Les résultats expérimentaux montrent que la méthode est très efficace en terme de coût calculatoire, espace et accès mémoire. Les rapports débit/distorsion de notre méthode au fil de l'eau sont très proches de ceux de la méthode avec chargement de l'objet en totalité avant traitement, mais présente l'avantage de nécessiter un espace mémoire beaucoup plus réduit.

Abstract *– In this paper we introduce a method allowing to perform a Scan-based Wavelet Compression of 3D semi-regular multiresolution meshes. This method consists in processing the data progressively during the acquisition while reducing considerably memory usage. The experiments show that the method is very effective in terms of memory usage and access. Moreover the proposed algorithm allows reducing the complexity of processing from O(N²) to O(N).*

1. Introduction

Les maillages constituent un outil puissant pour modéliser des objets 3D complexes grâce à leur double nature géométrique et combinatoire (les positions des sommets et la connectivité). Bien que de nombreuses alternatives existent pour la modélisation de formes surfaciques, les maillages sont aujourd'hui omniprésents et des efforts considérables sont développés pour le traitement numérique de la géométrie opérant essentiellement sur des maillages triangulaires [7]. Le problème de traitements au « fil de l'eau » se pose alors lorsqu'il s'agit de comprimer de très grands volumes de données avec un minimum de ressources mémoire [2], [5], [8]. En effet, la transformée en ondelettes des maillages 3D multirésolutions nécessite l'acquisition et le chargement complet de l'objet en mémoire avant son traitement. Sachant que les maillages 3D de grande précision ont des tailles énormes dépassant plusieurs millions de points et que les systèmes de traitement ont des capacités mémoire limitées, très vite se pose le problème de traitement lié à ce type de données.

Dans cet article, nous nous intéressons à la compression dans le domaine ondelettes de très grands maillages au fur et à mesure de leur acquisition. La transformée en ondelettes d'un signal est obtenue par une convolution du signal d'entrée par un ou plusieurs filtres de longueur finie. Le fait que ces filtres aient un support fini permet de calculer les coefficients d'ondelettes à l'aide d'un nombre fini d'échantillons du signal. Les méthodes de calcul de la transformée en ondelettes au « fil de l'eau » utilisent cette propriété pour calculer les coefficients haute et basse fréquences de la transformée avec une quantité de mémoire minimale. Nous proposons en premier lieu ici un algorithme de calcul de la transformée au « fil de l'eau » mise en oeuvre par schéma lifting. Cette méthode consiste à effectuer un traitement local de l'objet en fonction du mode d'acquisition 3D considéré tout en contraignant le coût mémoire à être minimal. La transformée en ondelettes générée est identique à celle que l'on obtiendrait si l'on avait la connaissance de la totalité de l'objet 3D à l'instant t.

Par la suite nous proposons une méthode de compression progressive utilisant la transformée en ondelettes au « fil de l'eau » pour les maillages 3D semi -régulier. Cette méthode consiste à faire le traitement au fur et à mesure de l'acquisition des données tout en réduisant considérablement l'espace mémoire utilisé.

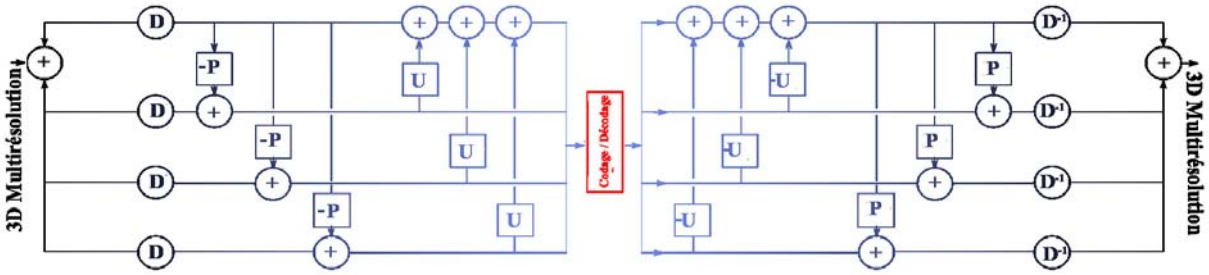


Fig. 3. *Schéma Lifting d'analyse et synthèse à quatre canaux correspondant au filtre de Butterfly.*

2. Ondelettes sur maillages 3D

L'idée principale de l'analyse multirésolution sur un maillage 3D semi-régulier (cf. figure 2) est la représentation hiérarchique du maillage par une sous-bande basse résolution plus simple et une collection de détails géométriques appelées coefficients d'ondelettes, nécessaires pour retrouver la fonction originale. La transformée en ondelettes réputée pour sa forte décorrélation est plus facile à implémenter en utilisant un schéma Lifting. En effet, le calcul est rapide, efficace et la transformée inverse est directement déduite [3]. Dans le cas du Butterfly, le signal d'entrée est sous-échantillonné et divisé en 4 ensembles disjoints au moyen d'une transformée polyphase. Ces sous-signaux seront prédits puis mis à jour pour la réduction de l'écart entre la fonction originale et ses prédictions. Pour un maillage multirésolution, l'application du schéma lifting est répétée n−1 fois successivement sur le signal résultant, n étant le nombre niveau de résolution (cf. figure 3).

La topologie de l'objet étant aléatoire, le filtre doit suivre la forme de l'objet en collant au maillage. L'extraction des points de filtrage est possible grâce à une méthode de parcours du maillage de voisin en voisin. L'étape de calcul du voisinage est très coûteuse en nombre d'opérations. La méthode de calcul classique est de complexité $O(N^2)$. Et vue que les maillages utilisés comptent des millions de triangles cette méthode classique coûtera au moins quatre heures de calcul sur un P4 à 2Ghz. La procédure de recherche de voisinage proposée effectue ce calcul en une seconde maximum, elle est donc 10^4 fois plus rapide que la méthode « classique » (Les détails de la méthode de calcul de voisinage sont donnés dans [9]). Notre approche permet de réduire cette complexité de $O(N^2)$ à $O(N)$. En effet, au lieu de rechercher les voisins dans tout le maillage, nous établissons une procédure de localisation de la plus petite surface les englobant (pour plus de détails voir [4]).

3. Implémentation au « fil de l'eau »

3.1 Acquisition spirale

Pour pouvoir traiter un objet au fur et à mesure de son acquisition, il faut le diviser en zones puis envoyer à chaque instant (ou demande) la zone nécessaire pour la transformée, puis la quantification et le codage. Dans le cas d'images 2D, la décomposition de l'objet est simple, il suffit de diviser l'image en ligne ou en blocs [8]. Pour la vidéo, l'unité est un GOP (Group Of Images) [2]. Dans ces deux cas la division est plus évidente que pour un objet 3D semi-régulier. L'idée que nous proposons dans le cas de maillages semi-réguliers multiéchelles est de prendre pour unité un triangle de basse fréquence. Pour ce, il faut disposer d'un outil permettant un déplacement orienté sur le maillage irrégulier de sorte que le parcours de l'objet en totalité ne laisse pas de parties non traitées, ce qui est délicat, vu que la topologie de l'objet est très variée.

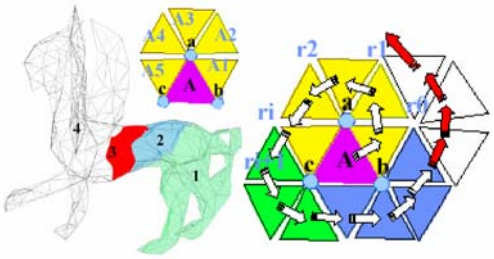


Fig. 4. *Méthode du parcours du maillage.*

Nous proposons ici d'utiliser un balayage (« scanning ») spiral pour permettre de parcourir l'objet 3D dans sa totalité [8]. Le choix du parcours influe directement sur la taille maximale de la mémoire utilisée. (L'influence de la méthode du parcours du maillage sur la mémoire nécessaire est détaillée dans [9]). Lors de l'application de la compression au fil de l'eau, nous avons constaté que la méthode d'acquisition spirale (figure 4) est la moins coûteuse en mémoire. Elle assure à chaque instant t, la présence de la totalité du voisinage d'un triangle partiellement traité nécessaire pour compléter son traitement et son effacement de la mémoire. L'algorithme spiral suivra le même sens d'acquisition du scanner.

3.2 Principe de la méthode

La méthode proposée se compose de deux parties essentielles : La transformée en ondelettes au fil de l'eau et le codage au fil de l'eau. L'acquisition spirale envoie un lot de triangles de basses fréquences avec tous leurs détails vers la transformée en ondelettes. Cette dernière en traite une partie et laisse le reste en attente d'autres informations nécessaires. La partie traitée est envoyée par la suite à un tampon. Lorsque ce dernier atteint un niveau choisit déterminant la taille maximale de mémoire utilisée, il envoie tous ses triangles vers la quantification et le codage. Les différentes étapes de la méthode proposée sont présentées dans la fig. 5 :

- *Acquisition spirale* : L'entrée étant l'objet 3D, l'acquisition spirale est celle responsable de l'envoi de

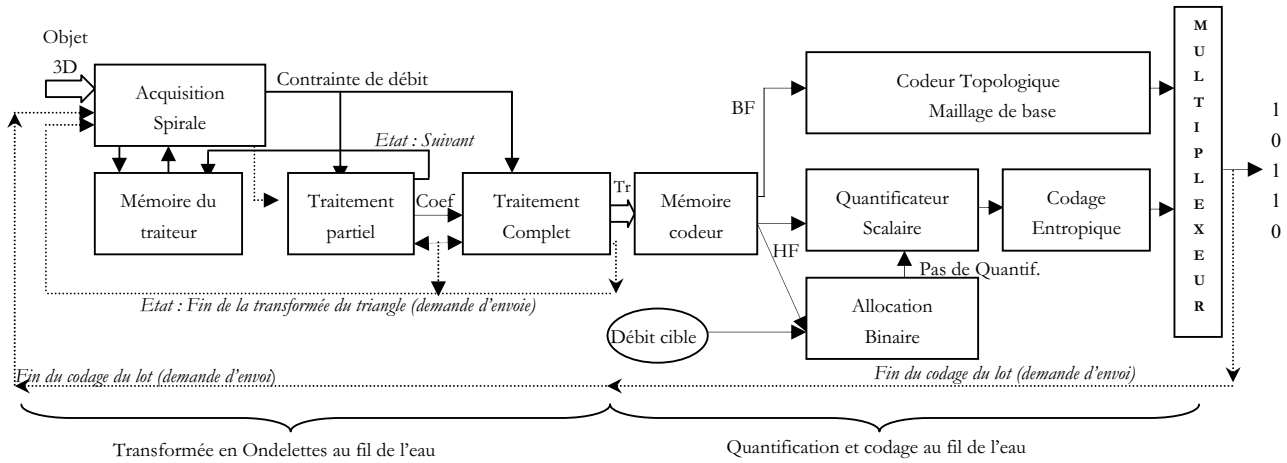


Fig. 5. *Schéma de la méthode proposée de compression au fil de l'eau des maillages 3D*

triangles aux autres blocs. Le rendement de la méthode est fortement lié à la manière d'acquérir les données.

- *Mémoire du traiteur* : C'est la mémoire réservée par le système d'exploitation pour la transformée.
- *Traitement partiel* : pour pouvoir appliquer la transformée en ondelettes en totalité sur un triangle, il faut connaître tout son voisinage, ce qui n'est pas possible car nous ne disposons pas toujours des données voulues à chaque instant t. Une phase partielle est donc nécessaire pour profiter de la présence de quelques voisins du triangle à traiter et ainsi faire un précalcul. A chaque fois qu'un nouveau triangle se trouve en mémoire, il déclenche la poursuite des traitements partiels dont il fait partie et commence sa transformée partielle. Le traitement partiel d'un triangle de basse fréquence implique le traitement complet de toute sa descendance (tous les niveaux de résolutions) sauf les triangles de bords de ce même triangle (pour plus de détails voir [4]).
- *Traitement total* : Lorsqu'un triangle n'est plus en attente, il est pris en charge par le bloc de traitement complet qui achève sa transformée, envoie ses coefficients puis l'efface de la mémoire. Lorsque toutes les attentes d'un triangle sont présentes en mémoire, l'unité de Traitement complet effectue la transformée en totalité.
- *Allocation binaire* : Il permet de calculer les pas de quantification optimaux qui minimisent l'erreur de reconstruction lors du codage en fonction d'un débit de consigne. Cette étape est répétée à chaque réception d'un nouveau lot de triangles envoyé par la transformée en ondelettes au fil de l'eau.
- *Quantification* : Les données géométriques de chaque lot de triangles sont quantifiées avec les pas calculés lors de l'allocation binaire. Nous avons choisi une quantification scalaire uniforme.
- *Codage entropique* : Un codage entropique des sous-bandes de coefficients quantifiés permet d'obtenir la trame binaire. C'est un codeur arithmétique contextuel par plans de bits qui exploite la corrélation spatiale intra et inter- résolution des coefficients d'ondelettes 3D à l'aide de contextes originaux adaptés à la géométrie des maillages. L'objet compressé est alors l'ensemble de toutes les trames binaires de tous les lots de triangles.

3.3 ETAT DE LA MEMOIRE

Vu que tout triangle partiellement traité est gardé en mémoire jusqu'à la fin de son traitement, il faut faire en sorte que les triangles attendus soient acquis en priorité pour achever le traitement des attentes et ainsi éviter de gonfler la mémoire. La méthode de lecture spirale proposée assure qu'à chaque acquisition, le traitement d'au moins un triangle est achevé et le triangle correspondant est ainsi libéré de la mémoire. Nous estimons que cette méthode influant directement sur la taille du Buffer, est efficace pour réduire au maximum l'utilisation de la mémoire. En effet, notre méthode assure à chaque tour spiral qu'un certain nombre de triangles soit totalement traités et effacés de la mémoire et que les nouveaux soient partiellement traités, pour une utilisation minimale de la mémoire.

La figure 6 illustre une estimation de la taille de mémoire maximale utilisée dans le cas du fil de l'eau en comparaison avec celle nécessaire si tout l'objet était conservé en mémoire. La figure 4 représente un exemple de l'état de la mémoire au cours d'une étape de la transformée : La partie en noir (4) n'est pas encore scannée, celle en vert (1) est déjà traitée et effacée de la mémoire et celle en bleu (2) est en cours de traitement total et va être effacée avant le tour spiral suivant. La partie en rouge (3) est partiellement traitée, gardée en mémoire, et est nécessaire pour le traitement en cours de la partie en bleu (2). A une étape donnée, seulement la partie partiellement traitée (3) est gardée en mémoire, d'où notre gain en mémoire comparé à la méthode classique qui nécessite la présence de la totalité de l'objet en mémoire avant son traitement.

4. Résultats expérimentaux et conclusion

Dans ce papier nous avons introduit une méthode qui permet d'effectuer une compression par ondelettes « au fil de l'eau » de maillages 3D multirésolutions semi-réguliers. Cette méthode consiste à faire le traitement au fur et à mesure de l'acquisition des données tout en réduisant considérablement l'espace mémoire utilisée.

4.1 Transformée en ondelettes « au fil de l'eau »

Les expériences montrent que la transformée en ondelettes est très efficace en terme de coût et accès mémoire, de l'ordre de 20,25% de la totalité de l'objet et sous une borne maximale de 33,7 %. La figure 6 représente une estimation de la taille maximale de la mémoire utilisée dans le cas du fil de l'eau en comparaison avec la méthode où le chargement de tout l'objet avant son traitement est nécessaire. Par exemple, la méthode proposée n'utilise que 17% de l'objet Rabbit pour avoir la même transformée en ondelettes.

4.2 Performance de la compression « au fil de l'eau »

La figure 7 représente une comparaison de la méthode proposée « au fil de l'eau » avec la méthode nécessitant la connaissance de l'objet en totalité. L'objet est découpé en plusieurs lots de triangles et une allocation de débit suivi d'un codage sont appliqués sur chaque lot séparément de sorte que la somme des débits soit égale au débit cible. Les lots choisis ici représentent 50% et 25 % de la totalité de l'objet cad ½ et ¼ en nombre de triangles total. Le rapport signal à bruit de l'objet final sera inférieur comparé à la méthode utilisant 100% de l'objet. Ce pendant les résultats expérimentaux montrent que pour notre méthode cette différence est très petite et de l'ordre de 1 dB dans le cas d'une compression à bas débit et presque nulle au delà de 2 bits/vertex. En contrepartie nous utilisons 25% de la taille de l'objet, soit un gain de 4 fois la taille mémoire qui serait nécessaire pour traiter l'objet complet. Les rapports débit/distorsion (PSNR) de notre méthode au fil de l'eau sont très proches de ceux de la méthode avec chargement de l'objet en totalité avant traitement ; la même qualité de compression avec un espace mémoire beaucoup plus réduit (figure 7). De plus la visuelle des objets décompressés demeure inchangée (cf Figure 8)

Références

[1] F.Payan et M. Antonini "Multiresolution 3D mesh Compression.", IEEE ICIP Septembre 2002

[2] C.Parisot,M. Antonini et M. Barlaud "3D scan based wavelet transform for video coding." , IEEE WMSP 2001

[3] A.Khodakvosky, P.Schroder et W. Sweldens "Progressive geometry compression", SIGGRAPH 2000

[4] A.Elkefi, M. Antonini et C. BenAmar " Compression de maillages 3D multirésolution" Rapport Interne I3S numéro ISRN I3S/RR- 2003-30-FR

[5] M.Isenburg et S.Gumholh "Out-of-core compression for gigantic meshes", SIGGRAPH, Juillet 2003

[6] H.Hoppe "progressive meshes". In ACM SIGGRAPH 2003

[7] P.Alliez et C.Gotsman"Recent Advances in Compression of 3D Meshes", CORESA, 2003

[8] C.Chrysafis et A.Ortega "Line Based Reduced Memory Wavelet Image Compression", IEEE TIP, mars 2000, pages 378-389

[9] A. Elkefi, M. Antonini, and C. Ben Amar. "3D scan-based wavelet transform for multiresolution meshes", EUSIPCO 2004, Vienne, Autriche

[10] F. Payan and M. Antonini. Model-based bit allocation for normal mesh compression. In IEEE International Workshop on Multimedia and Signal Processing, Siena, Italie, 29 septembre, 01 octobre 2004.

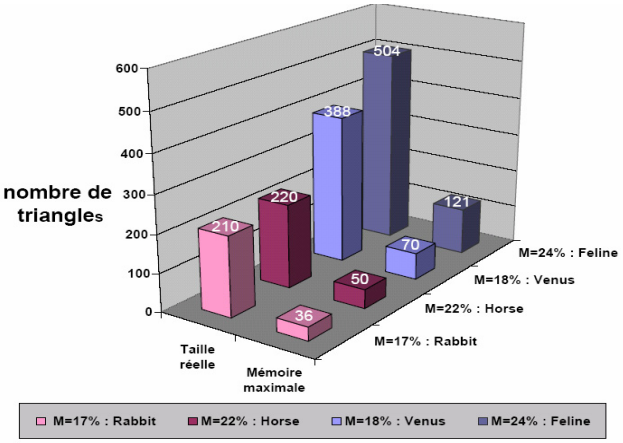


FIG. 6. *Estimation de la taille mémoire utilisée dans le cas « fil de l'eau » en comparaison avec celle nécessaire si tout l'objet était conservé en mémoire.*

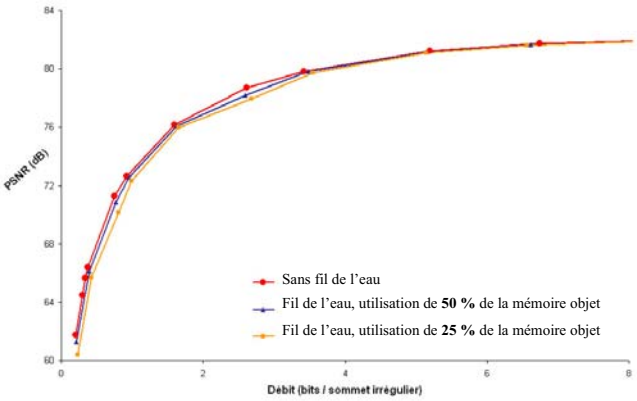


FIG. 7. *Courbe PSNR (dB) en fonction du débit bits/vertex pour l'objet Venus. Comparaison de la méthode au « fil de l'eau » avec l'approche « classique » nécessitant la connaissance de l'objet total.*

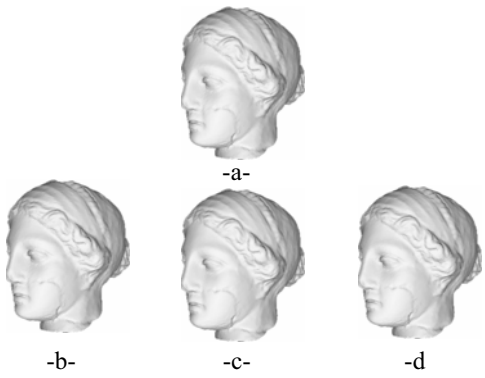


FIG. 8. *Comparaison visuelle entre les différentes méthodes.*

a- Objet original.

b- Compressé à 1 bit/vertex avec 100% de mémoire utilisée.

c- Compressé à 1 bit/vertex avec 50% de mémoire utilisée.

d- Compressé à 1 bit/vertex avec 25% de mémoire utilisée

Diffusion géométrique pour le masquage d'erreurs de quantification et de transmission sur des images JPEG couleur

Pascal BOURDON, Bertrand AUGEREAU, Christian CHATELLIER, Christian OLIVIER

Laboratoire SIC, FRE-CNRS 2731
Bât SP2MI, Bvd Marie et Pierre Curie, BP 30179,
F-86962 Futuroscope Chasseneuil Cedex, France
Tel : (33) 5 49 49 65 95 ; Fax : (33) 5 49 49 65 70
{bourdon, augereau, chatellier, olivier}@sic.univ-poitiers.fr

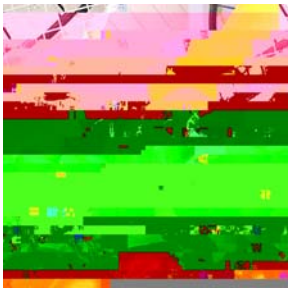
Résumé – Nous proposons ici une méthode de masquage d'erreurs de transmission *basée contenue*. Contrairement aux schémas classiques de correction d'erreurs de type FEQ/ARQ, notre méthode ne nécessite pas l'ajout de données de contrôle, et exploite directement la redondance spatiale de l'image source. Elle consiste en effet à interpoler les zones valides de l'image reçue dans les zones corrompues, à l'aide d'un processus de diffusion sous contraintes géométriques couplé à une approche multi-résolution. Outre les erreurs de transmission, ce modèle de diffusion nous permet de masquer également les erreurs de quantification (artefacts de compression).

Abstract – In this paper, we propose a *content-based* transmission error concealment method. Unlike classical FEC/ARQ error correction schemes, our method requires no extra control data, and take advantage of spatial redundancies contained in the source image. It consists in interpolating error-free decoded spatial information into corrupted areas, using a coupled multi-resolution/geometry-driven diffusion process. Beside transmission errors, this diffusion model can also be used to conceal quantization errors (compression artifacts).

1 Introduction

Malgré l'efficacité évidente de nouveaux standards de compression d'images, basés notamment sur des techniques de décomposition en ondelettes, comme JPEG2000 ou SPIHT, JPEG demeure l'un des standards les plus utilisés en télécommunications. Cependant, un problème récurrent dans tout système de communications est la perte ou dégradation possible des informations émises. Si aucune précaution n'est prise pour garantir une certaine qualité de service, l'un des inconvénients majeurs d'une transmission JPEG sur canal bruité est que la présence d'erreurs, aussi faible le Taux d'Erreurs Binaires (TEB) soit-il, peut avoir des conséquences désastreuses sur la qualité de l'image décodée (problèmes de désynchronisation, blocs manquants, *etc*) [1]. La raison d'une telle sensibilité aux erreurs est l'utilisation de codes à longueur variable. Bien que ces codes permettent d'atteindre des taux de compression très élevés, ils introduisent également une forte dépendance entre les mots de code. Du fait de cette dépendance, une simple erreur binaire aura tendance à se propager, et à rendre impossible la reconstruction correcte des données jusqu'à ce qu'un marqueur de synchronisation soit trouvé, en supposant que de tels marqueurs aient été insérés dans le flux de données. A titre d'exemple, nous pouvons observer, sur la figure 1, le résultat d'une simulation de transmission JPEG sur canal gaussien ($SNR = 11.0dB$, modulation MAQ-16) sans (Fig. 1.a), et avec (Fig. 1.b) marqueurs de synchronisation.

Comme nous pouvons le constater sur la figure 1.b, bien que l'utilisation de marqueurs de synchronisation permette de limiter la propagation des erreurs de transmission, elle ne les corrige pas pour autant. C'est pourquoi nous proposons ici un



(a) Flux désynchronisé

(b) Flux resynchronisé

FIG. 1: *Resynchronisation de flux JPEG erronés ($TEB \approx 6.0e^{-4}$)*

algorithme de post-traitement, dont l'objectif n'est pas la restitution d'un flux binaire identique bit pour bit au flux émis, mais la restitution d'une image visuellement fidèle à l'originale (nous parlerons donc de *masquage* d'erreurs, et non de correction). Celui-ci consiste à interpoler les informations spatiales de l'image décodées sans erreurs dans les zones corrompues, en exploitant leurs propriétés géométriques [2]. Outre JPEG, cette méthode peut d'ailleurs s'appliquer à d'autres schémas de compression par blocs, comme MPEG.

Cet article est organisé de la façon suivante : dans la section 2, nous définissons le principe d'une diffusion sous contraintes géométriques et présentons notre modèle de reconstruction, qui se base sur le modèle d'*inpainting* par Variation Totale de Chan et Shen [3]. Puis, nous explicitons dans la section 3 le fonctionnement de notre approche multi-résolution. Enfin, des résultats de simulation seront proposés dans la section 4.

2 Diffusion géométrique

2.1 Principe

La reconstruction automatique de données manquantes, ou *inpainting*, se révèle comme une nouvelle tendance dans le domaine du traitement d'images par Equations aux Dérivées Partielles (EDP) [4, 3]. Il s'agit de reconstituer l'intégralité d'une ou de plusieurs zones perdues d'une image, en utilisant les données valides présentes au voisinage de ces zones. Bien que de nombreuses méthodes de reconstruction de données par EDP se basent aujourd'hui sur des modèles du 3^{eme} ordre, plus performants, les modèles du 2^{nd} ordre, proches des modèles EDP de diffusion sous contraintes géométriques [5], peuvent eux aussi donner de bons résultats en *inpainting* [6]. De plus, il a été démontré qu'un modèle de diffusion par courbure moyenne (2^{nd} ordre) était équivalent à une opération de filtrage médian [7]. Cette propriété nous semble particulièrement intéressante, puisqu'elle nous permet de proposer une approximation simple de notre modèle, à base de filtres médians, laquelle pourrait être implantée facilement sur carte DSP.

Soit $I(x,y) = I$ une image fixe en niveaux de gris, représentée par une fonction de $\Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ qui associe au pixel $(x,y) \in \Omega$ son niveau de gris $I(x,y)$; Ω est le domaine de définition de l'image. L'EDP de régularisation de Perona-Malik peut s'écrire de la façon suivante :

$$\begin{cases} \dfrac{\partial I(x,y,t)}{\partial t} = c_\xi I_{\xi\xi} + c_\eta I_{\eta\eta} \\ I(x,y,0) = I_0(x,y) \end{cases} \qquad (1)$$

où I_0 définit une version bruitée de l'image I, $I_{\xi\xi}$ et $I_{\eta\eta}$ les dérivées secondes de I dans les directions orthogonales ξ et η, et ξ et η respectivement les directions du gradient et de son vecteur orthogonal.

Une propriété intéressante du vecteur unitaire ξ est qu'en chaque point (x,y), celui-ci est tangent à la courbure du contour de l'image. Ainsi, pour $c_\eta = 0$, l'équation (1) décrira la diffusion tangentielle pondérée (par c_ξ) de l'intensité lumineuse $I(x,y)$ le long d'un contour, durant un temps t. Intuitivement, la direction ξ nous permettrait donc d'interpoler des structures existantes à l'intérieur de zones manquantes, comme nous pouvons l'observer sur la figure 2.

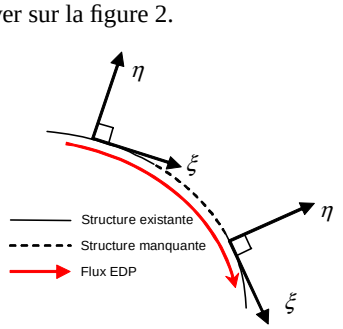


FIG. 2: *Interpolation par diffusion le long d'une courbure*

2.2 Modèle d'inpainting par Variation Totale

Dans [3], Chan et Shen proposent un modèle d'*inpainting* basé sur l'approche de restauration d'images par minimisation de Variation Totale (TV : *Total variation*) de Rudin *et al* [8]. Soit D la région de l'image à restaurer. $I_0|_{\Omega\setminus D}$ définit la portion incomplète, observée ou mesurée, de l'image originale sur le domaine Ω. Chan et Shen posent alors leur problème de reconstruction dans un cadre bayésien, et le présentent comme un problème de minimisation d'énergie. L'équation d'Euler-Lagrange qui lui est associée est la suivante :

$$\frac{\partial I}{\partial t} = \nabla . \left[\frac{\nabla I}{\|\nabla I\|} \right] + \lambda_e (I - I_0) \qquad (2)$$

valable sur l'intégralité du domaine Ω, où $\lambda_e = \lambda(1 - \chi_D)$ est le multiplicateur de Lagrange étendu, et χ_D la fonction caractéristique, ou masque, du domaine d'*inpainting* D.

L'équation (2) nous montre qu'un flux de chaleur purement anisotrope, orienté le long des courbures de l'image, est appliqué à l'intérieur du domaine d'*inpainting* D, dans le but de prolonger les lignes isophotes arrivant à ses frontières. Par ailleurs, nous démontrerons, à l'aide de résultats de simulation, que suivant les valeurs de λ, ce flux peut également être appliqué à l'intégralité de l'image, dans le but de masquer des erreurs de quantification (artefacts de compression).

2.3 Diffusion par tenseur

Tschumperlé *et al.* [9] ont récemment remarqué qu'une EDP basée sur des laplaciens orientés pouvait s'écrire de la façon suivante :

$$\frac{\partial I}{\partial t} = \text{trace}\,(\mathbf{TH}) \qquad (3)$$

où $\mathbf{H}$ est la matrice hessienne de I et $\mathbf{T} \in \mathbb{R}^{2\times 2}$ une matrice 2×2 définie positive qui définit les orientation et importance du lissage en chaque point de l'image au cours du processus de régularisation. $\mathbf{T}$ d'ailleurs appelé *tenseur de diffusion*.

L'un des principaux avantages liés à l'utilisation de la formulation de Tschumperlé *et al.* pour implanter le modèle de *TV-inpainting* est son extension rapide au cas d'images couleur, et le fait qu'elle conduise à de meilleurs schémas de résolution numérique, comparé aux formulations basées sur l'opérateur $div(.)$. Soit $\mathbf{I} : \Omega \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$ une fonction vectorielle qui associe au pixel $(x,y) \in \Omega$ ses 3 composantes spectrales primaires $\mathbf{I}(x,y) = [I_1(x,y), I_2(x,y), I_3(x,y)]^T$ dans l'espace couleur Rouge-Vert-Bleu (RGB). Notre modèle de *TV-inpainting* couleur peut s'écrire de la façon suivante :

$$\begin{cases} \dfrac{\partial I_i}{\partial t} = \text{trace}\,(\mathbf{TH}_i) + \lambda_e (I_i - I_{0_i}) \\ I_i(x,y,0) = I_{0_i}(x,y) \end{cases} \quad \text{for} \quad i = 1,2,3 \qquad (4)$$

avec $\mathbf{T} = \frac{1}{\sqrt{\lambda_+^*}} \theta_-^* \theta_-^{*\,T}$, où λ_+^*, λ_-^*, θ_-^*, et θ_+^* sont respectivement les valeurs et vecteurs propres d'une version lissée du tenseur de structure multispectral de DiZenzo $\mathbf{G} = \sum_{i=1}^{3} \nabla I_i \nabla I_i^T$

(lisser le tenseur de DiZenzo nous permet d'éviter d'éventuels problèmes de stabilité théorique [10], et d'évaluer la géométrie locale à partir d'un voisinage plus étendu).

3 Un schéma multi-résolution

3.1 Motivations et principe

L'analyse multi-résolution d'une image consiste à explorer celle-ci à travers plusieurs niveaux de description (obtenus par sous-échantillonnage), de la plus faible résolution, qui représente l'image dans une version simplifiée et grossière, à des résolutions plus élevées, sur lesquelles vont apparaître des structures plus fines. Nous pouvons établir un parallèle avec la théorie de l'espace échelle, et l'utiliser pour modéliser les premières phases de la vision humaine. Nos motivations concernant l'élaboration d'un processus d'*inpainting* multi-résolution étaient les suivantes : un modèle d'*inpainting* doit certes prolonger les structures de l'image dans les zones manquantes, mais il doit le faire de la même façon qu'un observateur humain le ferait : des structures globales (basse résolution) aux plus fines (haute résolution). Cette méthode a deux avantages principaux : elle devrait d'abord nous permettre d'accélérer sensiblement le processus de reconstruction, les zones manquantes étant initialisées à chaque niveau de résolution par des valeurs *a priori* proches de la solution (dans beaucoup d'algorithmes de *inpainting*, ces zones manquantes sont initialisées par un bruit gaussien, afin d'assurer la convergence de l'algorithme vers une bonne solution). De même, le fait de considérer en premier lieu les structures globales de l'image devrait nous permettre de vérifier le principe de connectivité de la perception humaine mis en évidence par Chan et Shan [3] (Fig. 3), sans modifier pour autant notre modèle EDP.

FIG. 3: *Le TV-inpainting Simple-résolution (S-R) ne vérifie pas le principe de connectivité*

3.2 Définition d'une grille multi-résolution

Notre objectif est donc de réduire les zones manquantes de l'image en réduisant sa taille, afin d'éviter les connections longue-distance. De petites zones d'*inpainting* (idéalement des zones de 1 pixel, puisque dans ce cas précis chaque pixel voisin constitue une donnée image valide) combinées à l'utilisation d'un tenseur de structure lissé devraient nous permettre de calculer des flots de diffusion cohérents. Nous pourrons alors restaurer les structures globales, et utiliser ce résultat afin d'initialiser les prochains niveaux de résolution, jusqu'à la restauration du dernier niveau (image pleine résolution).

Les transferts entre grilles de différentes résolutions seront définis à l'aide des fonctions $(.)_{\downarrow 2}$ et $(.)_{\uparrow 2}$: la fonction $(.)_{\downarrow 2}$ dénote l'opérateur de restriction, qui déplace la *matrice image* (nous sommes à présent dans le cas discret) d'une grille fine à une grille grossière, tandis que la fonction $(.)_{\uparrow 2}$ dénote l'interpolation, qui consiste à effectuer l'opération inverse. Dans notre cas, les transferts inter-grilles vont dépendre de la présence ou non de données valides. L'opération de restriction est effectuée à l'aide d'une méthode de *full weighting* : celle-ci consiste à remplacer 4 pixels par un seul dont la valeur résulte d'un calcul de moyenne. Dans notre approche, les points de l'image appartenant à D ne seront pas pris en compte dans les calculs de moyenne. L'opération d'interpolation, quant à elle, ne consiste qu'à initialiser les parties manquantes des grilles fines par une simple prolongation (*sample and hold*) des grilles grossières (Fig. 4).

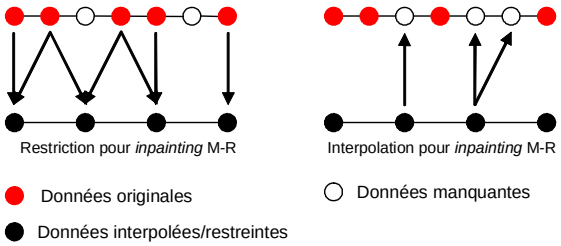


FIG. 4: *Méthodes de restriction et d'interpolation*

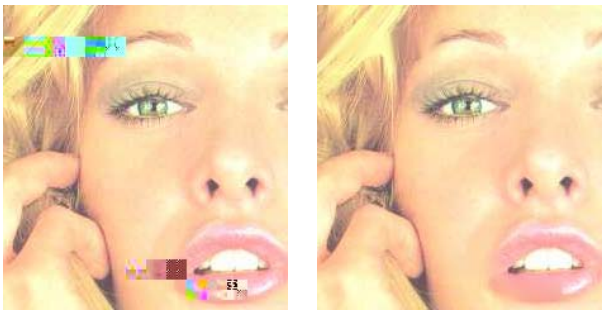
Notre algorithme fonctionne de la façon suivante :

- La fonction caractéristique χ_D du domaine D est sous-échantillonnée jusqu'à ce qu'elle n'indique plus aucune zone de *inpainting*, ou qu'elle ait atteint se taille minimale. Nous notons L le nombre total de niveaux de résolution;
- La matrice $\mathbf{I}$ est restreinte $L-1$ fois. Le niveau $L-1$ sera le plus petit niveau de résolution et $\mathbf{I}_{L-1} = (\mathbf{I})_{\downarrow 2^{L-1}}$ la représentation de l'image qui lui est associée;
- A moins qu'elle ne contienne aucune zone d'*inpainting*, l'image $\mathbf{I}_{L-1}$ est restaurée à l'aide du modèle de l'équation (4);
- L'image $\mathbf{I}_{L-1}$ restaurée est utilisée pour initialiser l'image de résolution supérieure $\mathbf{I}_{L-2} = (\mathbf{I}_{L-1})_{\uparrow 2}$. Chaque image $\mathbf{I}_l$ sera initialisée à l'aide de l'image $\mathbf{I}_{l+1}$ avant diffusion, jusqu'à $l = 0$.

4 Résultats de simulation

4.1 Masquage d'erreurs de transmission

La figure 5 nous montre un exemple de résultat obtenu par notre algorithme de reconstruction ($\lambda = 1$) sur l'image JPEG "Tiffany" (512×512, couleur) pour une transmission 16-QAM sur un canal gaussien ($SNR = 12.0dB$, 4.6% de données perdues). Comme nous pouvons le constater, malgré la perte de blocs de pixels de dimensions 16×16, l'algorithme a réussi à récupérer une bonne partie des structures corrompues. Des mesures de PSNR viennent par ailleurs confirmer la qualité du résultat, celui-ci passant de 22,69dB à 29,55dB.



(a) Reçue (22,69dB) (b) Reconstruite (29,55dB)

FIG. 5: *Reconstruction de l'image "Tiffany" (plan rapproché)*

4.2 Masquage d'erreurs de transmission et de quantification

Dans cette section, nous montrons que dans le cas d'un codage bas débit, connu pour engendrer des erreurs de quantification visibles à l'oeil humain, il est possible d'utiliser les propriétés de régularisation de notre modèle pour masquer de telles erreurs. Dans notre exemple, l'image "Peppers" (512×512, couleur) a été encodée à l'aide d'un codeur JPEG à un débit de $0.4bpp$. Afin de coupler *inpainting* et régularisation, la valeur du multiplicateur de Lagrange λ est passée de 1 to 0,3 au cours du processus de diffusion pleine-résolution (niveau 0). La figure 6 nous montre un résultat obtenu pour une transmission 16-QAM sur canal gaussien ($SNR = 11.0dB$, 6.1% de données perdues). Nous remarquons alors qu'au delà de la reconstruction des zones corrompues, l'orientation des flots de diffusion le long des lignes d'isophotes de l'image nous a permis de régulariser ses structures locales, et atténuer ainsi les effets de blocs et effets de Gibbs. Cette fois, le PSNR est passé de 20,96dB à 25,73dB.

(a) Reçue (20,96dB) (b) Reconstruite (25,73dB)

FIG. 6: *Reconstruction et régularisation de l'image "Peppers" (plan rapproché)*

5 Conclusion

Dans cet article, nous avons présenté une nouvelle méthode de masquage d'erreurs de transmission sur des images JPEG couleur par diffusion multi-résolution sous contraintes géométriques. Notre algorithme s'inspire d'un modèle d'*inpainting* par Variation Totale, qui nous permet d'élaborer un processus d'interpolation directionnelle à partir d'une base mathématique robuste. En interpolant les informations de voisinage d'une zone corrompue de l'image par diffusion géométrique, nous avons pu en effet restaurer une bonne partie des structures qui la composent. De même, l'utilisation d'une approche multi-résolution nous permet non-seulement d'accélérer le processus de diffusion, mais aussi d'effectuer des connections longue-distance, en reproduisant un comportement similaire à celui d'un observateur humain (c-à-d des structures globales aux plus fines). Au delà des erreurs de transmission, nous avons montré que notre modèle pouvait aussi être utilisé pour masquer d'éventuelles erreurs de quantification (artefacts de compression), grâce à ses propriétés de régularisation. Remarquons que cette méthode ne peut pas être utilisée pour reconstruire correctement de larges zones texturés. C'est pourquoi nos travaux futurs auront pour objectif de combiner celle-ci à des méthodes de synthèse de textures. De même, une extension au cas de séquences d'images à partir d'analyses de flots optiques, ainsi que l'implantation physique de cette méthode sur carte DSP, sont à présent à l'étude.

Références

[1] V. DeBrunner, L. DeBrunner, L. Wang, and S. Radhakrishnan, "Error control and concealment for image transmission," *IEEE Communications Society Surveys and Tutorials*, vol. 3, no. 1, pp. 2–9, 2000.

[2] B. Haar Romeny, *Geometry-Driven Diffusion in Computer Vision*, Kluwer Academic Publishers, 1994.

[3] T.F. Chan and J. Shen, "Mathematical models for local nontexture inpaintings," *SIAM Journal on Applied Mathematics*, vol. 62, no. 3, pp. 1019–1043, 2002.

[4] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 2000, pp. 417–424, ACM Press/Addison-Wesley Publishing Co.

[5] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 7, pp. 629–639, 1990.

[6] V. Caselles, J.M. Morel, and C. Sbert, "An axiomatic approach to image interpolation," *IEEE Transactions on Image Processing*, vol. 8, no. 3, pp. 376–386, March 1998.

[7] F. Guichard and J.M. Morel, "Introduction to partial differential equations in image processing," in *IEEE International Conference on Image Processing*, Washington, DC, 1995, p. Tutorial Notes.

[8] S. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, pp. 259–268, 1992.

[9] D. Tschumperlé and R. Dériche, "Diffusion PDE's on vector valued images, local approach and geometric viewpoint," *IEEE Signal Processing Magazine - Special issue on Mathematical Methods in Imaging*, vol. 19, no. 5, pp. 15–25, 2002.

[10] L. Alvarez, P-L. Lions, and J-M. Morel, "Image selective smoothing and edge detection by nonlinear diffusion. ii," *SIAM Journal on Numerical Analysis*, vol. 29, no. 3, pp. 845–866, 1992.

Interleaved S+P : Décomposition pyramidale entrelacée à contexte de prédiction enrichi

Marie BABEL, Olivier DÉFORGES, Joseph RONSIN

IETR CNRS UMR 6164 Groupe Image / INSA de Rennes
20, av. des Buttes de Coësmes, CS 14315, 35043 Rennes, France
{mbabel,odeforge,ronsin}@insa-rennes.fr

Résumé – La progressivité et les fonctionnalités du type codage par Région d'Intérêt constituent des composantes essentielles d'un codeur d'images efficace. L'état de l'art, dans le contexte de la compression sans perte, est représenté par les techniques S+P et CALIC. Dans un contexte pyramidal, l'algorithme ici proposé, l'Interleaved S+P, non seulement affiche des performances supérieures à ces méthodes, mais également procure les fonctionnalités avancées souhaitées. Une décomposition pyramidale originale, associée à un schéma MICD et fondée sur la méthode LAR (Locally Adaptive Resolution) est décrite dans cet article. Cette solution de codage met en œuvre la transformée en S de telle sorte que, pour chaque phase d'estimation, un contexte de prédiction enrichi soit disponible. Pour ce faire, la reconstruction sans perte d'une image est effectuée en deux étapes successives, où la première fournit une image LAR basse résolution de très bonne qualité.

Abstract – Scalability and functionalities such as Region of Interest coding constitute essential components of efficient image coder. In lossless compression context, state-of-the-art techniques are S+P and CALIC ones. We propose in a pyramidal context a new algorithm, the Interleaved S+P, that achieves better results than the previous methods, and provides also the desired functionalities. An original pyramidal decomposition used in association with a DPCM scheme and based on the LAR method (Locally Adaptive Resolution) is described. This solution adapts the S-Transform so that a refined prediction context is available for each estimation process. The lossless reconstruction of the image is thus realized in two successive steps, where the first one provides a LAR low-resolution image of very good quality.

1 Introduction

Du fait de la généralisation de l'utilisation des réseaux à bas-débits du type Internet, les méthodes de compression d'images tendent à satisfaire au besoin de scalabilité de la source. En effet, des applications spécifiques, telle la télémédecine, nécessitent une transmission progressive des images en termes de résolution et de qualité, et ce au moyen d'un unique codeur.

De ce fait, bien que les techniques opérant en pleine résolution de l'image soient les plus répandues dans le domaine de la compression sans perte (CALIC [1]), les codeurs à structure pyramidale continuent à faire l'objet de recherches. Si l'utilisation conjointe de schémas MICD et de transformées réversibles est rarement mise en œuvre au sein d'un même procédé de codage, l'algorithme S+P [2] propose une décomposition multirésolution de l'image, associant la transformée en S et des étapes de prédiction simples.

La méthode LAR (Locally Adaptive Resolution), fondée sur une représentation de l'image à taille de blocs variable, définit une technique efficace de compression avec pertes [3]. Nous avons introduit dans [4, 5] une approche pyramidale à redondance minimale (le LAR-APP), où les étapes de prédiction successives tirent à la fois parti d'un voisinage à 360 ° et d'une modélisation de contexte implicite. Cet article propose une solution originale de codage d'images : profitant de la description pyramidale du LAR-APP, une implantation particulière de la transformée en S conduit à la définition de l'Interleaved S+P.

Après avoir évoqué les particularités du LAR simple (§2), nous donnons la méthode de construction entrelacée de la pyramide Interleaved S+P (§3). La décomposition pyramidale est par la suite décrite en association avec les estimateurs mis en œuvre (§4). Enfin, les résultats de l'Interleaved S+P sont présentés en section 5.

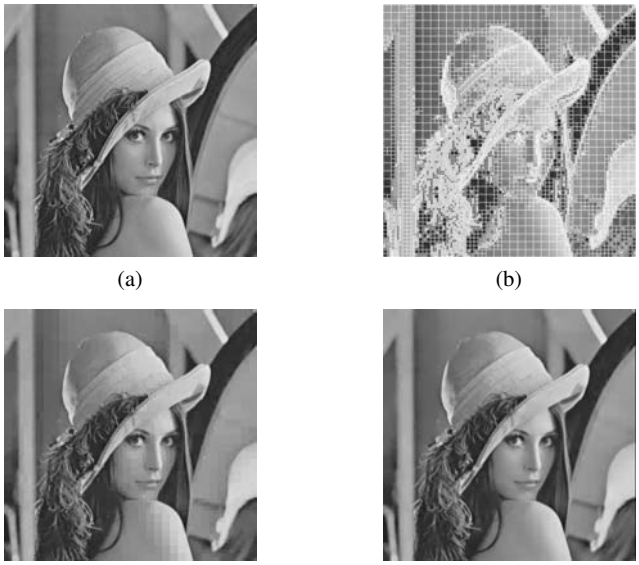
2 La méthode LAR pour les images en niveaux de gris

La méthode LAR (Locally Adaptive Resolution), définie en premier lieu pour la compression à bas débits d'images en niveaux de gris, est constituée d'un codec à deux couches : le codeur spatial, fournissant une image basse résolution, et le codeur spectral traitant de la texture. L'originalité de l'algorithme repose sur le principe suivant : la résolution locale, *i.e.* la taille des pixels, peut varier en fonction de l'activité locale. Ainsi, les pixels de petite taille se situent naturellement sur les contours de l'image, alors que les zones uniformes sont représentées par un bloc de grande taille (figure 1).

Cette grille de résolution spécifique, s'appuyant sur une topologie Quadtree déterminée par un gradient morphologique, a été en particulier exploitée dans l'élaboration d'une représentation en régions auto-extractible [6]. L'Interleaved S+P présenté dans cet article ajoute la propriété de scalabilité de codage de la source.

3 Construction de la pyramide Interleaved S+P

Dans cette section, les principes généraux de l'algorithme sont tout d'abord introduits, avant d'exposer la construction de



(a) (b) (c) (d)

FIG. 1 – (a) Image originale (b) Grille (0.032bpp). (c) Image basse résolution (0.4bpp) - Première descente de pyramide. (d) Image post-traitée.

la pyramide. Par la suite, nous notons z_0 et z_1 les coefficients transformés d'un couple (u_0, u_1) tels que

$$\begin{aligned} z_0 &= \lfloor (u_0 + u_1)/2 \rfloor, \\ z_1 &= u_1 - u_0, \end{aligned} \tag{1}$$

avec $\lfloor \cdot \rfloor$ opérateur d'arrondi à l'entier immédiatement inférieur.

3.1 Principes généraux

La décomposition pyramidale du LAR est issue de l'adaptation du prédicteur de Wu décrit dans [7]. Pour une image pleine résolution, les erreurs de prédiction sont encodées par trois échantillonnages successifs de l'image originale. Ainsi, pour un pixel donné, l'estimateur peut s'appuyer sur un contexte de configuration spatiale dite à 360 °, de telle sorte que l'entropie résultante soit considérablement réduite.

La première phase de l'Interleaved S+P consiste en l'application de la transformée en S 1D sur les deux vecteurs constitué chacun des deux pixels diagonalement adjacents d'un bloc 2×2 donné (figure 2). Par la suite, nous notons z_i^k, avec $i \in \{0,1\}$ et $k \in \{1,2,3\}$, les coefficients z_i codés par la k^{eme} passe de l'algorithme.

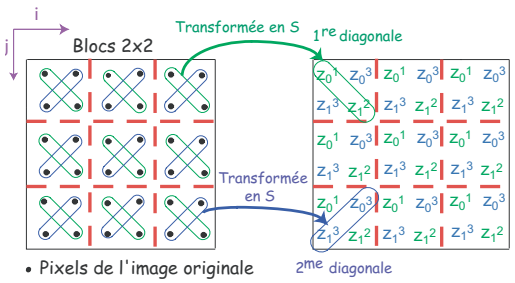


FIG. 2 – Application de la transformée en S

La première passe encode, par un MICD classique, l'image sous-échantillonnée formée par l'ensemble des valeurs de moyenne de deux pixels adjacents (coefficients z_0^1, cf figure 2). Par la suite, la seconde passe estime la valeur des coefficients z_1^2 afin de pouvoir reconstruire les deux pixels de la première diagonale. Pour cette seconde passe, le voisinage à 360 ° est constitué des valeurs déjà prédites (voisinage causal), ainsi que de l'ensemble des moyennes de diagonale issues de la première passe. Enfin, la troisième passe traite du reste de l'image originale, à savoir les coefficients transformés z_0^3 et z_1^3 : de nouveau, le contexte riche disponible permet d'estimer efficacement leur valeur.

3.2 Construction de la pyramide - Entrelacement

Soit I l'image originale de taille $N_x \times N_y$. La représentation multirésolution d'une image est donnée par l'ensemble $\{Y_l\}_{l=0}^{l_{max}}$, où l_{max} désigne le sommet de la pyramide et $l = 0$ la pleine résolution. Par extension de la méthode de Wu, quatre blocs $\frac{N}{2} \times \frac{N}{2}$ sont représentés au niveau supérieur par un bloc $N \times N$ de valeur égale à la moyenne des deux blocs sous-jacents de la première diagonale (première pyramide en S, figure 3) :

$$\begin{cases} l = 0, & Y_0(i,j) = I(i,j); \\ l > 0, & Y_l(i,j) = \left\lfloor \frac{Y_{l-1}(2i,2j)+Y_{l-1}(2i+1,2j+1)}{2} \right\rfloor, \end{cases} \tag{2}$$

avec $0 \le i \le N_x^l$, $0 \le j \le N_y^l$, où $N_x^l = N_x/l$ et $N_y^l = N_y/l$.

L'originalité de l'algorithme Interleaved S+P tient dans le fait qu'il est possible de **décrire la construction de la pyramide entière au moyen de deux pyramides entrelacées** (figure 3). La transformation de la deuxième diagonale d'un bloc 2×2 donné peut en effet aussi être vue comme la réalisation d'une seconde pyramide en S : l'estimation de la valeur d'un pixel dépend alors de celles présentes au niveau inférieur de la première pyramide en S.

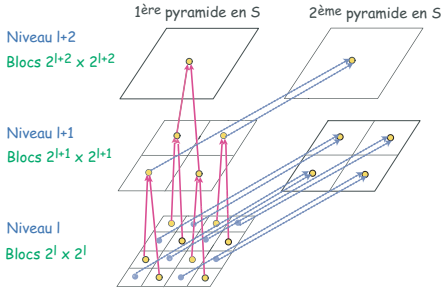


FIG. 3 – Construction de la pyramide par entrelacement de deux pyramides en S

4 Décomposition de la pyramide

Le processus de décomposition pyramidale de l'Interleaved S+P résulte de l'extension de la méthode de prédiction précédemment décrite à la pyramide entière. Dans cette section sont décrites l'ensemble des phases de prédiction.

4.1 Principes de la décomposition pyramidale

La reconstruction d'un bloc 2×2 situé à un niveau donné l est réalisée en deux passes. Tout d'abord, la transformée en S

s'applique sur les pixels $Y_l(2i,2j)$ et $Y_l(2i+1,2j+1)$:

$$z_0^{l,1}(2i,2j) = \left\lfloor \frac{Y_l(2i,2j)+Y_l(2i+1,2j+1)}{2} \right\rfloor, \quad (3)$$
$$z_1^{l,2}(2i+1,2j+1) = Y_l(2i,2j) - Y_l(2i+1,2j+1).$$

Nous remarquons immédiatement que le coefficient $z_0^{l,1}(2i,2j)$ est égal à la valeur du pixel reconstruit au niveau $l+1$ de la pyramide :

$$z_0^{l,1}(2i,2j) = Y_{l+1}(i,j). \quad (4)$$

De façon similaire, les pixels de la seconde diagonale sont transformés selon :

$$z_0^{l,3}(2i+1,2j) = \left\lfloor \frac{Y_l(2i,2j+1)+Y_l(2i+1,2j)}{2} \right\rfloor, \quad (5)$$
$$z_1^{l,3}(2i,2j+1) = Y_l(2i+1,2j) - Y_l(2i,2j+1).$$

Comme mentionné au paragraphe 2, la pyramide est alors décomposée selon deux descentes successives. Si la première descente vise à reconstruire l'image des blocs, la seconde encode l'image de texture : la grille LAR y est alors exploitée (voir figure 4). Pour ce faire, une image notée Siz contenant l'information de taille de blocs est élaborée à partir de notre partitionnement Quadtree. La figure 1.c donne l'image des blocs du LAR, résultat de la première descente partielle de notre pyramide.

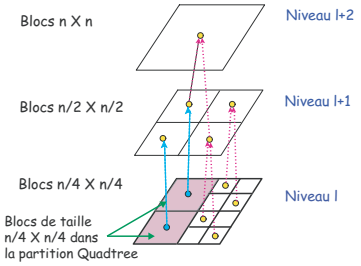


FIG. 4 – Décomposition : extraction de l'image LAR des blocs

4.2 Sommet de la pyramide - MICD classique

Au niveau le plus haut de la pyramide, les valeurs des pixels $Y_{l_{max}}(i,j) = z_0^{l_{max}-1,1}(2i,2j)$ subissent une unique phase d'estimation. Le prédicteur classique MED, utilisé dans la méthode de compression method LOCO-I [8], est alors mis en œuvre.

4.3 Définition des prédicteurs - Image des blocs

La reconstruction de l'image LAR basse résolution (image des blocs) induit, pour un niveau de la pyramide donné, le traitement des seuls pixels de taille $Siz(x\times l, y\times l) \leq 2^l$. Inversement, si $Siz(x\times l, y\times l) > 2^l$), la valeur du bloc est simplement recopiée de telle sorte que $Y_l(2i,2j) = Y_l(2i+1,2j) = Y_l(2i,2j+1) = Y_l(2i+1,2j+1) = Y_{l+1}(i,j)$.

Nous notons par la suite $\breve{z}$ la valeur estimée du coefficient transformé z.

Première pyramide en S. La valeur du coefficient $z_0^{l,1}$ étant déjà connue (niveau supérieur de la première pyramide en S), il nous suffit de procéder à l'estimation du coefficient transformé $z_1^{l,2}$, valeur gradient de la première diagonale d'un bloc 2×2 donné. Dans l'objectif d'obtenir l'image des blocs (zones de forte activité), nous avons remarqué qu'une estimation linéaire des coefficients demeure la plus efficace :

$$\begin{aligned}\breve{z}_1^{l,2}(2i+1,2j+1) = 2.1\Big[&0.9Y_{l+1}(i,j) + \frac{1}{6}\Big(Y_l(2i+1,2j-1)\\ &+Y_l(2i-1,2j-1)+Y_l(2i-1,2j+1)\Big)\\ &-0.05\Big(Y_l(2i,2j-2)+Y_l(2i-2,2j)\Big)\\ &-0.15\Big(Y_{l+1}(i,j+1)+Y_{l+1}(i+1,j)\Big) - Y_{l+1}(i,j)\Big)\Big].\end{aligned} \quad (6)$$

Deuxième pyramide en S. La troisième passe de l'algorithme de Wu adapté à notre problématique encode les coefficients transformés issus de la deuxième diagonale d'un bloc 2×2, appartenant de fait à la seconde pyramide en S. L'estimation de $z_0^{l,3}$ fait intervenir une prédiction inter- et intra-niveau :

$$\begin{aligned}\breve{z}_0^{l,3}(2i+1,2j) = \alpha_0\frac{1}{4}\Big(&Y_l(2i-1,2j+1)+Y_l(2i,2j+2)\\ &+Y_l(2i+2,2j)+Y_l(2i+1,2j-1)\Big)\\ &+\beta_0\hat{z}_0^{l,1}(2i,2j),\end{aligned} \quad (7)$$

où $(\alpha_0,\beta_0) = (0.25, 0.75)$, et $\hat{z}_0^{l,1}$ représente la valeur reconstruite du coefficient $z_0^{l,1}$. La valeur estimée $\breve{z}_1^{l,3}(2i,2j)$ se calcule selon :

$$\begin{aligned}\breve{z}_1^{l,3}(2i,2j+1) = \alpha_1\Big(&Y_l(2i-1,2j+1)+Y_l(2i,2j+2)\\ &-Y_l(2i+1,2j-1)-Y_l(2i+2,2j)\Big)\\ -\beta_1\Big(&Y_l(2i-1,2j)+Y_l(2i-1,2j+2)\\ &-Y_l(2i,2j-1)-\breve{Y}_l(2i,2j+1).\end{aligned} \quad (8)$$

avec $(\alpha_1,\beta_1) = (3/8, 1/8)$. $\breve{Y}_l(2i,2j+1)$ correspond au prédicteur de Wu [7] de la troisième passe appliquée au pixel $Y_l(2i,2j+1)$.

4.4 Informations de texture et prédicteurs

La réalisation de la deuxième phase de décomposition de la pyramide permet la reconstruction de la texture. Rappelons qu'à chaque niveau de la pyramide, seuls les pixels de taille $Siz(x\times l, y\times l) > 2^l$ subissent une étape de prédiction, les autres ayant été traités lors du premier processus de descente. Il est de nouveau possible de distinguer le cas des valeurs issues de la première pyramide en S (première diagonale), de celles de la deuxième pyramide en S (seconde diagonale).

Première pyramide en S. L'information de texture se caractérise par une faible activité locale. Les valeurs de gradient $z_1^{l,2}$ sont difficilement appréciables à travers un processus linéaire. C'est pourquoi nous nous appuyons sur des opérateurs non linéaires de type médian.

Soit $m_e(u_1,u_2,\ldots,u_n)$ la valeur médiane d'un ensemble $(u_1,u_2,\ldots,u_n)$ de n valeurs. La valeur estimée du coefficient

$z_1^{l,2}(2i,2j)$ situé dans une zone de texture est donnée par

$$\tilde{z}_1^{l,2}(2i+1,2j+1) = \frac{1}{4}\Big(m_e\big(Y_l(2i-2,2j), Y_l(2i,2j-2), Y_l(2i-1,2j-1)\big) + m_e\big(Y_{l+1}(i+1,j), Y_{l+1}(i,j+1), Y_{l+1}(i+1,j+1)\big)\Big). \quad (9)$$

Deuxième pyramide en S. Le contexte intra-niveau disponible pour l'estimation des coefficients transformés $z_0^{l,3}$ et $z_1^{l,3}$ s'avère suffisant pour une prédiction simple linéaire. Ainsi, les valeurs de moyennes $z_0^{l,3}$ sont traitées par l'application de la relation 7, où $(\alpha_0, \beta_0) = (0.37, 0.63)$. Quant aux coefficients $z_1^{l,3}$, leur estimation est obtenue après application de l'équation 8, avec le couple de valeurs $(\alpha_1, \beta_1) = \left(\frac{1}{4}, 0\right)$.

5 Résultats - Modélisation de contexte implicite

L'état de l'art des méthodes de compression d'images sans perte désigne le codeur CALIC [1] comme étant l'un des plus efficaces en termes d'entropie du premier ordre (taux de compression) ainsi que de complexité. L'algorithme S+P [2] constitue quant à lui une référence en matière de schéma de compression progressif. Nous avons donc procédé à des tests comparatifs avec ces précédents schémas de codage. Dans cette étude, l'encodage entropique (par codage arithmétique adaptatif) n'est pas implanté. Le tableau 1 expose les entropies du premier ordre obtenues après application de l'Interleaved S+P, CALIC et S+P. Notre méthode dépasse largement le simple S+P. De plus, les résultats de l'Interleaved S+P équivalent en moyenne à ceux de CALIC ; pour la majorité des images de cette série de mesures, notre algorithme surpasse même ce schéma de codage.

Grâce à notre sous-échantillonnage non uniforme, l'entropie est considérablement réduite, comme l'indique le tableau 1. En effet, l'entropie globale peut être diminuée lorsque plusieurs classes de symboles suivant la même loi peuvent être isolées. Notre méthode implique une séparation immédiate des lois : pour chaque niveau de la pyramide, nous distinguons les erreurs de prédiction issues de l'image LAR des blocs de l'image de texture, la passe 2 de la passe 3, les coefficients z_0 des coefficients z_1. De ce fait, grâce à notre représentation basée contenu, une modélisation de contexte implicite est réalisée.

	Entropie (bpp)			
Image	Interleaved S+P - Sans partition	**Interleaved S+P**	CALIC	S+P
Barb2	5.21	4.96	4.93	5.04
Cafe	5.66	5.43	5.37	5.42
Gold	4.82	4.64	4.65	4.73
Hotel	4.91	4.66	4.57	4.97
Lena	4.46	4.30	4.33	4.33
Peppers	4.74	4.57	4.58	4.67
tools	5.78	5.52	5.53	5.73
us	4.01	3.55	3.60	3.78
Moyenne	4.95	**4.70**	**4.70**	4.83

TAB. 1 – Entropie du premier ordre de la méthode Interleaved S+P (avec partition ou non), CALIC et S+P.

6 Conclusion

Cet article a présenté une technique originale de compression conjointe avec et sans perte. Fondée sur la méthode LAR, cette solution de codage utilise la transformée en S de telle sorte qu'un contexte de prédiction enrichi soit disponible pour chaque étape d'estimation. La connaissance de la structure Quadtree est suffisante pour améliorer de façon significative le taux de compression sans perte, par le regroupement et le codage séparé des informations de même nature (coefficients transformés, image LAR basse résolution, contours...). Les performances de cet algorithme Interleaved S+P dépassent significativement celle de la transformée classique S+P, et sont globalement équivalentes à CALIC. En supplément, la technique proposée permet non seulement la reconstruction progressive d'images, mais aussi fournit une carte de segmentation gratuite à la fois au codeur et au décodeur : une application immédiate en est la compression localement sans perte d'une image, par la définition d'une zone d'intérêt.

Nos recherches actuelles portent sur l'adaptation du codage LAR à la vidéo. D'autres travaux concernent l'élaboration d'une approche conjointe source-canal, en tant que solution pour la transmission sécurisée d'images médicales à travers les réseaux à faibles bandes passantes du type Internet.

Références

[1] X. Wu and N. Memon and, "A Context-based, Adaptive, Lossless/Nearly-Lossless Coding Scheme for Continuous-Tone Images (CALIC)," *International Standards Organization working document, ISO/IEC SC29/WG 1/N256*, vol. 202, 1995.

[2] A. Said and W. Pearlman, "Reversible Image Compression via Multiresolution Representation and Predictive Coding," in *Visual Communication and Image Processing*. SPIE, 1993, vol. 269, pp. 664–374.

[3] O. Deforges and J. Ronsin, "Region of Interest Coding for Low Bit-Rate Image Transmission," in *Proc. ICME'2000*, July 2000, vol. 1, pp. 107 – 110.

[4] M. Babel, O. Déforges, and J. Ronsin, "Lossless and Lossy Minimal Redundancy Pyramidal Decomposition for Scalable Image Compression Technique," in *Proc. ICASSP'03*, Hong Kong, April 6-10 2003, Conference cancelled - Invited paper in ICME 2003, vol. III, pp. 249–252.

[5] Marie Babel, Olivier Déforges, and Joseph Ronsin, "Décomposition pyramidale à redondance minimale pour compression d'images sans perte," in *GRETSI'03*, Paris, France, Sept. 8-11 2003.

[6] O. Deforges and J. Ronsin, "Supervised Segmentation at Low Bit Rates for Region Representation and Color Image Compression," in *Proc. ICME'2002*, Aug. 26-29 2002, vol. 1, pp. 665 – 668.

[7] X. Wu, "Lossless Compression of Continuous-Tone Images, via Context Selection, Quantization and Modelling," *IEEE Trans. on Image Processing*, vol. 6, no. 5, pp. 656–664, May 1997.

[8] M.J. Weinberger, G. Seroussi, and G. Sapiro, "LOCO-I : a low complexity, context-based, lossless image compression algorithm," *Proc. DCT'96*, pp. 141–150, 1996.

Evaluation d'un codeur de parole à très bas débit

Marc PADELLINI[1,2], Geneviève BAUDOIN[1], François CAPMAN[2]

[1] ESIEE, ESYCOM Telecommunications Systems LaboratoryBP 99, 93162 Noisy-Le-Grand, CEDEX, France.
[2]THALES COMMUNICATIONS, 160, Bd de Valmy, BP 82, 92704 Colombes, CEDEX, France.
marc.padellini@fr.thalesgroup.com

Résumé – Les approches classiques de codage de la parole à la trame ne permettent pas d'assurer une qualité de restitution suffisante à des débits inférieurs à 600 bits/s. En dessous, une nouvelle classe de codeurs dite à très bas débit doit être utilisée pour permettre d'exploiter la corrélation entre les trames successives afin de maintenir une bonne intelligibilité de la parole. Cet article présente un schéma complet de codage à très bas débit à 500 bits/s, qui exploite une base de donnée de segments de parole déterminés de manière entièrement automatique. Une évaluation complète de cette approche est réalisée suivant trois tests différents (MOS, intelligibilité, intelligibilité « segmentale ») conjointement avec trois autres codeurs dont le débit se situe entre 3600 bits/s et 600 bits/s. Les tests montrent qu'une bonne intelligibilité ainsi qu'une bonne qualité peuvent être atteintes grâce à ce type d'approche.

Abstract *– Classical frame based speech coding is lacking of quality at rates under 600 bits/s. Below, a new type of coders must be used to achieve sufficient intelligibility, by exploiting the correlation between successive frames. This paper presents a very low bit rate speech coder at 500 bits/s, resorting on a speech segment database designed by a completely unsupervised process. A three test evaluation on MOS, intelligibility and "segmental" intelligibility has been performed jointly with three other coders operating at a bit rate ranging between 3600 bits/s and 600 bits/s. This evaluation reveals that good intelligibility as well as good quality can be achieved by using segment indexing techniques.*

1. Introduction

Les codeurs de parole à très bas débit standardisés comme le NATO STANAG 4591 à 1200 bits/s utilisent des schémas classiques de codage à la trame de type MELP (Mixed Excitation Linear Predictive vocoder), dont le fonctionnement a été étendu à des bas débits en optimisant la quantification de trames successives. Ceci montre les limites du codage de la parole par trames. En effet, on peut dire de manière générale qu'en dessous de 1000 bits/s l'intelligibilité de ce type de codeur est fortement dégradée.

Ces vingt dernières années, plusieurs approches de codage à très bas débit ont été proposées afin d'exploiter la corrélation entre plusieurs trames successives de parole. Hoshiya et al. [2] proposent de combiner reconnaissance et synthèse de la parole en utilisant des HMM (Hidden Markov Models). Le codage est similaire à un système de reconnaissance vocale : une séquence de phonèmes est déterminée à l'aide de HMM. Cette séquence permet de restituer la parole en concaténant les modèles HMM correspondants, puis en générant une suite de coefficients cepstraux à partir de leur densité de probabilité d'émission. Ces coefficients permettent d'obtenir une approximation de l'évolution de l'enveloppe spectrale. Ce type d'approche permet d'atteindre des débits de l'ordre de 150 bits/s, cependant la qualité de ce type de codeur semble encore limitée [2].

Un autre approche consiste à transmettre directement l'évolution de l'enveloppe spectrale en utilisant par exemple un modèle polynomial sur des paramètres proches des formants [3]. Une bonne intelligibilité est atteinte entre 550 et 650 bits/s. Mais l'enveloppe spectrale n'est pas précise à cause du faible nombre de paramètres mis en jeu, ce qui limite la qualité de restitution. En revanche ce type de schéma permet un codage indépendant du locuteur.

Enfin une dernière approche s'inspire des techniques de reconnaissance et de synthèse vocale TTS (Text to Speech Synthesis). Le codeur reconnaît une séquence d'unités acoustiques dans le signal original. Pour chaque segment, il transmet le symbole correspondant à l'unité reconnue ainsi que des paramètres auxiliaires tels que les contours de fréquence fondamentale et d'énergie ainsi que la longueur du segment. La synthèse se fait par concaténation de représentants des unités élémentaires. Ces représentants (formes d'onde temporelles) sont stockés dans une base de données ou corpus de parole dont la taille est similaire à celle des systèmes TTS.

Ce type de codage utilise des phonèmes comme unités acoustiques. Potentiellement une bonne qualité peut être atteinte pour des débits de l'ordre de 400 bits/s [4] mais l'apprentissage doit être supervisé. D'autres types d'unités acoustiques peuvent être utilisées pour résoudre ce problème. Pour cela, des classes d'unités sont déterminées automatiquement par des méthodes purement statistiques. Une qualité se rapprochant d'un MELP à 2400 bits/s a été atteinte par KS. Lee et R. Cox [5] pour un débit de 580 bits/s. Le codeur étudié dans ce papier appartient à cette famille de codeurs, ils permettent d'offrir un codage large bande du signal de parole pour des débits inférieurs à 1 Kbits.

Cet article présente un schéma de codage très bas débit par indexation d'unités de taille variable appelé VLBR (Very Low Bit Rate coder). Il a été développé dans le cadre du projet RNRT SYMPATEX et a été comparé à différents codeurs bas débits dans une évaluation réalisée par le

CNAM. Les différents codeurs faisant l'objet de l'évaluation sont présentés dans la partie 3. Le protocole de test employé ainsi que les résultats obtenus sont présentés dans la partie 4.

2. Présentation du codeur VLBR

Ce schéma de codage se place dans la continuité des travaux présentés dans [6]. Il utilise une approche de reconnaissance par HMM (Hidden Markov Model) d'unités élémentaires de parole dans le codeur et une approche de synthèse par corpus dans le décodeur [11]. Ces unités correspondent à des segments de parole où le signal est stable. Le système utilise un corpus de parole de grande taille dans lequel le codeur recherche le segment le plus proche du segment à coder et dans lequel le décodeur lit le segment de synthèse à concaténer.

2.1 Phase d'apprentissage

Une phase d'entraînement des HMM est préalablement nécessaire pour déterminer les unités élémentaires de parole. Pour cela 64 HMM sont entraînés de manière itérative. Leur topologie est la suivante : trois états, une évolution de type gauche-droite avec une gaussienne modélisant la probabilité d'émission. Pour initialiser l'apprentissage, une première transcription est nécessaire. Elle est obtenue en deux étapes, par une étape de segmentation, puis une étape de classification des segments. Le corpus de parole est segmenté en réalisant un dendrogramme qui regroupe les trames successives les plus proches afin de former des segments ancrés sur des zones stables du signal. Puis ces segments sont classés en minimisant la distance cumulée entre les segments et 64 centroïdes obtenus par quantification vectorielle du corpus. Après l'entraînement des HMM, un algorithme de Viterbi est utilisé pour segmenter et classer conjointement l'ensemble du corpus. Tous les segments de même classe et de même sous-classe sont regroupés dans une base de donnée. La sous-classe représente le contexte d'une unité, elle est définie par la classe de l'unité la précédent dans le temps.

2.2 Phase de codage

Le codeur est représenté sur la figure 2. Il effectue l'extraction au fil de l'eau de 16 paramètres cepstraux de type LPCC (Linear Predictive Cepstral Coefficients), ainsi que le calcul de l'énergie sur des fenêtres de 20ms avec 10ms de recouvrement. Un algorithme de Viterbi à horizon limité est utilisé pour segmenter et classer conjointement les segments à l'aide des HMM.

2.2.1 Sélection des unités

Pour pouvoir sélectionner l'unité de la base la plus proche de l'unité à coder, un processus de sélection à deux étapes a été proposé dans [7]. Tout d'abord les 16 unités - de la même classe et de la même sous-classe - les plus proches en terme de pitch moyen sont présélectionnées. Puis l'unité qui combine la meilleure corrélation avec l'unité à coder est sélectionnée. Une corrélation cumulée sur le profil de pitch, le profil d'énergie et l'enveloppe spectrale moyenne est utilisée comme critère de sélection. Ceci permet de garder l'ensemble des unités du corpus tout en ayant une quantité fixe de bits alloués à l'index de l'unité sélectionnée.

2.2.2 Correction de la prosodie

Un modèle appelé HSD pour (Harmonic Stochastic Model) est utilisé pour manipuler l'unité sélectionnée. C'est un modèle de type Harmonique plus Bruit basé sur [8]. Il permet de réaliser des transformations sur l'échelle de temps et sur le pitch avec une bonne qualité. Deux paramètres de correction de la prosodie sont calculés en normalisant linéairement la longueur de l'unité sélectionnée avec celle de l'unité à coder:

- Un paramètre de correction linéaire permet de corriger la pente du profil de pitch.
- Un gain permet de corriger le niveau moyen du profil d'énergie.

Ces paramètres sont quantifiés sur une échelle logarithmique.

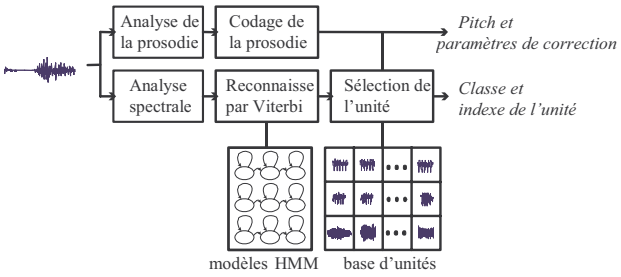


FIG. 2: Principe de codage VLBR.

2.3 Phase de décodage

Pendant la phase de décodage, les unités sont récupérées dans la base d'unités grâce à l'indice de classe, l'index de l'unité et le pitch moyen transmis. Les unités sont corrigées grâce aux paramètres de correction de la prosodie, puis elles sont concaténées pour restituer le signal de parole.

2.4 Table d'allocation des bits

Le débit moyen obtenu est d'environ 500 bits/s avec 170 bits/s réservés aux indices de classes et 330 bits/s réservés au codage de la prosodie. La table d'allocation des bits est présentée sur le tableau 1. La version testée ici est dépendante du locuteur et traite des signaux en bande élargie.

TAB 1: Table d'allocation des bits pour une unité

Paramètre	**Nombre de Bits alloués**
Classe (64)	6 bits
Index de l'unité (16)	4 bits
Total Information spectrale	*10 bits par segment*
Longueur du segment (3-18)	4 bits
Pitch moyen	5 bits
Correction du pitch	5 bits
Correction de l'énergie	5 bits
Total Information Prosodie	*19 bits/segment*
Total	**29 bits/segment**

3. Présentation des codeurs évalués

Cinq codeurs ont été retenus pour l'évaluation :

- Le codage PCM (Pulse Code Modulation) à une fréquence d'échantillonnage de 16000Hz car il permet

de constituer une référence du signal de parole non dégradé.

- L'analyse/synthèse harmonique (HSD) car elle correspond à la qualité maximale que l'ont peut espérer obtenir avec le codeur VLBR puisqu'elle est mise en jeu dans la manipulation des unités.
- Le codeur HSX-WB (Harmonic Stochastic eXcitation Wide-Band) au débit de 3600 bit/s. Il a été développé par THALES Communications sur la base décrite dans [9] et étendu à un fonctionnement en bande élargie [50 7000Hz]. Il offre une très bonne qualité sonore et une complexité de traitement réduite face au MELP.
- Le codeur MELPe (Enhanced Mixed excitation Linear Predictive vocoder) décrit dans [1] standardisé NATO STANAG-4591, à un débit de 2400 bit/s. Le codage porte sur le signal de parole en bande étroite [50 4000Hz].
- Une version du codeur MELP à 600 bit/s. Elle a été retenue pour une prochaine standardisation STANAG 4591.

4. Evaluation

Le codeur VLBR a d'abord été entraîné durant une phase d'apprentissage (Cf 2.1) séparément sur 10 locuteurs de la base BREF [10]. La quantité de parole utilisée est de l'ordre d'une heure de parole. 70 phrases par locuteur ont été réservées pour constituer des données de test.

4.1.1 Matériel de parole utilisé :

Sont codés par les 5 codeurs de l'évaluation des phrases qui constitueront 3 matériels de test:

- 70 phrases prononcées par 10 locuteurs. Ils ont été sélectionnés pour leur diversité dans la base de corpus lus de parole BREF (voir [10]). (M1)
- 40 phrases n'ayant pas de sens logique, extraites du corpus CNAM lues par 1 locuteur. (M2)
- 8 séquences de 28 mots lues par 1 locuteur. (M3)

Treize auditeurs ont été recrutés parmi des étudiants faisant des études supérieures. Trois tests différents ont été réalisés sur les quatre codeurs lors de l'évaluation .

4.1.2 Test subjectif MOS :

Le matériel de test (M1) est utilisé sur 13 auditeurs. Le test porte sur deux critères: la qualité globale et l'intelligibilité. 50 phrases sont diffusées (une phrase est tirée aléatoirement pour chaque locuteur et chaque codeur). Les auditeurs attribuent une note de 1 à 5 à chacun des extraits sonores et ce sur un critère seulement.

Pour tester la fatigue auditive des auditeurs, le test est doublé en alternant de critère. Ainsi les 13 auditeurs ont participé à 4 séances d'écoute.

Les résultats obtenus sont présentés sur la Figure 3 et 4, avec des intervalles de confiance à 95%. Les deux critères donnent le même classement, cependant la courbe d'intelligibilité est simplement translatée vers le haut ce qui montre que les codeurs sont plus performants en intelligibilité.

Le codeur VLBR surpasse le MELP à 600 bits/s mais reste de mauvaise qualité, alors que le MELPe à 2400 bits/s est de qualité moyenne. Le codeur HSX-WB et l'Analyse/Synthèse HSD sont perçus de manière équivalente et sont de bonne qualité.

4.1.3 Test d'intelligibilité par retranscription :

Le matériel de test (M2) est utilisé pour 11 auditeurs. Les locuteurs doivent retranscrire ces 40 phrases qui n'ont pas de sens logique. Il n'y a pas d'a priori possible, l'ordre des phrases et le codeur utilisé sont déterminés aléatoirement. Sur la Figure 5 est représenté le nombre de phonèmes non ou mal reconnus par phrase.

Ce test place le VLBR et le MELPe à 2400 bits/s à des niveaux comparables de performance en terme d'intelligibilité, et devant le MELP à 600 bits/s. En revanche le codeur HSX-WB et l'Analyse/Synthèse HSD donnent des niveaux comparables à la référence non dégradée.

4.1.4 Test d'intelligibilité « segmental » :

Le matériel de test (M3) est utilisé sur 11 auditeurs. C'est un test de rimes portant sur des mots monosyllabiques de type consonne/voyelle/consonne. L'auditeur entend un mot et doit choisir le mot qu'il a reconnu parmi deux mots qui lui sont présentés visuellement sur une feuille. Une séquence de 28 mots issue du même codeur doit être reconnue. Les codeurs et les séquences sont alternées aléatoirement de manière à ce que l'auditeur ait à reconnaître les 8 séquences de mots codés par les 6 codeurs, c'est à dire 48 séquences. Le nombre d'erreurs par séquence est représenté Figure 6.

Ce test place le VLBR et le MELPe à 2400 bits/s à des niveaux comparables (recouvrement de leurs intervalles de confiance), et devant le MELP à 600 bits/s. Le HSX-WB est proche de l'Analyse/Synthèse HSD mais les auditeurs le distinguent clairement par rapport à l'original.

5. Conclusion

Cette article a présenté les performances générales de codeurs à bas débit. Il montre que le codeur VLBR permet d'atteindre des niveaux d'intelligibilité comparables à ceux du codeur MELPe à 2400 bits/s pour un débit 5 fois plus faible. Ceci s'explique par l'apport du codage en bande élargie et par la représentation de l'évolution spectrale des trames par segments. Le codeur HSX-WB à 3600 bps bénéficie aussi de l'apport de la bande élargie ce qui lui confère une très bonne qualité globale de codage.

En revanche la qualité globale du VLBR reste à améliorer pour concurrencer les autres approches de codage et justifier les contraintes qui lui sont liées. Elle est toutefois légèrement supérieure à celle du codeur MELP à 600 bits/s. La concaténation sur des unités de type diphones devrait pouvoir réduire les discontinuités propres à ce type d'approche. Des travaux sur la robustesse du codeur VLBR dans des environnements bruyant sont en cours ainsi que son extension à un codage indépendant du locuteur.

Ce travail a été réalisé dans le cadre du projet RNRT SYMPATEX.

Références

[1] T. Wang, K. Koishida, V. Cuperman, A. Gersho, J.S. Collura. *A 1200/2400 bps coding suite based on MELP.* SCW'02, pp 90-92, 2002.

[2] T. Hoshiya, S. Sako, H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura. *Improving the performance of HMM-based very low bit rate speech coding.* Proc. ICASSP, pp800-803, 2003

[3] P. Zolgaghari, T. Robinson. *Speech coding using mixture of gaussians polynomial model.* Proc. Eurospeech. 1999

[4] M. Felici, Al. *Very low bit rate speech coding using a diphone-based recognition and synthesis approach.* Trans. IEE Electronic letters 9(34), pp 859-860, 1998.

[5] K.S. Lee, R.V. Cox. *A segmental speech coder based on a concatenative TTS.* Trans. Speech Communication 38, pp 89-100. 2002.

[6] J.Cernocky, G.Baudoin, G.Chollet, *Segmental vocoder – going beyond the phonetic approach*, Proc. ICASSP-98, pp. 605-608, 1998.J. Cernocky.

[7] M. Padellini, F. Capman, G. Baudoin. *Very low bit rate (VLBR) speech coding around 500 bits/sec*. Proc. EUSIPCO, Vienne, 2004.

[8] I. Stylianou, *Modèles Harmoniques plus Bruit combinés avec des Méthodes Statistiques, pour la Modification de la Parole et du Locuteur.* Thèse ENST, 1996.

[9] P. Gournay, F. Chartier. *A 1200 bits/s speech coder for very low bit rate communications.* IEEE Workshop on Signal Processing Systems (SiPS), Boston, 1998.

[10]L.F.Lamel, J.L.Gauvain, M.Eskenazi, *BREF, a large vocabulary spoken corpus for French*, Proc. EUROSPEECH-91, Genoa, Italie, 1991.

[11]G. Baudoin, F. El-chami, "Corpus based very low bit rate speech coding", *ICASSP'03 International Conference on Acoustics, Speech & Signal Processing,* 6 - 10 Avril 2003 Hong Kong.

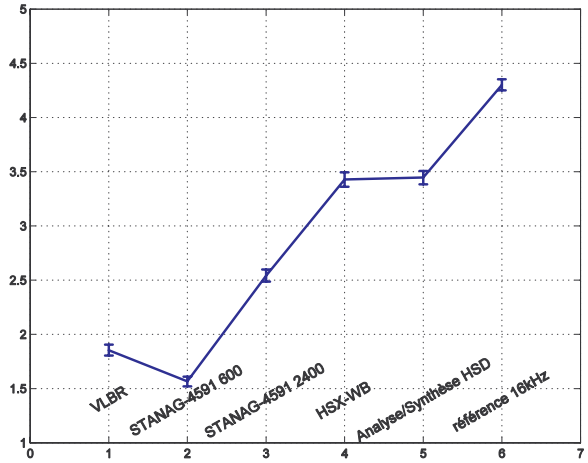


FIG. 3: Courbe test MOS, Critère: qualité globale

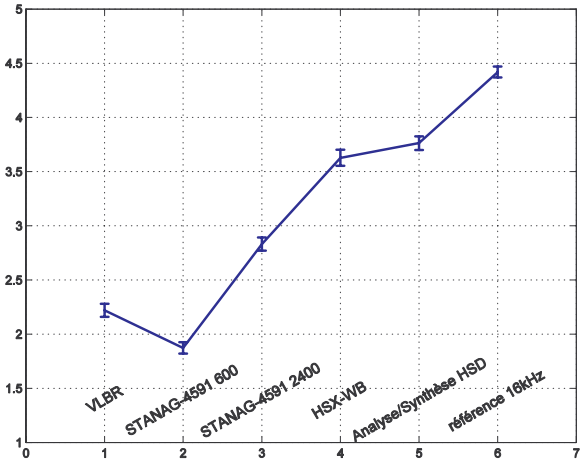


FIG. 4: Courbe test MOS, Critère: intelligibilité

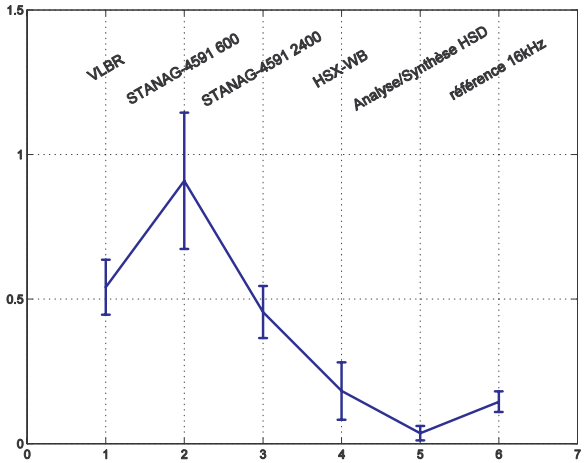


FIG. 5: Courbe relative au nombre d'erreur de transcription de phonèmes par phrase

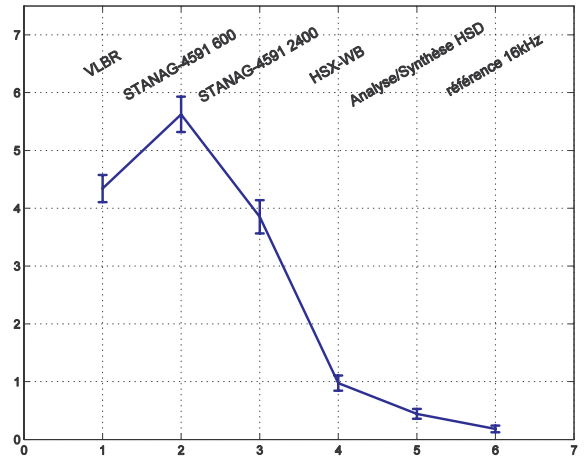


FIG. 6: Courbe relative au nombre d'erreur de reconnaissance par séquence de test.

Métriques de distorsion pour l'analyse comparative de schémas de filigranage 3D

Patrice RONDAO ALFACE, Mathieu DE CRAENE, Benoît MACQ

Université catholique de Louvain
Laboratoire de Télécommunications et Télédétection
Place du Levant 2, B-1348 Louvain-La-Neuve, Belgique
rondao@tele.ucl.ac.be, decraene@tele.ucl.ac.be, macq@tele.ucl.ac.be

Résumé – Dans cet article, nous traitons de la problématique de la mesure de distorsion des maillages 3D dans le cadre de l'analyse comparative de schémas de filigranage. Cette mesure est nécessaire afin de classifier les types de déformations acceptables et de déterminer des seuils de tolérance. Nous proposons dans cette optique deux approches distinctes et complémentaires. La première consiste en une métrique perceptive globale basée sur l'analyse des différences entre les vues 2D correspondantes de l'objet 3D et de sa version déformée. Nous montrons par nos résultats que le choix de l'information mutuelle comme critère de comparaison de ces vues projetées permet de mieux détecter de faibles distorsions à la limite de l'imperceptibilité, courantes dans le cadre du filigranage. La deuxième approche complète la première en analysant les différences locales entre maillages via l'estimation de l'énergie de déformation. Cette stratégie trouve ses fondements notamment dans le calcul de la paramétrisation planaire ou sphérique des surfaces maillées. Cette étude se conclut par la comparaison des résultats de ces méthodes et des métriques 3D de l'état de l'art.

Abstract – Three-dimensional image quality assessment causes new challenges for a wide set of applications and particularly for emerging 3-D watermarking schemes. New metrics have to be drawn for the distortion measurement from an original 3-D surface to its deformed version. Such metrics are necessary namely in the case of a watermarking scheme to determine whether a distortion is acceptable or not. We propose two complementary strategies to tackle this issue. First, we propose a global perceptive metric based on the analysis of the differences between correspondent 2D views of the original and deformed versions of the 3D model. We show that the mutual information criterion better captures slightly perceptible distortions to which watermarking schemes must resist. Secondly, we focus on local deformations by estimating the distortion energy, extending works on distortion minimization for planar and spherical parameterizations. Finally results of both methods are discussed and compared to the state of the art.

1 Introduction

Estimer la différence ou la similarité entre deux surfaces maillées 3D reste un problème ouvert et encore relativement peu exploré. L'importance de cette estimation s'avère néanmoins capitale dans de nombreuses applications telles la compression, la transmission progressive, l'indexation, ou le filigranage. Les premières métriques de distorsion ont été développées dans deux directions différentes mais complémentaires. La première est motivée par la nécessité de mesurer l'impact des déformations *globales* causées par les algorithmes de compression 3D[6]. La seconde direction de recherche, à l'inverse, s'est concentrée sur les déformations *locales* dans le cadre des schémas de simplification des maillages 3D[7]. Les outils issus de ces approches distinctes partagent néanmoins les mêmes fondements et se basent en général sur une estimation de la distance eucilidienne. Si ce choix permet de développer des algorithmes simples et rapides, il ne permet cependant pas de capturer la distorsion perceptive, exigence plus forte du domaine du filigranage. En effet, dans le cadre du benchmark de schémas de filigranage 3D, il est nécessaire de capturer la différence perçue par l'utilisateur entre le maillage original et sa version déformée. Etant donné le manque actuel d'outils performants de mesure de distorsion en 3D, la comparaison de schémas de filigranage demeure ardue et donc peu populaire.

Nous proposons en ce sens deux stratégies complémentaires. La première approche vise la mesure des différences (globales) de rendu entre version altérée et originale l'objet, essentielle dans de nombreuses applications où l'utilisation principale de l'objet 3D passe par sa visualisation 2D. La comparaison, via un critère basé sur l'information mutuelle, de nombreuses vues 2D sélectionnées automatiquement dans des conditions de rendu standard, permet de dessiner une métrique globale qui capture les légères modifications peu perceptibles auxquelles sont confrontés les schémas de filigranage. La deuxième approche, quant à elle, consiste en l'estimation de l'énergie de déformation pour étudier localement les distorsions apportées par les attaques courantes dans le domaine du filigranage.

Il est en outre important de souligner que cet article se concentre sur la mesure de la distorsion entre un *maillage original* et sa *version déformée* et non sur la mesure de qualité du maillage lui-même après traitement comme réalisé par Corsini et al. [4].

Cet article est organisé comme suit. La première section fait le point sur les métriques existantes et leurs limitations. Les sections 3 et 4 présentent nos contributions. Tout d'abord une métrique basée sur la mesure de la distorsion entre de mul-

tiples paires de vues 2D correspondantes et ensuite une autre métrique qui complète la première par l'analyse des déformations locales, basée sur l'estimation de l'énergie distorsion 3D. Nos résultats sont illustrés et commentés en section 5.

2 Etat de l'art

Les premières métriques 3D ont vu le jour dans le domaine de la simplification des maillages, outil de base notamment des célèbres schémas de multirésolution [6, 7]. Ces métriques ont été développées dans le souci de déterminer localement la distorsion provoquée par l'extraction d'un point ou d'une arête du maillage. Les métriques les plus usuelles dans ce cadre sont la très célèbre distance de Hausdorff[3, 8], les différences de volume[1], les différences locales de courbure[11] et les fonctions d'erreur locale quadratiques QEM[7]. Dans le cas particulier des schémas de compression sans modification de la connectivité, le PSNR et le Laplacien géométrique[10] se basent sur une somme normalisée des distances euclidiennes entre points correspondants des deux maillages à comparer ainsi que sur un critère de régularité de la surface. Parallèlement à ces travaux, Reddy[13] a proposé une mesure basée sur le modèle de vision humaine et défini un schéma de simplification qui dépend du point de vue, les zones visibles gardant plus de détails que les autres. Cependant, il n'existe encore à l'heure actuelle aucune mesure psycho-visuelle permettant d'estimer le caractère imperceptible des attaques et de l'insertion d'un filigrane.

Afin de permettre un benchmark efficace de schémas de filigranage, il convient de développer un outil qui mesure la distorsion subie par l'entièreté du maillage tout en assurant un contrôle local de la distorsion. Nous proposons d'allier deux stratégies différentes pour résoudre ces deux contraintes. La première est une métrique globale et est basée sur la visualisation de l'objet 3D, la seconde est de type locale et basée sur l'estimation de l'énergie de distorsion.

3 Distorsion visuelle dans le rendu

Dans cette section, nous étudions les distorsions des surfaces maillées 3D par la comparaison de projections 2D de la surface maillée par de multiples points de vue. En pratique, nous discrétisons de manière régulière la sphère de rayon minimal circonscrite au maillage et positionnons la caméra en ces points. Dans le cadre de nos expériences, nous avons testé plusieurs nombres de vues et avons déterminé un intervalle de confiance de 64 à 256 vues. Les conditions d'illumination sont choisies de manière standard par rapport à la position de la caméra. Cette méthode a été initialement explorée par Lindstrom et al [12] pour la simplification des maillages. Cependant, les auteurs n'utilisent qu'un critère de moindres carrés sur des images de luminance et utilisent cette technique pour déterminer des distorsions locales. Nous étendons cette étude en proposant un critère plus adapté à de faibles distorsions et à une possible extension de la méthode à des conditions de rendu différentes pour comparer les vues correspondantes des deux surfaces maillées : l'information mutuelle [15]. Ce critère mesure la dépendance statistique entre chaque paire d'images correspondantes. De plus, l'information mutuelle permet de re-synchroniser les vues entre elles si l'on n'a pas effectué une étape préliminaire de coregistration des deux maillages 3D à comparer. La section 5 expose nos résultats et les compare aux autres métriques, celle de Linstrom et al., le PSNR, le Laplacien géométrique et la distance de Hausdorff pour quelques attaques typiques de filigranage.

4 Energie de déformation

Les premières études sur l'énergie de déformation des surfaces maillées sont apparues avec le problème de la paramétrisation et du placage de texture [5]. Paramétrer un maillage 3D revient à le projeter sur le plan 2D en minimisant les distorsions qui sont mesurées par les différences d'angles, de longueurs ou d'aires des triangles de ce maillage. Diverses mesures de distorsion ont été proposées : énergie harmonique (préservation des propriétés globales), conforme (préservation des angles uniquement), authalique (préservation des aires uniquement), énergie de stretch etc. L'extension à la mesure de distorsion entre deux maillages tous deux plongés dans l'espace 3D exige quelques modifications. En effet, il faut tout d'abord partitionner la surface en sous-maillages de connectivité identique pour ensuite calculer l'application bijective d'énergie minimale pour chaque paire de sous-maillages correspondants. Le travail pionnier de Schreiner et al.[14] utilise des points caractéristiques choisis à la main pour réaliser cette partition à des fins de morphing entre objets 3D différents. Obtenant un maillage de base de faible résolution, qu'ils développent ensuite en *progressive meshes*, les auteurs calculent l'énergie de *stretch* et en dérivent une métrique symétrique. L'application obtenue en minimisant l'énergie non-linéaire de stretch requiert un difficile et long processus d'optimisations locales successives pour chaque résolution successive des sous-maillages. Nous proposons de simplifier le calcul de cette métrique en choisissant l'énergie harmonique et en remaillant les sous-maillages correspondants avec une connectivité régulière de subdivision par la méthode des *Normal Meshes* [9].

4.1 Remaillage

Le remaillage se construit de la manière suivante. En commençant par un triangle de base, on subdivise celui-ci en trouvant pour le milieu de chacune de ses arêtes, l'intersection entre la normale du triangle partant de ce point et la surface maillée de départ. En cas d'intersections multiples, nous choisissons automatiquement la plus proche au sens euclidien. L'absence d'intersection ne se présente que dans le cas de sous-maillages comportant des trous topologiques, ce que nous évitons par construction de la partition initiale. En pratique, à l'issue du remaillage chaque sous-maillage possède un nombre de points supérieur au nombre de points original. Ce remaillage, s'il est assez fin pour que chaque triangle de départ contienne au moins un point du remaillé, se justifie dans le cadre de l'application bijective harmonique qui est continue et linéaire par morceaux sur les triangles du maillage.

4.2 Energie harmonique

Notons A et B deux sous-maillages correspondants. L'énergie harmonique de déformation de l'application de A dans B est donnée par :

$$H_{AB} = \frac{1}{2} \sum_{(i,j)\in A} (\cot \alpha_{A_{i,j}} + \cot \beta_{A_{i,j}})(B(i) - B(j))^2, \quad (1)$$

où (i, j) est une arête dans A et B, $\alpha_{A_{i,j}}$ et $\beta_{A_{i,j}}$ sont les angles qui voient l'arête (i, j) dans A et $B(i)$ et $B(j)$ sont les images des points 3D i et j de A dans B. Pour obtenir une métrique symétrique, nous posons $H(A, B) = max(H_{AB}, H_{BA})$. L'énergie ici quadratique est aisément minimisée par la résolution d'un système linéaire contrairement aux optimisations locales de l'énergie de stretch. Nous montrons en outre qu'elle permet de déterminer localement les distorsions apportées par des transformations 3D. Les résultats sont illustrés et commentés à la section 5.

5 Résultats

Cette section présente nos résultats empiriques pour les deux métriques proposées dans cet article. Les figures 1, 2 et 3 comparent notre métrique basée sur les projections 2D avec les métriques usuelles de l'état de l'art pour quelques attaques typiques de filigranage : addition de bruit gaussien, lissage et décimation. Ces résultats montrent que l'information mutuelle capture mieux de petites variations de l'objet (ce qui est le cas de figure du filigranage) que le critère de moindres carrés proposé dans [12]. Intuitivement, ceci s'explique par l'allure de la courbe de l'information mutuelle en fonction de l'intensité des attaques. Celle-ci est fortement décroissante et de courbure positive pour des intensités d'attaque faibles, contrairement à la courbure généralement négative des moindres carrés. L'impact des conditions de rendu sur la métrique doit cependant être étudié plus en détails et comparé notamment avec le rendu de Phong moins courant mais de meilleure qualité[2]. La distorsion visuelle du modèle *bunny* peut être perçue dans les figures 4 et 5 pour les attaques testées. Enfin, la figure 6 illustre nos résultats préliminaires pour la détection de distorsion locale basée sur l'estimation de l'énergie de déformation. On peut y voir le résultat du remaillage par sudivision d'un sous-maillage triangulaire sur le modèle *head* ainsi que l'énergie de distorsion pour une attaque de lissage Laplacien.

6 Conclusions

Nous avons proposé deux nouvelles métriques pour l'application du benchmark de schémas de filigranage 3D. Les résultats expérimentaux sont probants mais nécessitent encore de nombreux tests. La suite des travaux concernera l'étude de l'impact des conditions de rendu sur notre métrique basée sur les projections et l'étude de bornes sur l'approximation du calcul de l'énergie harmonique en fonction de la finesse du remaillage.

7 Remerciements

Les deux premiers auteurs sont financés par la bourse FRIA de la Communauté Française de Belgique. Nous remercions également l'Action de Recherches Concertées CRYPTOMARK de la Région Wallonne.

Références

[1] P. Alliez and F. Schmitt, *Mesh Approximation Using a Volume-Based Metric*, In Proc. of Pacific Graphics, pp. 292-301, 1999.

[2] G. Bishop and D.M. Weimer, *Fast Phong Shading*, Proc. of the 13th Annual Conference on Computer Graphics and Interactive Techniques, pp. 103-106, 1986.

[3] P. Cignoni, C. Rocchini and R. Scopigno, *Metro : Measuring Error On Simplified Surfaces*, Computer Graphics Forum, Vol. 17, No. 2, pp. 167-174, 1998.

[4] M. Corsini, E. Drelie Gelasca and T. Ebrahimi, *A Multi-Scale Roughness Metric For 3D Watermarking Quality Assessment*, in Workshop on Image Analysis for Multimedia Interactive Services, April 13-15, Montreux, Switzerland, 2005.

[5] M. Desbrun, M. Meyer and P. Alliez, *Intrinsic Parameterizations Of Surface Meshes*, Computer Graphics Forum. 17(2), pp. 167-174, 2002.

[6] M. Eck, T. De Rose, T. Duchamp, H. Hoppe, M. Lounsbery and W. Suetzle, *Multiresolution Analysis Of Arbitrary Meshes*, ACM SIGGRAPH, pp. 173-182, 1995.

[7] M. Garland, *Quadric-Based Polygonal Surface Simplification*, Ph.D. dissertation, Computer Science Department, Carnegie Mellon University, CMU-CS-99-105, May 1999.

[8] A. Gueziec, *Meshsweeper : Dynamic Point-To-Polygonal-Mesh Distance And Applications*, IEEE Trans. on Visualization and Computer Graphics, pp. 47-61, 2001.

[9] I. Guskov, K. Vidimce, W. Sweldens and P. Schröder,*Normal Meshes*, Computer Graphics Proc. (SIGGRAPH 2000), pp. 95-102, 2000.

[10] Z. Karni and C. Gotsman, *Spectral Compression Of Mesh Geometry*, Computer Graphics (Proc. of SIGGRAPH), pp. 279-286, 2000.

[11] S. J. Kim, S.J. Kim and C. H. Kim, *Discrete Differential Error Metric For Surface Simplification*, In Proc. of Pacific Graphics, pp. 276-283, 2002.

[12] P. Lindstrom and G. Turk, *Image-Driven Simplification*, ACM Transactions on Graphics, Vol. 19, No. 3, pp. 204-241, 2000.

[13] M. Reddy, *Perceptually Optimized 3D Graphics*, IEEE Computer Graphics and Applications, Vol. 21, No. 5, pp. 68-75, 2001.

[14] J. Schreiner, A. Asirvatham, E. Praun and H. Hoppe, *Inter-Surface Mapping*, ACM SIGGRAPH, 2004, to appear.

[15] P. Viola and W. M. Wells, *Alignment By Maximization Of Mutual Information*, Fifth Int. Conf. On Computer Vision, pp. 16-23, 1995.

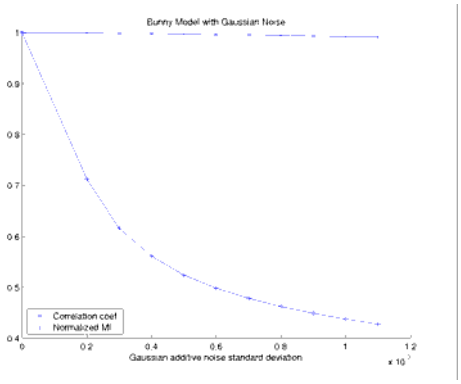


FIG. 1 – Attaque par bruit gaussien en fonction du paramètre α pour le modèle *bunny*. Les métriques de similarité ont été calculées par les moindres carrés et l'information mutuelle.

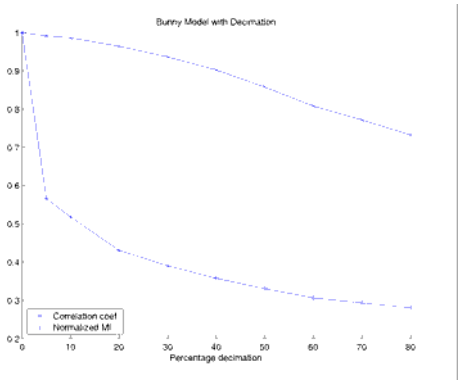


FIG. 2 – Attaque par décimation en fonction du pourcentage de points restant pour le modèle *bunny*. Les métriques de similarité ont été calculées par les moindres carrés et l'information mutuelle.

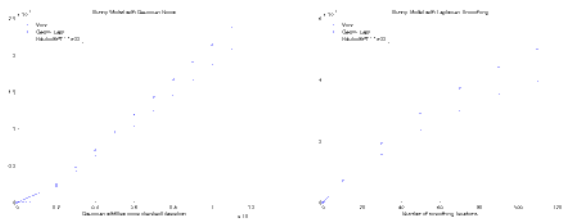


FIG. 3 – A gauche, attaque par bruit gaussien additif et à droite par lissage Laplacien pour le modèle *bunny*. Les métriques de distorsion 3D représentées sont le PSNR, le Laplacien géométrique et la distance de Haussdorff.

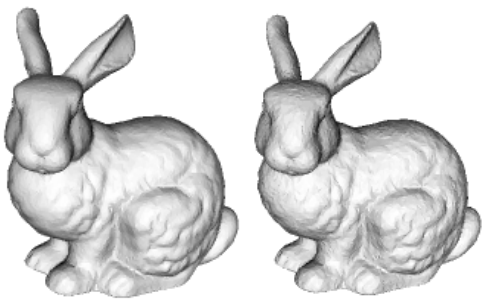


FIG. 4 – Modèle *bunny* représenté pour l'attaque de bruit gaussien avec des valeurs du coefficient α égales à $\alpha = .0005$ et .001 respectivement.

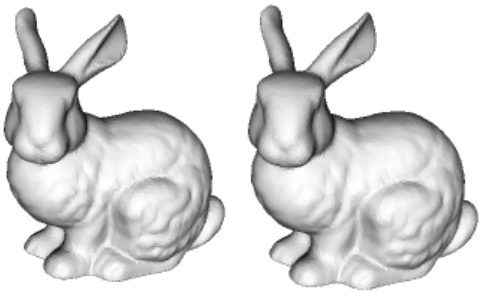


FIG. 5 – Modèle *bunny* représenté pour l'attaque de lissage laplacien avec 70 et 110 itérations respectivement.

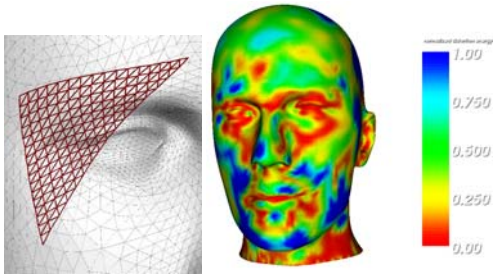


FIG. 6 – A gauche détail du processus de remaillage et à droite énergie de déformation harmonique locale pour le modèle *head* et sa version lissée avec 200 itérations.

Un Algorithme de Tatouage d'Images Numériques reposant sur les Changements d'Horloge Périodiques

Vincent MARTIN[1], Marie CHABERT[1], Bernard LACAZE[2]

[1]ENSEEIHT/IRIT, Institut National Polytechnique de Toulouse
2 Rue Camichel, BP 7122, 31071 Toulouse Cedex 7, France

[2]INSA/IRIT

Vincent.Martin,Marie.Chabert,Bernard.Lacaze@enseeiht.fr

Résumé – Cet article s'intéresse à l'insertion d'un ou plusieurs tatouages dans une image numérique selon la technique d'étalement de spectre. L'algorithme proposé utilise les Changements d'Horloge Périodiques, qui sont une alternative originale aux méthodes classiques de tatouage inspirées des communications multi-utilisateurs par étalement de spectre. Cet article fournit une étude théorique des performances de la méthode proposée, ainsi que des résultats expérimentaux présentant une comparaison détaillée avec un algorithme de référence, en fonction de diverses conditions d'insertion et de diverses attaques. Un masque perceptuel et l'utilisation des propriétés statistiques locales de l'image sont pris en compte dans les résultats théoriques et expérimentaux.

Abstract – This paper deals with the embedding of single or multiple watermarks into digital data. The proposed algorithm is based on Periodic Clock Changes, which are an alternative to classical pseudo-noise modulation methods. This study presents theoretical results and an experimental comparison between the performance of the proposed algorithm and of the classical one, in terms of embedding conditions and of robustness to several attacks. The theoretical and experimental results take into account the use of a perceptual mask and the use of local statistical properties of the image.

1 Introduction

La protection de l'intégrité des images, sons et vidéos numériques et la gestion de leurs droits d'auteurs suscite actuellement un grand intérêt. Le tatouage numérique consiste à insérer une marque dans les composantes perceptuelles du document numérique avant sa diffusion auprès des usagers. Cette marque est constituée d'un ou plusieurs messages secrets. La technique d'insertion peut être publique mais utilise une clé secrète. La marque doit être imperceptible pour l'usager, robuste aux attaques et doit pouvoir être détectée et décodée en réception par une personne munie de la clé. Dans le cadre du tatouage aveugle étudié dans cet article, l'image originale n'est pas nécessaire à la détection. Certaines techniques de tatouage, telles que l'étalement de spectre, s'inspirent d'une analogie avec les communications numériques [6]. En effet, l'insertion et le décodage d'un tatouage peuvent s'interpréter comme la transmission d'un message dans un canal bruité. Le bruit représente alors l'image hôte et les attaques.

Cet article étudie l'insertion additive et simultanée d'un ou plusieurs messages dans une image numérique, selon le principe du tatouage par étalement de spectre [10]. A l'insertion, on étale le spectre du message afin de le dissimuler et au décodage, on étale le spectre du bruit pour mieux l'éliminer. Par analogie avec les méthodes d'accès à répartition par code, on peut effectuer du tatouage multiple. Celui-ci est en effet nécessaire dans certaines applications comme le *fingerprinting*, où chaque utilisateur est identifié par une empreinte unique dans le but de détecter les copies illégales du document. Pour améliorer les performances en réception, différents domaines d'insertion sont possibles. Par ailleurs, l'utilisation d'un masque basé sur les caractéristiques psychovisuelles de l'image permet d'assurer l'imperceptibilité du tatouage.

Cette étude se focalise sur la technique d'étalement en elle-même. L'étalement de spectre est le plus souvent obtenu en modulant le message par une séquence pseudo-aléatoire appelée séquence directe. La séquence directe fait office de clé secrète. Cette technique sera notée dans la suite DS-CDMA (pour *Direct Sequence Code Division Multiple Access*). Dans cet article, les Changements d'Horloge Périodiques (PCC, pour *Periodic Clock Changes*) sont proposés comme technique alternative d'étalement de spectre. Les PCC possèdent des propriétés d'étalement intéressantes et ont été appliqués avec succès aux communications multi-utilisateurs [8].

La partie 2 présente le principe général des PCC et propose un algorithme de tatouage basé sur cette technique. Dans la partie 3, on compare au travers de simulations les performances au décodage de DS-CDMA et PCC par rapport au bruit introduit par l'image hôte et à diverses attaques.

2 Changements d'Horloges Périodiques

2.1 Définitions et propriétés

Un filtre Linéaire Périodique Variant dans le Temps (LPTV) est un filtre dont la réponse impulsionnelle est une fonction $h(n,k)$ T-périodique ($T \in \mathbb{N}$) du temps indexé par $n \in \mathbb{N}$. Sa fonction de transfert $H_n(\omega)$ est définie par :

$$H_n(\omega) = \sum_{k=-\infty}^{+\infty} h(n,k) e^{-ik\omega}, \qquad H_n(\omega) = H_{n+T}(\omega) \quad (1)$$

Les filtres LPTV ont été appliqués à l'entrelacement, à l'égalisation aveugle et aux communications par étalement de spectre [3].

Soit f une fonction T-périodique de n. Dans un cadre stochastique, si $Z = \{Z(n), n \in \mathbb{Z}\}$ est un processus stationnaire alors on appelle Changement d'Horloge Périodique le nouveau processus

$$U(n) = Z(n - f(n)), \qquad f(n) = f(n+T) \tag{2}$$

Les PCC sont des cas particuliers de LPTV ($H_n(\omega) = e^{-i\omega f(n)}$).

Supposons désormais que $f(n)$ est une permutation aléatoire T-périodique définie par $f(n) = \underline{n} - q_{\underline{n}}$, où q est une permutation de $(0, 1, 2, ..., T-1)$ et $\underline{n}$ le reste de la division euclidienne de n par T ($n = \overline{n}T + \underline{n}$). Le PCC inverse est $f^{-1}(n) = \underline{n} - q_{\underline{n}}^{-1}$. Pour T suffisamment grand, le spectre de V s'approche de celui d'un bruit blanc [8].

Les communications multi-utilisateurs utilisant les PCC transmettent pour chaque message M_j le résultat d'une permutation aléatoire f_j de période donnée T. L'application successive de deux PCC quelconques $f_i \circ f_j$ est un PCC et étale le spectre. Seul le PCC inverse f_j^{-1} permet de retrouver le spectre d'entrée. Les performances des PCC et du DS-CDMA ont été comparées dans [11] pour les communications multi-utilisateurs, en fonction du nombre d'utilisateurs. Les estimations du TEB (Taux d'Erreur Binaire) montrent des résultats similaires pour un grand nombre d'utilisateurs et une légère supériorité des PCC pour un faible nombre d'utilisateurs.

2.2 Application au tatouage

Les LPTV utilisés dans le cadre du tatouage doivent blanchir le spectre, être inversibles, sûrs cryptographiquement et éventuellement former un ensemble orthogonal pour le tatouage multiple. Les permutations aléatoires périodiques remplissent toutes ces conditions pour un faible coût calculatoire.

La littérature contient plusieurs références à l'utilisation des permutations aléatoires dans le cadre du tatouage, le plus souvent comme entrelaceur préalable du message pour améliorer la sécurité [5], mais aussi pour l'entrelacement d'une information à spectre coloré dans le cadre du tatouage audio asymétrique [1].

Changements d'Horloge Périodiques Monodimensionnels :

Le tatouage W_j est obtenu en appliquant un PCC T_{1D}-périodique f_j (la clé secrète) à une version redondante M'_j du message M_j : $W_j = f_j(M'_j)$, avec

$$m'_j(l + (p-1)L) = m_j(l), l \in \{1, ..., L\}, p \in \{1, ..., P\} \tag{3}$$

La redondance ainsi introduite permet d'obtenir de bonnes performances en réception malgré le rapport élevé entre la puissance du document et celle du tatouage (DWR, pour *Document to Watermark Ratio*) qu'impose la contrainte l'imperceptibilité.

Après transmission l'image reçue est l'image tatouée et attaquée $I'_W = I + B + \alpha \sum_{j=1}^{J} W_j$, où I est l'image support et B le bruit. Lors du décodage, le PCC inverse est appliqué à I'_W et étale le spectre de I et de B. La Fig.1 montre les propriétés d'étalement des PCC sur l'exemple d'une image numérique (Lena).

Les propriétés statistiques de l'image ou des coefficients de sa transformée permettent d'établir des stratégies de décodage

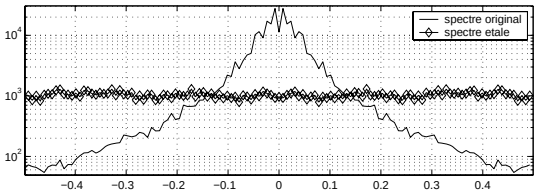


FIG. 1 – Propriétés d'étalement des PCC (Lena)

optimal [5]. En particulier, le décodage par corrélation ou par filtrage adapté est optimal dans le cas de distributions gaussienne de l'image et du bruit. Le message reçu est $\hat{M}_j$:

$$\begin{aligned} \hat{m_j}(l) &= \frac{1}{\alpha P} \sum_{p=1}^{P} (f_j^{-1}(i'_W))(l + (p-1)L) \\ &= m_j(l) + \frac{1}{\alpha P} \sum_{p=1}^{P} (f_j^{-1}(i+b))(l + (p-1)L) \end{aligned} \tag{4}$$

Soient $\sigma_I^2, \sigma_B^2, \sigma_{\text{MAI}}^2$ les variances dues respectivement à l'image support, au bruit et aux interférences multi-utilisateurs (dans le cas du tatouage multiple). D'après le théorème Central-Limite, $\hat{M}_j$ est gaussien de moyenne $E[\hat{M}_j] = M_j + \mu(I)$ (où $\mu(I)$ est la moyenne de I) et de variance $(\sigma_I^2 + \sigma_B^2 + \sigma_{\text{MAI}}^2)/\alpha^2 P$. Alors $\hat{m_j}(l)$ est une statistique suffisante pour le décodage et la règle de décision est

$$D_j(l) = [\text{signe}(\hat{m_j}(l) - \mu(I_W))]_{l \in \{(1,...,L\}}. \tag{5}$$

Le théorème Central-Limite montre que

$$\text{BER} \simeq Q(\alpha\sqrt{P}/\sqrt{\sigma_I^2 + \sigma_B^2 + \sigma_{\text{MAI}}^2}) \ , \tag{6}$$

où $Q(x) = \int_x^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du$. Les résultats expérimentaux présentés dans la partie 3 vérifient cette estimation théorique des performances.

Ce récepteur est préférable dans les conditions du tatouage ($\sigma_{\text{MAI}}^2 << \sigma_{I_W}^2$) à d'autres récepteurs (décorrélateur par exemple) utilisés dans le cadre de la détection multi-utilisateurs [12] et repris dans des publications concernant les permutations aléatoires [9].

Il existe $T!$ permutations équiprobables de période T. Dès que $T > N/8$, 1D-PCC offre donc de meilleures garanties sur l'entropie de la clé que la modulation de type DS, où l'on peut générer 2^N clés équiprobables.

Changements d'Horloge Périodiques Bidimensionnels :

Le tatouage est le résultat de l'application successive d'une permutation f_j^1 sur les colonnes et d'une seconde permutation f_j^2 sur les lignes du message redondant, pris sous forme matricielle. L'insertion et le décodage suivent le même principe que pour 1D-PCC. A la réception, on applique successivement les PCC inverses $(f_j^2)^{-1}$ et $(f_j^1)^{-1}$ sur l'image tatouée. L'objectif est d'obtenir des performances similaires à celles de 1D-PCC pour une période plus faible : l'association de deux PCC au décodage devrait davantage réduire la corrélation spatiale présente dans l'image support.

3 Simulations

3.1 Implantation

Domaine d'insertion : W peut être inséré dans la luminance I ou dans une transformation inversible de I (DFT, DCT [10],

transformée en ondelettes [2]...). DS et PCC ont été comparés dans le domaine spatial (luminance) (L-DS, L-PCC) et dans celui de la Transformée en Cosinus Discrète (DCT) par blocs 8x8 (DCT-DS, DCT-PCC), qui est le plus utilisé du fait de son rôle dans le format de compression JPEG.

Masque perceptuel : l'insertion dans le domaine de la DCT par blocs 8x8 (DCT-DS, DCT-PCC) bénéficie de recherches sur l'analyse perceptuelle effectuées en compression d'image, car on l'utilise dans le format JPEG. Le masque perceptuel inspiré des travaux de Ahumada et Peterson [4] a été choisi dans le domaine de la DCT. Dans le domaine de la DCT, la variance de l'image support σ_I^2 dans (6) doit être recalculée en une nouvelle variance $\tilde{\sigma_I}^2$ pour tenir compte de ce masque qui limite notamment l'insertion à 22 coefficients choisis dans les moyennes fréquences (où la variance est moindre).

Préfiltrage au détecteur : dans le domaine spatial, l'image n'est pas stationnaire. Cependant, on peut estimer ses moments locaux à l'aide d'un filtre de Wiener [5] afin d'obtenir une estimation $\hat{I}$ de l'image originale au moment du décodage. L'image préfiltrée est $I''_W = I'_W - \hat{I}$ et σ_I^2 dans (6) devient $\sigma^2_{I'_W - \hat{I}}$, ce qui réduit considérablement l'influence du bruit dû à l'image.

Paramètres : les performances de DS et PCC sont comparées par simulations sur un ensemble de 5 images test. Dans le domaine de la DCT, cet ensemble est complété par des images hôtes I générées aléatoirement : les coefficients de chaque bloc suivent des distributions gaussiennes généralisées [5]. Pour une estimation précise du TEB, des messages sont générés aléatoirement jusqu'à ce qu'au moins 100 bits erronés aient été observés. Les valeurs par défaut ont été fixées à $N = 2^{18}$, $L = 100$ bits, $J = 1$, $T_{1D} = 2^{12}$, $T_{2D} = 2^6$ et DWR situé au seuil d'imperceptibilité (DWR=36 dB en moyenne).

3.2 Performances intrinsèques

Influence de la puissance d'insertion : lorsque le DWR augmente, l'imperceptibilité du tatouage est meilleure mais le TEB augmente également, pour les trois algorithmes comparés (cf Fig.2).

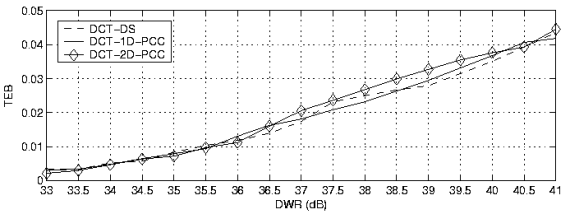


FIG. 2 – Performance au décodage par rapport à DWR (images simulées)

Choix de la période de la permutation : la période de la permutation influe sur les propriétés d'étalement, d'orthogonalité (dans le cas du tatouage multiple) et le coût calculatoire. Dans le domaine de la luminance, la corrélation entre les pixels joue un rôle important, ce qui n'est pas le cas dans le domaine transformé. On montre expérimentalement que les périodes $T_{1D} = 2^{12}$ et $T_{2D} = 2^6$ sont optimales pour une image comportant $N = 2^{18}$ pixels.

Influence du débit du message et du tatouage multiple : les performances de DS et PCC augmentent avec le nombre P d'échantillons insérés pour chaque bit d'information (Fig.3). Pour le tatouage multiple, les échantillons correspondant aux différents messages sont superposés, tout en respectant la contrainte d'imperceptibilité. Les PCC offrent des propriétés d'orthogonalité similaires à DS (cf Fig.4).

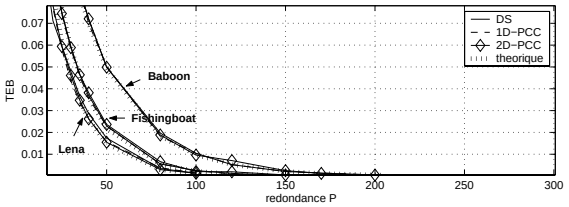


FIG. 3 – Performance au décodage par rapport à la redondance P

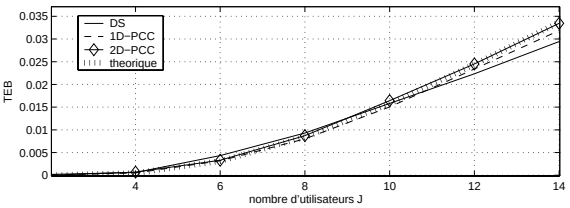


FIG. 4 – Performance au décodage par rapport au nombre d'utilisateurs J

3.3 Robustesse par rapport au bruit

Les performances de DS et PCC sont comparées en fonction du DWR (rapport entre la puissance de l'image hôte et celle du tatouage), puis du WNR (rapport entre la puissance du tatouage et celle du bruit, ou *Watermark to Noise Ratio*) dans le cas d'un bruit gaussien additif ou multiplicatif. Les PCC offrent une aussi bonne résistance au bruit que DS (cf Fig.5 et Fig.6).

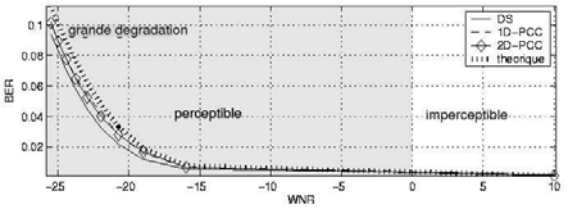


FIG. 5 – Robustesse au bruit additif

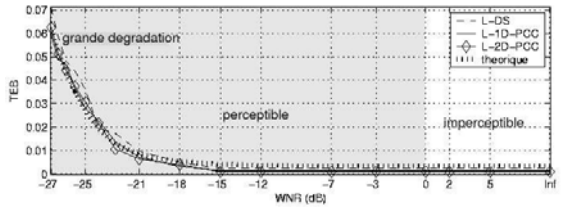


FIG. 6 – Robustesse au bruit multiplicatif

3.4 Robustesse à des attaques sophistiquées

Attaques désynchronisantes : les attaques dites géométriques (rognage, rotation, translation...) provoquent une désynchronisation entre l'image tatouée et la clé (code ou permutation). En effet, le calcul de la corrélation (DS) ou de la permutation inverse (PCC) conduisent à un décodage totalement erroné lorsqu'il est effectué sur des vecteurs ou matrices légèrement décalés. Plusieurs solutions (qui sortent du cadre de cette étude) utilisent l'insertion dans des domaines transformés appropriés ou l'insertion d'un signal de synchronisation [7]. Les deux méthodes peuvent être appliquées indifféremment à PCC ou à DS. Cependant, il n'existe pas encore de méthode de synchronisa-

tion robuste à des transformations géométriques locales et non affines et à l'élimination du signal de synchronisation.

Changement d'échelle, filtrage de Wiener, compression JPEG : les performances des trois algorithmes sont comparées par rapport à trois attaques sophistiquées classiques. Le changement d'échelle consiste à diminuer par 4 la taille de l'image transmise. A la réception, l'image est agrandie à sa taille originale par interpolation bilinéaire. L'attaque par filtrage de Wiener consiste à considérer que le tatouage est un bruit additif, et que l'image est le signal à débruiter. Les PCC offrent une aussi bonne résistance à ces attaques que DS (cf Fig.8, Fig.9). Les algorithmes sont plus robustes dans le domaine transformé.

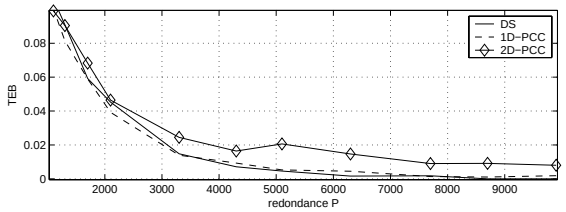


FIG. 7 – Robustesse au changement d'échelle en fonction de la redondance

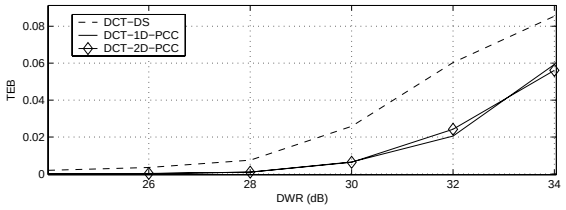


FIG. 8 – Robustesse au filtrage de Wiener

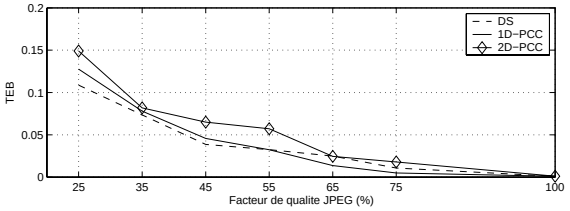


FIG. 9 – Robustesse à la compression JPEG

4 Conclusion

Les permutations aléatoires ont été introduites dans cette étude dans le cadre théorique général des PCC. Un tatouage par étalement de spectre PCC a été proposé et comparé à l'algorithme de référence DS, dans le domaine spatial et celui de la DCT par blocs. Les performances de PCC et DS sont globalement similaires et correspondent aux propriétés usuelles des techniques d'étalement de spectre : une grande robustesse au bruit additif et multiplicatif mais une vulnérabilité aux attaques désynchronisantes. Comme attendu, l'insertion dans le domaine de la DCT permet une meilleure robustesse aux attaques sophistiquées. 2D-PCC semble préférable à 1D-PCC du fait de son coût calculatoire plus faible et de ses meilleures performances pour une valeur adéquate de la période de la permutation.

A l'aide d'arguments théoriques et expérimentaux, cette étude montre que l'étalement par PCC offre des résultats globalement équivalents à ceux des techniques d'étalement classiques utilisant une modulation par un pseudo-bruit, ainsi qu'un bon niveau de sécurité. Elle justifie également l'emploi des permutations aléatoires en tant que technique d'étalement alternative dans les algorithmes existants. Les PCC basés sur les permutations aléatoires sont très simples de concept, d'implantation et de calcul et peuvent être introduits dans divers algorithmes de tatouage, concernant divers domaines transformés. On pourrait également conseiller l'utilisation d'étalement par PCC pour le tatouage de sons ou vidéos, où la redondance serait plus importante et la périodicité serait mieux exploitée. Les propriétés de filtres LPTV plus généraux, qui permettent d'effectuer simultanément un étalement et un modelage du spectre, sont à l'étude.

Références

[1] T. Furon, P. Duhamel. An asymmetric watermarking method. *IEEE. Trans. on Signal Proc.*, 51(4) :981–995, 2003.

[2] D. Kundur, D. Hatzinakos. Digital Watermarking Using Multiresolution Wavelet Decomposition. *IEEE ICASSP'98*, 5 :2659–2662, 1998.

[3] D. McLernon. One-dimensional Linear Periodically Time-Varying structures : derivations, interrelationships and properties. *IEE Proc.-Vis. Image Signal Proc.*, 146(5) :245–252, 1999.

[4] A.J. Ahumada, H.A. Peterson. Luminance-model-based DCT quantization for color image compression. *Proc. SPIE on Human Vision, Visual Proc., and Digital Display III*, 1666 :365–374, 1992.

[5] J.R. Hernández, F. Pérez-González. Statistical analysis of watermarking schemes for copyright protection of images. *IEEE Proc., Special Issue on Identification and Protection of Multimedia Information*, 87(7) :1142–1166, 1999.

[6] J.G. Proakis. *Digital Communications*. McGraw, NY, 4th edition, 2001.

[7] S. Pereira, T. Pun. Fast robust template matching for affine resistant image watermarking. *IEEE Transactions on signal proc.*, 51(4) :1045–1053, 2003.

[8] B. Lacaze, D. Roviras. Effect of random permutations applied to random sequences and related applications. *Signal Processing*, 82 :821–831, 2002.

[9] M. Coulon, D. Roviras. MMSE Joint Detection for an Asynchronous Spread-Spectrum System Based on Random Permutations. *IEEE ICASSP'04, Proc.*, 2 :17–21, 2004.

[10] I.J. Cox, J. Kilian, F.T. Leighton, T. Shamoon. Secure spread spectrum watermarking for multimedia. *Image Processing, IEEE Trans. on*, 6(12) :1673–1687, 1997.

[11] D. Roviras, B. Lacaze, N. Thomas. Effects of Discrete LPTV on Stationary Signals. *IEEE ICASSP'02, Proc.*, 2 :1127–1220, 2002.

[12] S. Verdu. *Multiuser Detection*. Cambridge University Press, 1998.

Stéganographie d'images couleur dans l'espace L*a*b* par modulation de distance

Gaël CHAREYRON[1], Dinu COLTUC[2], Alain TRÉMEAU[1]

[1]Laboratoire LIGIV - Université Jean Monnet
10 rue Barrouin, 42000 Saint Etienne, France

[2]Laboratoire FIE - Université Valahia Targoviste
BP 16-296, Bucarest, Roumanie

chareyron@ligiv.org, tremeau@ligiv.org
coltuc@valahia.ro

Résumé – Dans ce papier, nous proposons une nouvelle approche de stéganographie vectorielle afin de cacher un message dans une image couleur par modulation de distances dans l'espace $L^*a^*b^*$. Afin de minimiser les artefacts visuels, nous introduisons une méthode qui permet de moduler la force de l'insertion en fonction d'un critère d'invisibilité donné. Le principal atout de cette approche est de pouvoir marquer une image couleur avec un message d'au moins 8 bits par pixel sans diminuer la qualité de l'image résultante.

Abstract – In this paper, we propose a vectorial approach of steganography in order to hide a message in a color image by modulation of distances computed in the $L^*a^*b^*$ color space. In order to minimize the visual artefacts, we introduce a method which enables to modulate the weight of the embedding in function of a given invisibility criteria. The main interest of this approach is to be able to watermark a color image with a message with at least 8 bits per pixel without decreasing the quality of the resulting image.

1 Introduction

Dans ce papier nous nous intéressons au problème de cacher un message dans une image couleur par modulation de distances dans l'espace $L^*a^*b^*$. Alors que les méthodes usuelles séparent les trois plans couleurs et utilisent une approche scalaire pour insérer le message afin de minimiser les artefacts visuels [6, 4, 5, 3, 1], nous proposons une approche vectorielle qui permet de moduler la force de l'insertion en fonction d'un critère d'invisibilité donné. L'idée est d'insérer une signature en chacun des pixels à partir du calcul de la distance couleur qui sépare chaque paire de 2 pixels consécutifs définie par une clé secrète. L'extension à des groupes de p pixels est également abordée. Cette signature repose sur une sélection de couleurs qui vérifient une certaine contrainte liée à la minimisation de l'écart couleur qui sépare la couleur d'origine de la nouvelle couleur. Le principal atout de cette approche, confirmé par plusieurs séries de tests, est que l'on peut notamment marquer une image couleur avec un message d'au moins 8 bits par pixel sans diminuer trop significativement la qualité de l'image résultante.

2 Principe du schéma d'insertion

La technique d'insertion que nous proposons consiste à insérer un message dans un ensemble donné de n groupes de p pixels consécutifs. Afin d'illustrer la capacité des diverses schémas d'insertion qui peuvent être définis selon cette approche, nous avons choisi de présenter différentes stratégies basées sur le marquage d'une suite donnée de couples de pixels (i.e. $p = 2$).

Afin d'avoir un certain niveau de sécurité, nous proposons de sélectionner à partir d'une clé secrète l'ensemble des $n \times p$ pixels à marquer. En effet, un très grand nombre de partitions en groupes de p pixels peut être défini pour un ensemble donné de n groupes. Par contre, l'ordre dans lequel sont traités ces p pixels à l'intérieur de chaque groupe n'a pas de répercutions sur le nombre de combinaisons. Cette clé a pour principale fonction de définir le chemin qui aura été suivi dans l'image afin de marquer celle-ci. Différentes possibilités peuvent bien entendue être utilisées afin de constituer cet ensemble, par exemple une suite chaînée de n groupes consécutifs ou un ensemble donné aléatoire de n groupes indépendants. Nous avons considéré que le cardinal de l'ensemble des n groupes de p pixels consécutifs à marquer devait être égal au nombre P de pixels de l'image, et que tout pixel devait contribuer à p groupes afin que la capacité du schéma d'insertion soit maximal.

Soit I le tableau, de dimension $n \times p$, des pixels qui auront été marqués. Notons $j = 1, \ldots, n \times p$ l'indice de parcours du tableau I. Considérons à titre d'exemple le cas d'une liste chaînée avec $p = 2$, alors tout pixel P_i recensé dans le groupe j doit nécessairement être également recensé dans le groupe $j - 1$. Enfin notons également C la palette des couleurs de l'image d'origine.

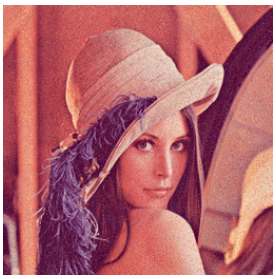
Le principe que nous avons utilisé consiste à calculer dans un premier temps la distance couleur qui sépare ces p pixels, puis d'utiliser une fonction de modulation sur cette distance afin de modifier ces pixels.

L'espace de travail dans lequel le calcul des distances couleur est effectué est l'espace $L^*a^*b^*$, ce dernier étant par définition plus à même (au sens de la perception visuelle) de décrire correctement de petites différences de couleur [7].

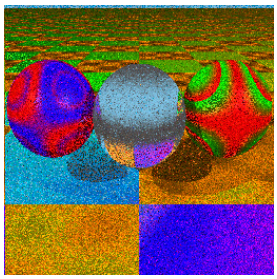
2.1 Cas où $p = 2$

Pour illustrer l'approche que nous avons développée, illustrons tout d'abord un des cas d'étude les plus simples; celui de l'insertion à la chaîne de deux pixels consécutifs pour lequel chaque pixel est marqué en fonction de son prédécesseur.

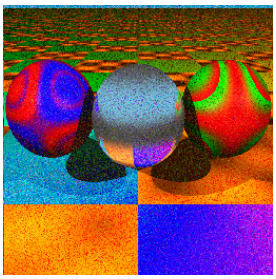
(a) Insertion d'un message sans ajout de couleurs (PSNR=29,06).



(b) Insertion d'un message avec ajout de couleurs (PSNR=30,86).



(c) Insertion d'un message sans ajout de couleurs (PSNR=18,12).



(d) Insertion d'un message avec ajout de couleurs (PSNR=23,02).

(e) Insertion d'un message sans ajout de couleurs (PSNR=28,10).

(f) Insertion d'un message avec ajout de couleurs (PSNR=30,77).

FIG. 1: *Comparaison de la qualité de l'image obtenue sans insertion de nouvelles couleurs ou avec (boite $3 \times 3 \times 3$), pour un message de 8 bpp.*

Considérons dans un premier temps que nous voulons insérer un message m de longueur $\log_2 M$ bits (e.g. 8 bits pour $M = 256$) entre un pixel donné P_i et le pixel suivant P_{i+1}. Notons $\vec{O}_i$ la couleur du pixel P_i dans l'image d'origine et $\vec{M}_i$ la couleur modifiée après marquage. Notons $\Delta E_{ab}(\vec{O}_i,\vec{M}_i)$ l'écart couleur entre la couleur d'origine du pixel P_i et sa nouvelle couleur après marquage. Cette distance est calculée dans l'espace couleur $L^*a^*b^*$ à partir de la formule CIE 1976 ΔE_{ab} [7].

Pour chaque pixel P_i, il est possible de trouver une couleur pour laquelle :

$$\left[\Delta E_{ab}(\vec{O}_i,\vec{M}_i) \times c\right] \bmod M = m \qquad (1)$$

où $[.]$ représente la fonction partie entière et c est une constante d'echelle.

Le schéma d'insertion que nous avons développé consiste à marquer chaque pixel P_{i+1} de chaque couple de pixels (P_i,P_{i+1}), de façon à ce que tous les pixels de l'image soient marqués une et une seule fois.

Pour se faire deux stratégies peuvent être utilisées.

La première stratégie consiste à rechercher parmi l'ensemble C la couleur $\vec{M}$ qui vérifie l'expression (1), puis à changer la couleur du pixel P_{i+1} par cette valeur, i.e. $\vec{M}_{i+1} = \vec{M}$. Cette méthode permet de ne pas accroître le nombre de couleur de l'image générée tout en conservant la palette des couleurs de l'image originale.

La deuxième stratégie consiste a:

1. générer, dans l'espace des couleurs $L^*a^*b^*$, un boite englobante (e.g. un cube) centré sur la couleur $\vec{O}_{i+1}$ du pixel P_{i+1},
2. rechercher dans cette boite englobante la couleur $\vec{M}$ qui vérifie l'expression (1),
3. changer la couleur du pixel P_{i+1} par cette valeur, i.e. $\vec{M}_{i+1} = \vec{M}$.

Cette méthode réduit le champ d'investigation à un ensemble de couleurs plus cohérentes, ce qui revient à minimiser la perte de qualité résultante, au dépend éventuellement d'une augmentation du nombre de couleurs.

Quelque soit la stratégie utilisée, nous pouvons être confronté à la situation où plusieurs couleurs $\vec{M}$ satisfont aux conditions que l'on s'est fixé. Nous avons donc établi une contrainte supplémentaire qui consiste à ne retenir que le candidat qui minimise la distance $\Delta E_{ab}(\vec{O}_{i+1},\vec{M}_{i+1})$.

Dans l'hypothèse ou aucune couleur $\vec{M}$ ne satisfait aux conditions que l'on s'est fixé, cas qui ne peut se produire que dans le cadre de la seconde stratégie, il suffit alors d'augmenter la taille de la boite englobante jusqu'à obtenir au moins un candidat.

Ce processus est itéré tant que la liste du tableau I n'est pas entièrement parcourue, afin de marquer tous les pixels de l'image. A l'itération $i + 1$ le pixel P_{i+1} devient pixel pivot, il n'est alors pas modifié, par contre le pixel P_{i+2} est à son tour modifié en fonction de la couleur d'origine $\vec{O}_{i+1}$ du pixel P_{i+1}.

Les deux stratégies de marquage que nous avons proposé ne présentent pas les mêmes avantages. La première permet de ne pas rajouter de nouvelles couleurs à l'image, ce qui permet de garder un nuage couleur, une palette de couleurs, sensiblement identique à l'image originale. A contrario, la seconde stratégie permet la création de nouvelles couleurs qui vont perturber plus ou moins fortement le nuage couleur, tout en assurant une meilleure invisibilité perceptuelle de la marque (cf. images de la colonne de gauche/colonne de droite dans la Fig. 1). Plus les boites englobantes sont denses, i.e. plus il y a de candidats qui à la fois appartiennent à l'ensemble C et aux boites englobantes, plus le nuage couleur reste stable. A titre d'exemple, à l'inverse

des images « Lena » et « Mandrill» pour lesquels la distribution des couleurs est dense, l'image «Balls» est constituée de couleurs clairsemées.

Etant donné que l'on insère un message de taille $\log_2 M$ dans chaque pixels, i.e. dans P pixels, la capacité du schéma est égal à $\log_2 M$ bpp.

2.2 Cas où $p > 2$

Dans le cas ou $p > 2$, la capacité de l'image support diminue. Néanmoins l'utilisation d'un nombre de pixels plus important permet de diluer encore plus la signature dans l'image. Toute la difficulté consiste à définir un mode de marquage des pixels qui permet d'optimiser la qualité d'image tout en préservant la capacité.

3 Principe de détection, cas où $p = 2$

La détection du message s'effectue grâce à la clé secrète qui nous permet de retrouver dans quel ordre les pixels de l'image ont été marqués. Nous pouvons alors reconstituer le tableau I des couples de pixels marqués, puis à partir de l'expression $\Delta E_{ab}(\vec{O}_{i+1}, \vec{M}_{i+1})$ retrouver le message m qui a été inséré en chaque pixel P_{i+1}. L'intégralité du message est alors décodé lorsque tous les pixels de l'image ont été parcourus, i.e. tous les couples de pixels du tableau I.

4 Evaluation de la capacité et de la qualité

Afin de tester la capacité et de cette méthode, nous avons testé différentes tailles de message. Dans ce papier nous avons choisi de présenter quelques résultats illustratifs calculés à partir d'images tests de taille 512×512. Afin de tester l'invisibilité de l'insertion du message nous avons calculé tout d'abord le PSNR (cf. Fig. 2), puis le ΔE_{ab} moyen entre l'image d'origine et l'image marquée, ainsi que l'écart type associé (cf. Fig. 3). Les résultats que l'on a obtenu confirment que la mesure du ΔE_{ab} moyen est plus pertinente que la mesure du PSNR pour décrire la perte de qualité visuelle générée par un processus donné [2].

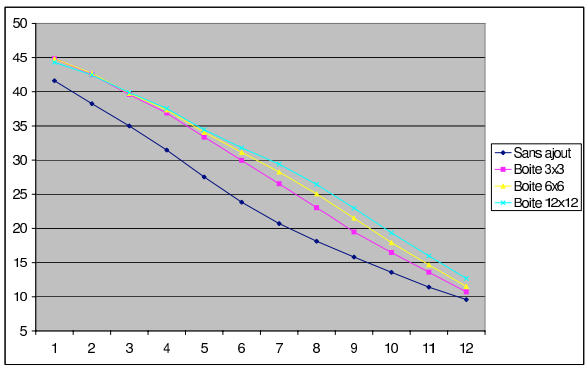
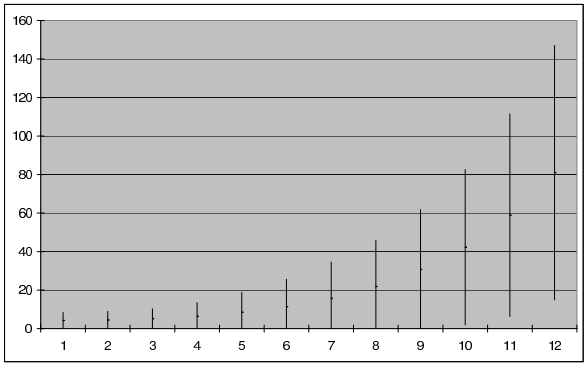
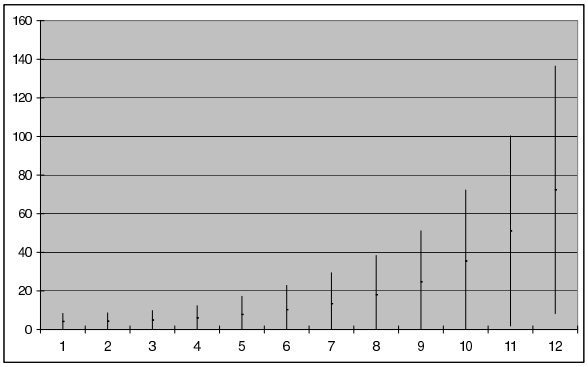


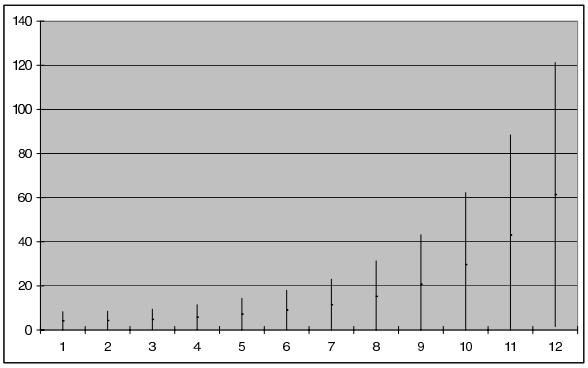
FIG. 2: *Evolution du PSNR en fonction de la quantité d'information insérée.*



(a) Pour une boite englobante $3 \times 3 \times 3$.



(b) Pour une boite englobante $6 \times 6 \times 6$.

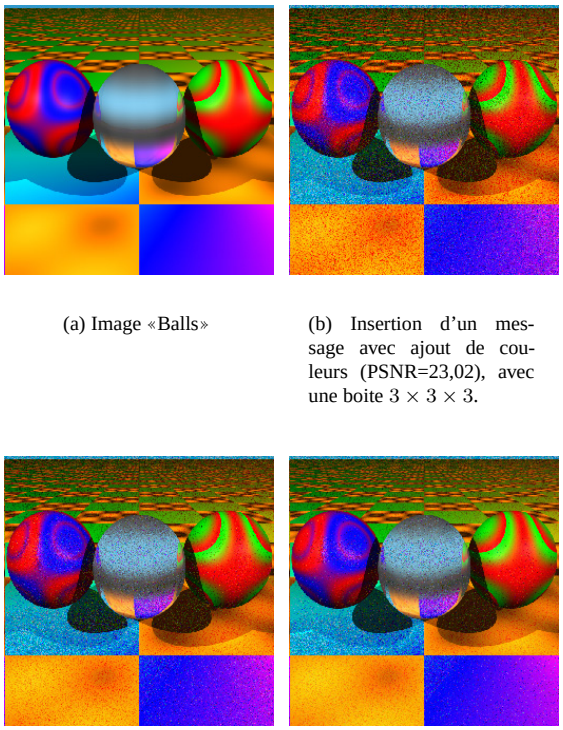


(c) Pour une boite englobante $12 \times 12 \times 12$.

FIG. 3: *Evolution du ΔE_{ab} moyen et de l'écart type associé en fonction de la quantité d'information insérée, pour différentes tailles de boites englobantes.*

Nous avons ainsi inséré entre 1 bit/pixel et 12 bit/pixels. Nous avons également utilisé des boites englobante de taille $3 \times 3 \times 3$ à $12 \times 12 \times 12$ dans l'espace des couleurs $L^*a^*b^*$ (cf Fig. 4). Au-delà d'une certaine taille, en l'occurrence $9 \times 9 \times 9$, tout ajout de couleurs devient perceptible, d'autant plus perceptible quand il intervient dans une zone localement homogène.

Afin de rester dans le cadre d'une relative invisibilité de la marque, nous n'avons pas considéré des boites englobantes de plus grande taille.



(a) Image «Balls»

(b) Insertion d'un message avec ajout de couleurs (PSNR=23,02), avec une boite $3 \times 3 \times 3$.

(c) Insertion d'un message avec ajout de couleurs (PSNR=25,06), avec une boite $6 \times 6 \times 6$.

(d) Insertion d'un message avec ajout de couleurs (PSNR=26,44), avec une boite $12 \times 12 \times 12$.

FIG. 4: *Qualité de l'image obtenue en fonction de la taille de la boite englobante, pour un message de 8 bpp.*

A supposer qu'aucune boite englobante de taille inférieure à $9 \times 9 \times 9$ centrée en une couleur donnée, ne satisfait aux conditions que l'on s'est fixé (dans le cadre de la seconde stratégie pour l'insertion du message), nous avons pris le parti de nous replacer dans le cadre de la première stratégie, i.e. d'étendre la recherche à l'ensemble C de toutes les couleurs de l'image originale.

La Figure 2 permet d'analyser comment varie le PSNR en fonction de la taille du message à insérer et de la taille des boites englobantes. On constate que, pour de petites capacités, plus la taille de la boite englobante est importante moins le PSNR est bon. Par contre, pour de plus grandes capacités, la qualité devient meilleure. En effet, plus la boite englobante est de taille importante plus on s'éloigne des couleurs originales de l'image d'où un PSNR moins bon pour les petites capacités. En conséquence, si l'on veut éviter l'apparition de fausses couleurs particulièrement visibles, il ne faut pas utiliser des boites englobantes de taille supérieure à du $9 \times 9 \times 9$.

De même, plus la longueur du message inséré dans l'image est important plus la qualité de l'image résultat diminue. Cela étant, si l'on veut éviter une perte de qualité trop importante (e.g. un PSNR inférieur à 30), la méthode proposée ne permet pas d'aller au-delà d'un message de taille 10 bpp, ce qui en soit est déjà conséquent. Au dela de 10 bpp le message est déformé. On ne trouve pas toujours un couple de couleur qui permet l'insertion du message.

5 Conclusion

Nous avons présenté une méthode de tatouage d'image couleur haut débit orienté vers la stéganographie et l'authentification d'image. Cette méthode permet une meilleure invisibilité du tatouage qu'une méthode traditionnelle utilisant une séparation des trois composantes. Les deux stratégies présentées permettent de contrôler le nombre de couleurs de l'image finale afin de passer inaperçu aux techniques de détection statistique.

Des travaux sont en cour pour généraliser cette approche à des groupes de p pixels ($p > 2$) de façon à minimiser plus encore la perte de qualité résultante, i.e. de lisser spatialement le bruit résultant qui se superpose à l'image.

Références

[1] J. Fridrich, R. Du, and M. Long. Steganalysis of lsb encoding in color images. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2000.

[2] D. Coltuc G. Chareyron and A. Trémeau. Watermarking and authentication of color images based on segmentation of the xyy color space. In *to be published in JIST*, 2005.

[3] N. Memon I. Avcibas and B. Sankur. Steganalysis using image quality metrics. In *IEEE Transactions on Image Processing, vol. 12*, pages 221–229, 2003.

[4] M. Goljan J. Fririch and R. Du. Reliable detection of lsb steganography in color and grayscale images. In *Proceedings of the ACM Workshop on Multimedia and Security*, pages 27–30, 2001.

[5] E. Kawaguchi M. Niimi, H. Noda and R. O. Eason. Luminance quasi-preserving color quantization for digital steganography to palette-based images. In *Proceedings of ICPR'02, Vol. 4*, page 102252, 2002.

[6] D. Hatzinakos P. Campisi, D. Kundur and A. Neri. Compressive data hiding: An unconventional approach for improved color image coding. In *JASP vol 2*, pages 152–163, 2002.

[7] A. Trémeau, C. Fernandez-Maloigne, and P. Bonton. *Images Couleur : de l'acquisition au traitement*. Dunod, 2004. Collection Science Sup.

Insertion d'informations numériques dans un signal : Application à la classification de données expérimentales en anémométrie laser Doppler

Alain LE DUFF[1], Bruno GAZENGEL[2]

[1]École Supérieure d'Électronique de l'Ouest, 4, Rue Merlet de la Boulaye, 49009 Angers cedex 01, France

[2]Laboratoire d'Acoustique de l'Université du Maine, Av. O. Messiaen, 72085 Le Mans cedex 09, France

alain.le_duff@eseo.fr, Bruno.Gazengel@univ-lemans.fr

Résumé – Ce travail propose d'évaluer une procédure de marquage de signaux expérimentaux en anémométrie laser à effet Doppler. La méthode adoptée est basée sur une technique d'étalement de spectre et des mesures de corrélation. L'impact du marquage sur la qualité des estimations des paramètres de la vitesse particulaire acoustique est mesuré de manière objective à l'aide d'une méthode basée sur la dérivée de la phase du signal.

Abstract – The aim of this work is to propose to evaluate a watermarking procedure for experimental signals in laser Doppler anemometry. The adopted method is based on a spread spectrum technique and correlation measurements. The impact of the watermark on the quality of the estimation of the particle acoustic velocity parameters is measured using a phase derivative based estimator method.

1 Introduction

La mesure de vitesse particulaire acoustique par anémométrie laser Doppler (ALD) génère des quantités de données considérables. Dans un contexte de traitement en temps différé de ce type de données il est indispensable de pouvoir leur associer un certain nombre d'informations concernant, par exemple, les paramètres du système d'acquisition (fréquence d'échantillonnage, date, etc.). Une solution consiste simplement à positionner ces renseignements dans l'entête du fichier accueillant les données issues de l'acquisition. Une alternative consiste à insérer directement ces informations dans les données acquises. Une telle méthode de classification des signaux d'acquisition permet d'envisager, par exemple, la mise en place d'une base de donnée de signaux ALD partagée par plusieurs laboratoire de recherche. Les données peuvent ainsi être échangées de manière discrète et indépendante du format des fichiers. Pour cela il est indispensable de choisir une méthode de marquage à la fois robuste et ne perturbant que très faiblement les mesures des paramètres de la vitesse acoustique. Dans ce papier nous proposons d'évaluer, à titre de faisabilité, une procédure de marquage basée sur une technique d'étalement de spectre et des mesures de corrélation. L'impact du marquage sur la qualité des estimations des paramètres de la vitesse acoustique est par la suite mesuré de manière objective à l'aide d'une méthode basée sur la dérivée de la phase du signal Doppler (PDBE - Phase Derivative Based Estimator).

2 Signaux Doppler

Le principe de la mesure de vitesse particulaire acoustique par ALD est décrit en détail dans [1]. Les expériences que nous menons supposent que la vitesse particulaire acoustique $V(t)$ se compose d'une composante continue (vitesse d'écoulement due à la convection naturelle) à laquelle se superpose une composante sinusoïdale (excitation acoustique) de fréquence $F_{\rm ac}$:

$$V(t) = V_{\rm ec} + V_{\rm ac}\cos(2\pi F_{\rm ac}t + \phi_{\rm ac}) \qquad (1)$$

Les paramètres du signal acoustique sont la fréquence acoustique $F_{\rm ac}$, supposée connue, la phase acoustique initiale $\phi_{\rm ac}$ et l'amplitude de la vitesse acoustique $V_{\rm ac}$ que l'on cherche à estimer. La vitesse de convection, $V_{\rm ec}$, supposée petite devant $V_{\rm ac}$, doit elle aussi être estimée. Dans l'hypothèse où une seule particule d'ensemencement traverse le volume de mesure, le signal Doppler prend la forme

$$s(t) = A(t)\left\{M + \cos\left[2\pi F_{\rm B}t + 2\pi D x(t) + \phi_0\right]\right\}. \qquad (2)$$

où $x(t)$ est la position de la particule ($V(t) = dx(t)/dt$) et $F_{\rm B}$ la fréquence porteuse (de l'ordre de 40 MHz). Dans la relation (2), $A(t)$ représente l'amplitude variable du signal avec

$$A(t) = K\exp\left\{-\left[\beta x(t)\right]^2\right\} \qquad (3)$$

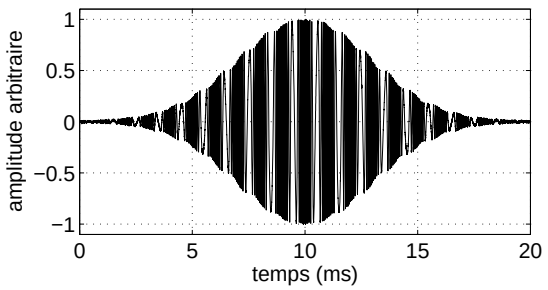


FIG. 1: *Composante en phase, $y_1(t)$, du signal Doppler.* $F_{\rm ac}$ = 500 Hz.

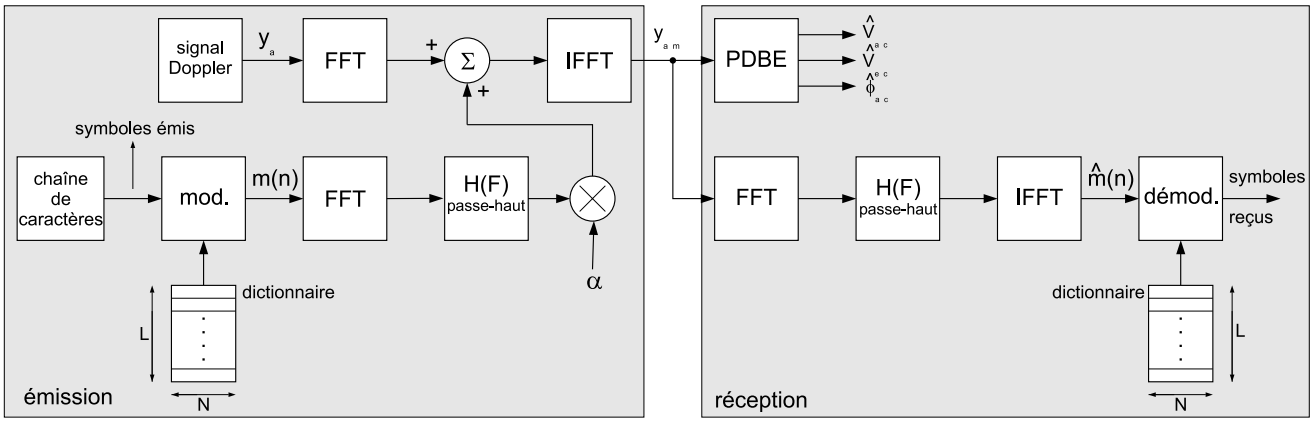


FIG. 2: *Système de marquage.*

où le terme $1/\beta$ dépend du système de mesure. La phase instantanée est donnée, quant à elle, par

$$2\pi Dx(t) + \phi_0 = 2\pi F_{\mathrm{D}} t + \phi_0 = \phi(t). \quad (4)$$

Par la suite la composante basse-fréquence $MA(t)$ est éliminée à par un filtrage passe-haut. L'équation (2) devient alors

$$s(t) = A(t)\cos\left[2\pi F_{\mathrm{B}} t + \phi(t)\right]. \quad (5)$$

Enfin, le spectre de puissance de $s(t)$ est ramené en bande de base à l'aide d'un démodulateur en quadrature analogique synchrone de la fréquence F_{B}. Comme les spectres de $A(t)$ et de $\cos\phi(t)$ présentent des supports fréquentiels disjoints une bonne approximation des signaux $y_1(t)$ et $y_2(t)$ peut être donnée par

$$y_1(t) = A(t)\cos\phi(t), \quad (6)$$
$$y_2(t) = A(t)\sin\phi(t). \quad (7)$$

Cette opération est analogue à une transformée de Hilbert. Après échantillonnage nous obtenons les signaux à temps discret $y_1(k)$ et $y_2(k)$ donnés par

$$y_1(k) = A(k)\cos\phi(k) + b_1(k) \quad (8)$$
$$y_2(k) = A(k)\sin\phi(k) + b_2(k) \quad (9)$$

Dans les relations (8) et (9), $b_1(k)$ et $b_2(k)$ sont deux bruits blancs, indépendants, à moyennes nulles et de variance identique. Les deux composantes $y_1(k)$ et $y_2(k)$ sont destinées à accueillir les informations numériques.

La figure 1 donne, à titre d'exemple, l'allure d'un signal d'observation $y_1(k)$ synthétisé numériquement.

3 Principes de la méthode de marquage

La procédure de marquage s'inspire d'une méthode décrite dans [2]. Son principe est illustré par la figure 2. L'information binaire est codée à l'aide d'un dictionnaire $\mathcal{D}$ associant à chaque symbole à transmettre (m-uplet binaire) une séquence aléatoire blanche gaussienne de longueur N. Ce dictionnaire est donc constitué de M signaux aléatoires, orthonormalisés par la procédure de Gram-Schmidt [3], entretenant une relation bijective avec les $M = 2^m$ symboles. Le signal ainsi obtenu est alors inséré par sommation fréquentielle (hautes-fréquences) avec le signal analytique

$$y_{\mathrm{a}}(k) = y_1(k) + jy_2(k) \quad (10)$$

associé aux signaux d'observation. Un coefficient de pondération α permet de contrôler la puissance de la marque. À la réception, les composantes HF du signal marqué sont soumises à un détecteur par corrélation. L'estimation de la valeur maximale du coefficient de corrélation entre le signal et les M signaux du dictionnaire, connu du récepteur, détermine alors les symboles émis les plus probables.

4 Estimation des paramètres de la vitesse

L'estimation de la vitesse dont le principe est détaillé dans un article précédant [4] s'effectue en trois phases : Un opérateur de type $\arctan$ permet d'extraire la phase instantanée $\phi(k)$:

$$\phi(k) = \arctan\left[\frac{y_2(k)}{y_1(k)}\right] \quad (11)$$

La vitesse est ensuite estimée en dérivant la phase ce qui conduit à

$$\widehat{V}(k) = \frac{1}{2\pi D}\frac{[\phi(k+1) - \phi(k-1)]}{2T_e} \quad (12)$$

En tenant compte de la relation (4) la vitesse estimée devient :

$$\widehat{V}(k) = V_{\mathrm{ec}} + \left(\frac{\sin 2\pi f_{\mathrm{ac}}}{2\pi f_{\mathrm{ac}}}\right) V_{\mathrm{ac}}\cos(2\pi f_{\mathrm{ac}} k + \phi_{\mathrm{ac}}). \quad (13)$$

Enfin une démodulation en quadrature, synchronisée par le signal d'excitation acoustique de fréquence F_{ac}, suivie d'une transformation de coordonnées cartésiennes en coordonnées polaires, permet d'extraire l'amplitude $\hat{V}_{\mathrm{ac}}$ et la phase $\hat{\phi}_{\mathrm{ac}}$ de la vitesse acoustique, $\hat{V}_{\mathrm{ec}}$ étant estimée, quant à elle, à l'aide d'un simple filtrage de type passe-bas. L'ensemble de cette procédure est résumé par la figure 3.

5 Résultats et conclusions

Les performances de la méthode de marquage sont évaluées à l'aide de 500 simulations de Monte-Carlo et à partir de signaux Doppler de synthèse. Pour chaque signal une marque de 310 bits est insérée avec une puissance variant de $0V^2$ (insertion nulle) à $5V^2$. Le signal Doppler recevant la marque est représenté par $y_1(k)$ et $y_2(k)$ échantillonnés chacun à une fréquence de 100 kHz et sur une durée de 80 ms ce qui correspond à 8000 échantillons.

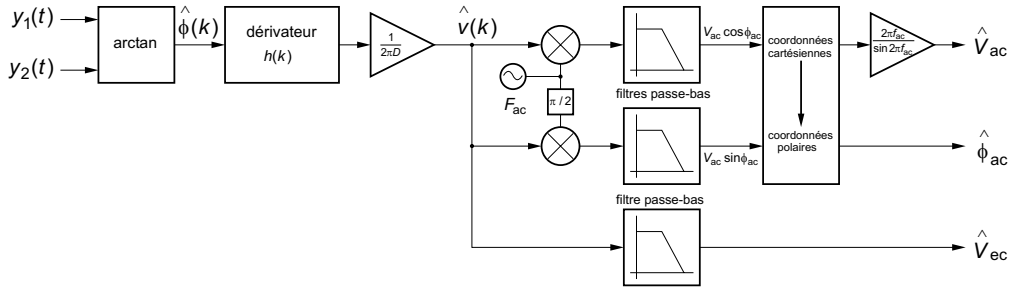


FIG. 3: *Estimation des paramètres de la vitesse acoustique.*

Dans un premier temps, l'évolution du Taux d'Erreur Binaire (TEB) moyen en fonction du rapport signal à bruit (RSB) du signal Doppler et de la puissance d'insertion α de la marque est étudié (figure 4 (a)). Dans un second temps les variances d'estimation des paramètres de la vitesse acoustique sont étudiées en fonction de α (figure 5). De plus, ces différentes variances sont comparées aux bornes de Cramer-Rao déterminées lors d'une précédente étude [1]. Les résultats présentés dans ce résumé se limitent aux valeurs suivantes : $V_{\rm ac} = 1000$ Hz, $V_{\rm ec} = 2$ mm/s et $\phi_{\rm ac} = 80°$. Le dictionnaire utilisé est, quant à lui, constitué de M = 8 vecteurs de longueur N = 64.

Dans la pratique les signaux Doppler observés présentent très fréquemment des RSB supérieurs à 15 dB. La figure 4 (b) montre alors l'évolution du taux du TEB en fonction de α pour RSB = 15 dB. Il est possible de constater que le taux d'erreur devient très faible à partir d'une puissance de $2V^2$ ($\alpha = 2$) (pas d'erreur observée pour 500 simulations). De même, dans des conditions identiques, la figure 6 montre qu'il est difficile d'observer un écart notable entre les variances des estimations des signaux marqués et de ceux qui ne le sont pas ($\alpha = 0$) pour des puissances de marque inférieures à $2,5V^2$ ($\alpha = 2,5$). Il apparait donc que, pour cette application, une puissance de marquage α proche de $2V^2$ ($\alpha \approx 2$) constitue une valeur optimale assurant un compromis entre un taux d'erreur faible et une bonne précision des estimations des paramètres de la vitesse acoustique.

Ce travail préliminaire montre que le marquage de données expérimentales en ALD est envisageable dans de bonne condition de robustesse et de précision des mesures.

References

[1] A. Le Duff, G. Plantier, J.C. Valière et R. Perdriau, *Acoustic Velocity Measurement in the Air by Means of Laser Velocimetry: Cramer-Rao Bounds and Maximum Likelihood Estimation*, IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP'2002, Orlando, may 2002.

[2] M.D. Swanson, B. Zhu, A.H. Tewfik, L. Boney. *Robust audio watermarking using perceptual masking*. Signal Processing, Vol. 66 pp. 337-355, 1998.

[3] J.G. Proakis. *Digital communications*. McGraw-Hill, 1995.

[4] A. Le Duff, G. Plantier, J.C. Valière, B. Gazengel, *Mesure de vitesse acoustique particulaire en champ libre par anémométrie laser Doppler*, 19ème colloque Gretsi sur le Traitement du Signal et des Images, Paris, septembre 2003.

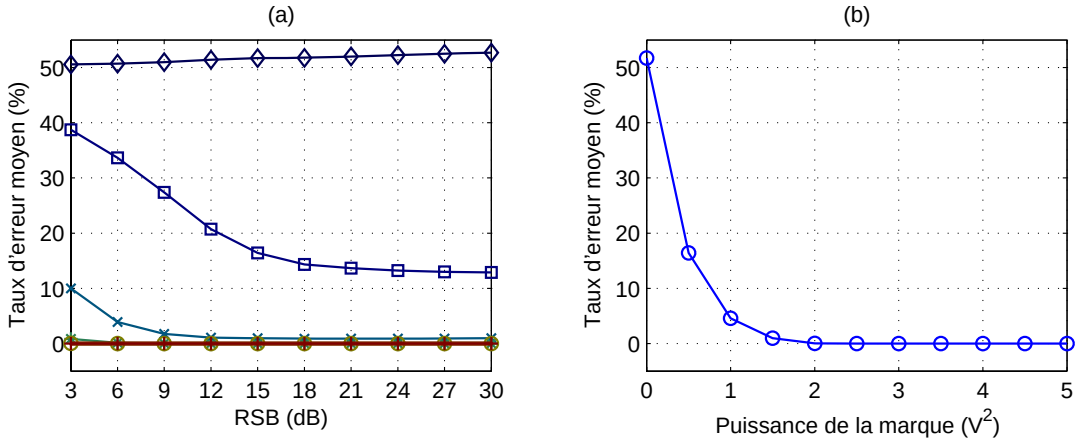


FIG. 4: (a) *Taux d'erreur en fonction du RSB pour différentes valeurs de* α *:* $-+-\ \alpha=5$, $-\circ-\ \alpha=4$, $-*-\ \alpha=3$, $-\times-\ \alpha=2$, $-\square-\ \alpha=1$, $-\diamond-\ \alpha=0$ *;* (b) *Taux d'erreur en fonction de* α *pour* RSB = 15 dB.

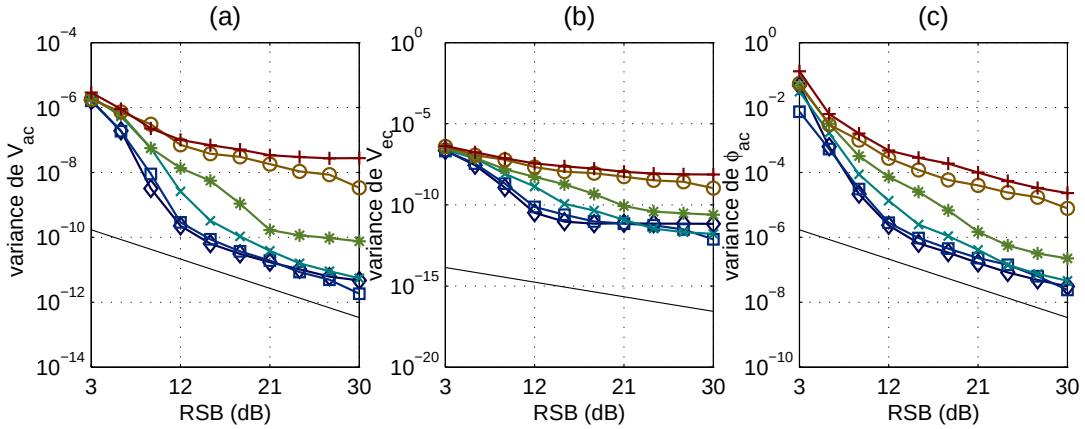


FIG. 5: *Variances d'estimation* - (a) $V_{\rm ac}$, (b) $V_{\rm ec}$, (c) $\phi_{\rm ac}$ - *en fonction du RSB pour différentes valeurs de* α *:* $-+-\ \alpha=0$, $-\circ-\ \alpha=1$, $-*-\ \alpha=2$, $-\times-\ \alpha=3$, $-\square-\ \alpha=4$, $-\diamond-\ \alpha=5$, *– CRB.*

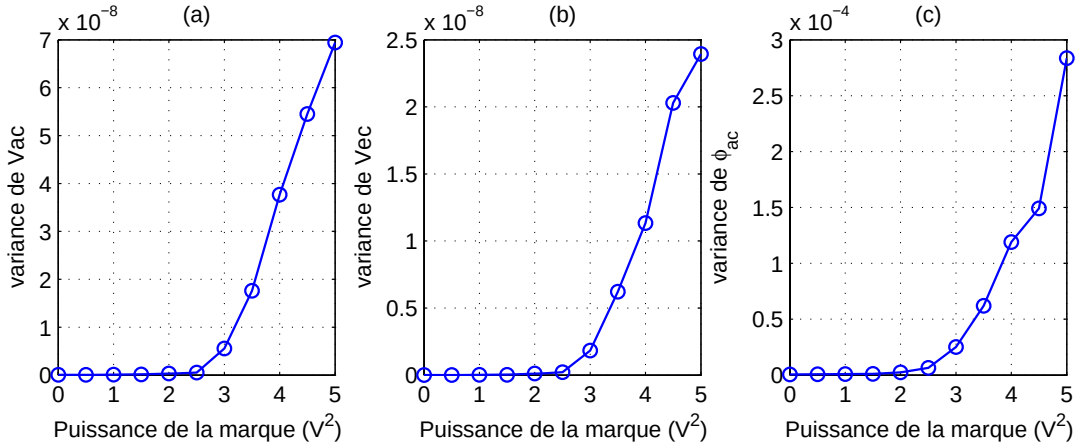


FIG. 6: *Variances d'estimation en fonction de la puissance de la marque pour* RSB = 15 dB - (a) $V_{\rm ac}$, (b) $V_{\rm ec}$, (c) $\phi_{\rm ac}$.

Transfert sécurisée d'une RI sans perte par une méthode d'insertion de données robuste à la compression JPEG

Philippe AMAT, William PUECH

Laboratoire LIRMM, UMR CNRS 5506, Université de Montpellier II
161, rue Ada, 34392 MONTPELLIER CEDEX 05, FRANCE
philippe.amat@lirmm.fr, william.puech@lirmm.fr

Résumé – Dans cet article, nous présentons une méthode d'insertion de données cachées originale permettant de protéger la haute résolution d'une région d'intérêt. Les pertes de résolution dans la région d'intérêt sont récupérées puis insérées par marquage dans l'image. L'image avec la région d'intérêt sans perte sera visible seulement après l'extraction et la reconstruction de celle-ci. La méthode d'insertion s'appuie sur les coefficients DCT et sur la modification de la matrice de quantification. Cette technique a été développée de manière à être robuste à la compression JPEG.

Abstract – In this article, we present an original method of data hidding witch allows protecting the high resolution of a region of interest. Resolution losses into region of interest are retrieved and embedded in the image. The image with the region of interest without loss will be visible only after the extraction and the reconstruction of this one. The data hidding method is based on DCT coefficients and quantization matrix modification. This method was developed so as to be robust with JPEG compression.

1 Introduction

Le transfert d'information sur les réseaux demande une forte compression des images. Il faut alors trouver un compromis entre rapidité de transfert et conservation d'une très bonne qualité de l'information. Généralement seules certaines régions de l'image présentent un intérêt majeur. De ce fait il n'est pas nécessaire de conserver une très haute qualité pour toute l'image [7, 5, 3, 6]. Une fonctionnalité semblable est offerte par le futur standard de compression JPEG 2000 : il est possible de conserver des régions d'intérêts (RI) sans perte en compressant le reste de l'image [2]. Dans cet article nous présentons une nouvelle méthode qui inclue par insertion de données cachées (IDC) l'information perdue d'une RI lors de la compression JPEG d'une image. L'information perdue représente les pertes engendrées par l'algorithme de compression JPEG dans une RI de l'image. Cette information est incluse dans les coefficients DCT par un algorithme d'IDC [1]. Dans le cas du JPEG2000 la RI sans perte est visible dès l'ouverture de l'image comprimée. Dans le cas de notre méthode, la RI sans perte est protégée à l'ouverture de l'image comprimée. En effet avec notre méthode, l'information complète de la RI n'est visible seulement qu'après l'extraction des données cachées et la reconstruction de la RI. La Section 2 détaille la méthode d'IDC adaptée de la méthode de [1]. Dans la Section 3 nous montrons les résultats obtenus sur l'image de Lena.

2 La méthode de protection proposée

2.1 Adaptation de la méthode de [1]

Dans la DCT les coefficients les plus importants pour la qualité de l'image sont généralement ceux correspondant aux basses fréquences. Les coefficients les plus quantifiés par la compression sont ceux correspondant aux hautes fréquences. La proposition de [1] est de cacher l'information dans les moyennes fréquences. Dans cette méthode les coefficients à marquer seront quantifiés par 1. Le Tableau 1 montre la table de quantification standard du JPEG pour un facteur de qualité de 50 avec la modification de 26 coefficients comme le suggère [1]. De ce fait, le taux de compression diminue alors que la qualité de l'image augmente. Dans cette méthode 2 bits du message sont marqués par substitution sur les deux bits de poids faible de 26 coefficients moyennes fréquences, soit une capacité de 52 bits par bloc. Cette méthode va donc nous permettre de dissimuler des messages de taille conséquente et surtout de connaître à l'avance la capacité de l'IDC. Pour notre problème, la quantité de données à insérer est variable. En effet, en fonction de la taille de la RI et du facteur de compression choisi la quantité de données varie. Il a donc fallu adapter cette méthode proposée par [1]. Dans un premier temps nous récupérons la totalité des pertes de la RI (Section 2.2). En connaissant la quantité de bits du message à cacher, nous pouvons ensuite moduler le nombre de 1 dans la matrice de quantification (MQ) (Section 2.4). La Figure 1 montre que ce processus a une phase d'optimisation, les modifications de la MQ affectant directement les pertes de données dans la RI. Cette méthode permet d'insérer une quantité de 1 plus optimale

que dans la méthode initiale de [1]. De cette façon le taux de compression est meilleur et la qualité de l'image est plus proche de celle du JPEG original.

TAB. 1 – Matrice de quantification modifiée pour un facteur de qualité de 50

16	11	10	16	1	1	1	1
12	12	14	1	1	1	1	55
14	13	1	1	1	1	69	56
14	1	1	1	1	87	80	62
1	1	1	1	68	109	103	77
1	1	1	64	81	104	113	92
1	1	78	87	103	121	120	101
1	92	95	98	112	100	103	99

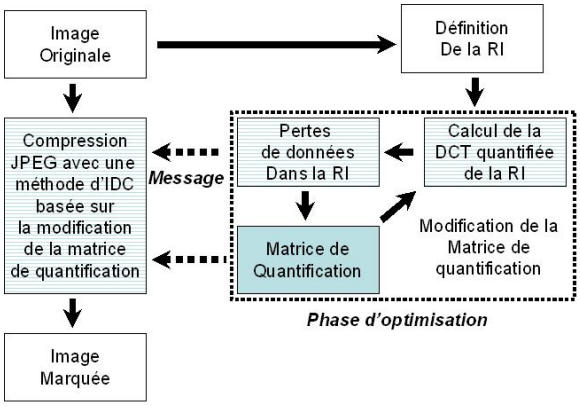


FIG. 1 – La méthode proposée de protection de données.

2.2 Récupération des pertes dans la RI

Afin de proposer une méthode de reconstruction d'une RI sans perte, il a fallu s'intéresser aux parties de l'algorithme qui engendrent des pertes et à la façon de les récupérer. Les pertes provenant de la quantification ne posent pas de problème quant à leur récupération. Le problème principal vient du fait que certaines pertes proviennent de l'arrondi de l'IDCT lors de la phase de décompression. Il faut aussi prendre en compte que la méthode d'IDC engendre également des pertes. La Figure 2 montre comment on arrive à calculer la totalité des pertes engendrées par la compression, la decompression et l'IDC sur chaque bloc de la RI. Pour les pertes dues a l'IDC on positionne à zéro tous les bits de poids faible dont on va se servir pour effectuer l'IDC (LSB0 et LSB1). De ce fait, lors de la phase de récupération des données cachées on replacera les bits à zéro après lecture. Cela nous permet d'avoir une méthode de calcul des pertes qui ne dépend pas du message à marquer et donc d'éviter que l'IDC fasse varier le nombre d'erreurs. Avec cette méthode on s'assure que les pertes sont bien la différence entre la RI de l'image originale et la RI de l'image que l'on va reconstruire.

2.3 Codage des pertes

Les pertes sous cette forme ne nous permettent pas d'effectuer l'IDC. Etant donné qu'elles vont être insérées puis récupérées sous la forme d'un flux binaire, il est nécessaire de les mettre en forme pour faciliter leur extraction. La méthode choisie est la méthode de Huffman [4] qui utilise des tables de préfixes pour encadrer les valeurs à coder. Cette méthode comporte un avantage dans notre cas car chaque section de coefficient sera codée avec le même nombre de bits. Ainsi il nous est d'autant plus facile, pendant la phase d'optimisation, d'évaluer la quantité de bits à insérer sans avoir à effectuer le codage de Huffman. Les tables de préfixes utilisées sont celles inclues dans le JPEG.

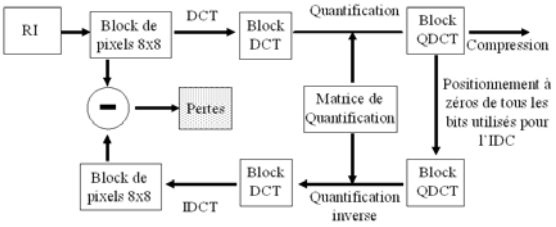


FIG. 2 – Récupération des pertes.

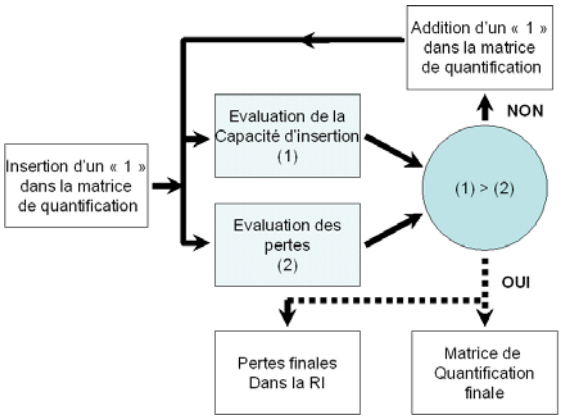


FIG. 3 – Modification de la matrice de quantification.

2.4 Modification de la matrice de quantification

La récupération des pertes, Section 2.2, nous permet d'évaluer précisément la quantité de données perdues. Cette quantité va nous permettre de calculer exactement la quantité de 1 à insérer dans la matrice de quantification afin de pouvoir réaliser l'IDC. La Figure 3 montre comment la MQ est modifiée afin d'avoir le nombre optimum de 1 nécessaire à l'IDC. Le fait d'insérer des 1 dans la MQ modifie la quantité de pertes de la RI. En effet, certains coefficients ne seront plus quantifiés. Cette opération est donc réalisée plusieurs fois afin que les pertes de données et le nombre de 1 dans la MQ soient optimums. Le résultat de cette optimisation nous concède ainsi la MQ qui sera

utilisée pour exécuter la compression JPEG et l'IDC.

2.5 Méthode d'IDC

L'IDC va être effectuée sur les coefficients DCT qui seront quantifiés par 1 (Tableau 1) [1]. L'IDC se fera par substitution des deux bits de poids faible LSB0 et LSB1. Il sera effectué dans tous les blocs de l'image même ceux contenus dans la RI, ceci est rendu possible grace à la mise à zéro des coefficients non quantifiés Section 2.2.

2.6 Extraction et reconstruction

Cette section traite de l'extraction des données cachées et de la reconstruction de la RI. La reconstruction de la RI se passe en deux étapes :

- La première étape consiste à récupérer les données cachées, pour cela il suffit de lire les bits de poids faible pour les coefficients dont la position dans le bloc correspond à un 1 dans la matrice de quantification (Celle-ci est contenue dans les entêtes du JPEG). Les premières données reçues sont la position et la taille de la RI. Tous les autres bits récupérés sont placés dans un vecteur binaire qui est ensuite décodé à l'aide de la table de Huffman afin de créer un nouveau vecteur de pertes. A la fin de cette opération nous obtenons une image décomprimée (mais ne contenant pas la RI reconstruite), les coordonnées ainsi que la taille de la RI et un vecteur de pertes.

- La deuxième étape est la reconstruction de la RI. A l'aide du vecteur de pertes on reforme des matrices 8×8 représentant les pertes dans chaque bloc de la RI. Les pertes récupérées étant des pertes dans le domaine spatial, il suffit de les ajouter aux blocs décomprimés correspondant de la RI.

3 Résultas expérimentaux

Nous avons appliqué notre méthode de protection sur l'image de Lena en choisissant une RI autour des yeux illustrée Figure 4. La taille de l'image originale est de 512×512 pixels et la taille de la RI est 144×80 pixels (4.39%). La Figure 5 montre l'image comprimée et marquée avec notre méthode pour un facteur de qualité de 10%. Le Tableau 2 compare les résultats de compression et de qualité entre l'algorithme JPEG et notre méthode pour différents facteurs de qualité. Nous observons que le taux de compression est moins significatif dans notre méthode, ceci est dû aux 1 insérés dans la matrice de quantification. Ces 1 changent également le facteur de qualité que nous avons choisi au début. La différence de qualité entre l'image comprimée avec l'algorithme JPEG et celle comprimée et marquée avec notre méthode, Figure 5, est inférieure a 2.2 dB. Ceci prouve que notre méthode d'insertion de données est invisible à l'oeil humain. Le Tableau 3 nous donne le nombre de 1 insérés dans la matrice de quantification ainsi que le nombres de bits utilisés pour effectuer l'IDC. La Figure 6 montre le résultat obtenu après l'extraction et la reconstruction de la RI. La Figure 7 montre les différences entre l'image originale et l'image reconstruite Figure 6, la zone noire prouve que la RI n'a subit aucune perte de données. La qualité de la RI après la reconstruction de l'image est identique à la RI de l'image originale (PSNR infini).

Tab. 2 – Comparaison des résultats obtenus entre JPEG et notre méthode.

FQ	10%	30%	50%
PSNR de l'image compressée JPEG (dB)	29.51	33.03	34.43
Taux de compression du JPEG	30.7	15.7	11
PSNR de l'image compressée avec notre méthode (dB)	27.39	32	32.56
Taux de compression avec notre méthode	6.1	5.8	5.6

Fig. 4 – Image originale avec sélection d'un RI.

Tab. 3 – Résultats obtenus avec notre méthode d'IDC.

Facteur de qualite (FQ)	10%	30%	50%
Nombre de 1 dans MQ	10	9	8
Bits insérés	72844	63295	59988

4 Conclusion

Dans cet article nous avons présenté une méthode permettant de sauvegarder une RI sécurisée sans perte dans une image comprimée avec l'algorithme JPEG. Elle améliore un service qui existe dans JPEG 2000 et mis en application dans l'algorithme JPEG. Elle est basée sur le travail

FIG. 5 – Image comprimée et marquée (FQ = 10).

FIG. 6 – Image avec RI reconstruite.

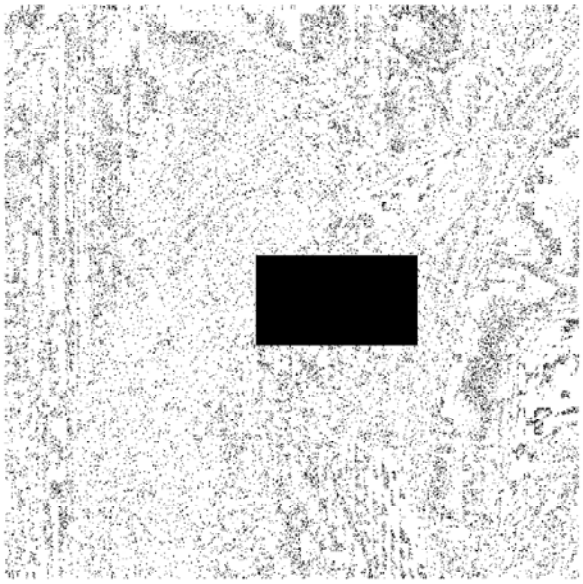


FIG. 7 – Différence entre image originale et image avec RI reconstruite.

de [1]. Nous avons amélioré leur méthode pour y inclure l'information d'une zone déterminée d'une image. Nous étendons actuellement cette méthode à des images couleurs.

Références

[1] C.-C. CHANG and T.-S. L.-Z. CHEN CHUNG, *A Steganographic Method Based Upon JPEG and Quantization Table Modification, Information Sciences, vol. 141, pp. 123-138, 2002.*

[2] C. Christopoulos, J. Askelöf M. Larsson, *Efficient Methods for Encoding Regions of Interest in the Upcoming JPEG2000 Still Image Coding Standard, IEEE Signal Processing Letters, Vol. 7, No. 9, pp. 247-249, September 2000.*

[3] S.B. GOKTURK, C. TOMASI, B. GIROD, and C. BEAULIEU, *Medical image compression based on region of interest, with application to colon CT images, Engineering in Medicine and Biology Society (EMBS), Proceedings of the 23rd Annual International Conference of the IEEE, vol. 3, pp. 2453 -2456, 2001.*

[4] Huffman, *A Method for the Construction of Minimum Redundancy Codes. In Proceedings IRE, vol. 40, pp. 1098-1101, 1962.*

[5] W. JIA, X. HE, and Q. LIN, *Echocardiography Sequencial Images Compression Based on Region of Interest, 2nd International Conference on Information Technology for application, ICITA, Harbin, China, Jan. 2004.*

[6] J. STROM, and P. COSMAN, *Medical Image Compression with Lossless Regions Of Interest, Signal Processing, vol. 59, No.2, pp.155-171, Jun. 1997.*

[7] A. WAKATAMI, *Digital Watermarking for RI Medical Images by Using Compresses Signature Image, In WIAMIS'04, Lisboa, Portugal, 2004.*

Performances d'un Cryptosystème Numérique à base de Chaos

V. GUGLIELMI[1], M. BONNEFONT[1], D. FOURNIER-PRUNARET[1], P. PINEL[1], A.K. TAHA[1]

[1]LESIA, Institut National des Sciences Appliquées, 135 avenue de Rangueil, 31077 Toulouse Cedex

veronique.guglielmi@insa-toulouse.fr

Résumé – Nous présentons les résultats de la cryptanalyse d'un système numérique de cryptographie, système réalisé sur DSP et basé sur l'utilisation de suites chaotiques générées par une récurrence discrète de dimension 2. Nous nous plaçons dans le pire scénario possible quant à la sécurité du chiffrement, ce qui ramène notre attaque à un problème d'estimation des paramètres d'une trajectoire chaotique. La sécurité du cryptosystème est grandement améliorée dès lors que l'on s'interdit la transmission de 2 points successifs de la trajectoire.

Abstract *– We present the results of the cryptanalysis of a digital cryptographic system, implemented on DSP and based on the use of chaotic suites issued from a two-dimensional discrete recurrence. We consider the worst case as regards the cipher security, which reduces our attack to estimating the parameters of a chaotic trajectory. Then, the security of the system is greatly improved if we prohibit the transmission of 2 successive points of the trajectory.*

1. Contexte

L'utilisation de signaux chaotiques pour sécuriser les transmissions est un sujet d'études depuis plusieurs années [1][2][3]. En effet, la grande sensibilité du chaos fait que la même récurrence discrète non linéaire permet de générer, en modifiant légèrement les valeurs de ses conditions initiales ou de ses paramètres, des suites non périodiques de points aux distributions analogues mais qui pourtant ne prendront en fait jamais les mêmes valeurs. Les attracteurs chaotiques peuvent ainsi être assimilés à des générateurs de variables pseudo-aléatoires, chaque suite étant considérée comme la réalisation d'un processus stochastique discret.

Nous avions implémenté sur DSP (Digital Signal Processors) un algorithme de cryptage reposant sur la difficulté à distinguer des trajectoires différentes d'un même attracteur chaotique (construit à partir de récurrences de dimension deux) [4]. Cette réalisation nous avait permis de vérifier le bon fonctionnement de notre cryptosystème et de montrer que toute attaque « à force brute » nécessiterait des temps de calculs prohibitifs [5].

Ici, nous présentons l'analyse cryptographique plus complète que nous avons depuis menée. Nous nous plaçons dans le contexte le pire possible, en ayant au préalable inhibé les niveaux supplémentaires de sécurité que possède le cryptosystème, et en simulant des attaques « à clair choisi ». Le principe de nos attaques dépasse alors le cadre spécifique de notre algorithme de cryptage : il s'agit désormais de retrouver les valeurs des paramètres d'une récurrence chaotique à partir de la seule connaissance de points issus de cette récurrence. Le problème posé est donc maintenant un problème d'estimation non linéaire, qui risque de se retrouver dans tout système de transmission « sécurisée » utilisant des signaux chaotiques en tant que suites pseudo-aléatoires. En effet, l'aspect pseudo-aléatoire des signaux chaotiques repose sur la difficulté à prédire exactement les valeurs prises par ces signaux. Cette difficulté repose entièrement sur la sensibilité du chaos par rapport aux paramètres ; elle n'existe donc plus dès lors qu'il est possible d'estimer les valeurs des paramètres de manière suffisamment précise pour pouvoir reconstruire la trajectoire correspondante de l'attracteur chaotique.

2. Description du cryptosystème

Supposons que le message en clair $\{p_n\}$ soit sous forme binaire, et considérons une récurrence non inversible amenant à des séquences chaotiques $\{s_n\}$. Pour l'implémentation de notre système, les récurrences paraissant pour l'instant les plus appropriées sont des récurrences de dimension égale au minimum à deux [6] :

$$\begin{cases} x_{n+1} = f(x_n, y_n, a, b) \\ y_{n+1} = g(x_n, y_n, a, b) \end{cases}$$

où (x,y) sont les variables réelles, f et g des fonctions non linéaires et a et b des paramètres réels.

Nous prenons pour la séquence chaotique $\{s_n\}$ une seule des deux coordonnées (x_n , y_n), et nous notons pour le point à l'instant n : $s_n = F^n(s_0)$, où s_0 représente la condition initiale de la récurrence. Considérons maintenant deux séquences différentes $\{s_n\} = F^n(s_0)$ et $\{t_n\} = F^n(t_0)$. La n-ième valeur cryptée c_n sera soit s_n soit t_n, selon la valeur du n-ième bit correspondant p_n du message en clair (message supposé sous forme binaire donc) : $c_n = s_n$ si $p_n = 0$; $c_n = t_n$ si $p_n = 1$.

Nous utilisons deux trajectoires du même attracteur chaotique pour les séquences de chiffrement $\{s_n\}$ et $\{t_n\}$, où s_0

et t_0 sont deux conditions initiales différentes choisies dans le bassin d'attraction de l'attracteur. De ce fait, par définition du chaos, les valeurs réelles prises par les suites $\{s_n\}$ et $\{t_n\}$ peuvent être très proches, et distribuées de la même manière. Mais, en même temps, s_n n'est jamais égal à t_n !

Le récepteur déchiffre le message par comparaison des valeurs du texte chiffré c_n avec celles des deux séquences s_n et t_n : $p_n = 0$ si $c_n = s_n$; $p_n = 1$ si $c_n = t_n$. Pour que ce déchiffrement soit possible, il est clair qu'il faut pouvoir reconstruire exactement les deux suites $\{s_n\}$ et $\{t_n\}$ au niveau du récepteur. Nous supposons, afin d'éviter le problème de la synchronisation du chaos entre émetteur et récepteur [7], que les caractéristiques du signal chaotique (paramètres et conditions initiales, en supposant donc l'équation de récurrence connue de tous) ont été échangées au préalable entre les interlocuteurs à titre de clés de communication.

La figure 1 présente le schéma complet du système de cryptage. Notre cryptosystème possède ainsi deux niveaux supplémentaires de sécurité, reposant l'un sur l'utilisation d'une troisième suite pseudo-aléatoire comme bruit additif, l'autre sur l'implémentation d'un masque binaire au niveau du codage en virgule flottante sur le DSP des points issus des suites chaotiques, mais ces deux niveaux sont inhibés pour les attaques présentées ici puisque nous nous plaçons dans le pire contexte possible quant à la sécurité.

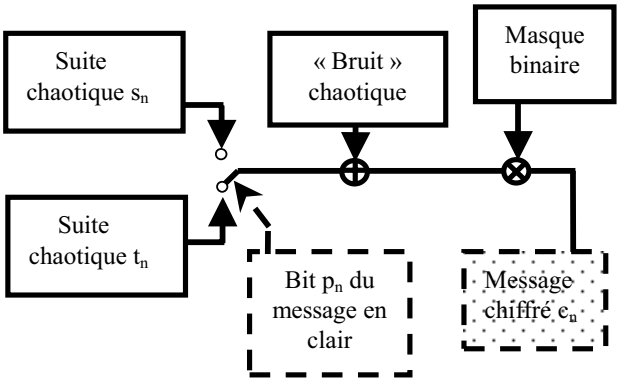


FIG. 1 : Schéma-bloc du codeur

3. Attaques numériques

Dans le cadre de cette communication, nous mènerons notre cryptanalyse en prenant la récurrence cubique suivante :

$$\begin{cases} x_{n+1} = y_n \\ y_{n+1} = a\left(-{x_n}^3 + x_n\right) + b\left(-{y_n}^3 + y_n\right) \end{cases}$$

où (x,y) sont les variables réelles, et a et b des paramètres réels.

Nous avons déjà pu voir par ailleurs [5] qu'il ne s'agit sans doute pas là de la récurrence la meilleure possible quant à la sécurité du système, parce qu'elle est linéaire par rapport aux paramètres a et b, et qu'il s'agit en fait d'une récurrence à une dimension mais de rang 2. Néanmoins, il nous a semblé intéressant de mener nos attaques par rapport à cette récurrence, relativement simple et bien connue, pour en dégager certains résultats sur la sécurité du cryptosystème valables quelle que soit la récurrence utilisée.

Les hypothèses faites pour l'attaque du système [8] sont celles correspondant à une attaque « à clair choisi » : l'attaquant dispose d'une machine de chiffrement, qu'il peut utiliser à sa guise. De ce fait, l'attaquant peut choisir de ne crypter par exemple que des "0", ce qui a pour effet de supprimer le basculement entre les deux suites chaotiques $\{s_n\}$ et $\{t_n\}$: la n-ième valeur cryptée c_n sera toujours égale à s_n pour tout instant n. On suppose de plus que l'attaquant connaît le modèle de récurrence utilisé, mais n'a par contre pas accès à la clé privée que constituent les paramètres a et b et la condition initiale s_0. L'attaque cherche alors à estimer les valeurs de a et de b à partir des points de la suite $\{s_n\}$.

3.1 Des solutions analytiques

La connaissance de 4 points consécutifs d'une même suite, ou même seulement la connaissance de 2 séries de 3 points consécutifs d'une même suite, amène à un système linéaire de 2 équations à 2 inconnues. En effet, supposons par exemple que nous connaissions 4 points successifs s_0, s_1, s_2, s_3 d'une même suite. Cela amène à :

$$\begin{cases} s_2 = b(s_1 - {s_1}^3) + a(s_0 - {s_0}^3) \\ s_3 = b(s_2 - {s_2}^3) + a(s_1 - {s_1}^3) \end{cases}$$

Le problème admet alors une solution analytique :

$$\begin{cases} a = \dfrac{s_3(s_1 - {s_1}^3) - s_2(s_2 - {s_2}^3)}{(s_1 - {s_1}^3)^2 - (s_2 - {s_2}^3)(s_0 - {s_0}^3)} \\ b = \dfrac{s_2(s_1 - {s_1}^3) - s_3(s_0 - {s_0}^3)}{(s_1 - {s_1}^3)^2 - (s_2 - {s_2}^3)(s_0 - {s_0}^3)} \end{cases}$$

Pour remédier à ceci, nous introduisons des décalages dans la récurrence (ce qui correspond à augmenter la dimension de la récurrence), afin d'interdire au pirate la connaissance de points successifs. A partir de la connaissance de 4 points d'indices non consécutifs d'une même suite : $s_0, s_{d1}, s_{d2}, s_{d3}$, où les décalages vérifient: $d3 > d2 > d1 \geq 1$, nous construisons désormais un système polynomial de 3 équations à 3 inconnues :

$$\begin{cases} s_{d1} = Q_1(a,b,s_0) \\ s_{d2} = Q_2(a,b,s_0) \\ s_{d3} = Q_3(a,b,s_0) \end{cases}$$

Les degrés des polynômes Q_1, Q_2, Q_3 varient de manière exponentielle par rapport aux décalages $d1, d2, d3$. Par exemple, pour le polynôme Q_1, pour tout décalage $d \geq 1$, on a :

- Degrés en a et b : $(3^{d-1}-1)/2$;
- Degrés en s_0 : 3^{d-1}

On peut donc obtenir simplement des polynômes de degrés élevés :

- $d=3$ amène aux degrés 4 en a et b et 9 en s_0
- $d=4$ amène aux degrés 13 en a et b et 27 en s_0
- $d=10$ amène aux degrés 9841 en a et b et 19683 en s_0.

Ceci d'une part empêche toute résolution analytique, et d'autre part amène à un système admettant non pas une solution unique mais bien au contraire un grand nombre de solutions parasites. L'existence de ces solutions parasites va compliquer considérablement la recherche des valeurs correctes des paramètres.

3.2 Par la méthode de Newton-Raphson

Avec les décalages introduits au paragraphe précédent, la résolution analytique du problème posé devient impossible, et il faut passer à une méthode numérique d'optimisation. Celle que nous mettons en œuvre est la méthode itérative de Newton-Raphson.

Dans un premier temps, on s'intéresse au cas où les 2 premiers points des 4 points $s_0, s_{d1}, s_{d2}, s_{d3}$ supposés connus sont en fait 2 points d'indices consécutifs (i.e. $d1=0$).

Le système à résoudre n'a alors plus que 2 inconnues au lieu de 3. La méthode de Newton-Raphson en permet la résolution, en amenant aux valeurs correctes de a et de b, et ce même pour des décalages importants (de l'ordre de 15 par exemple) et pour une initialisation quelconque (sans information a priori) de l'algorithme d'optimisation.

Par contre, dès lors que les 2 premiers points des 4 points connus $s_0, s_{d1}, s_{d2}, s_{d3}$ ne sont pas des points successifs (i.e. $d1 \geq 1$), l'attaque n'est que marginalement efficace car l'optimisation, même pour des valeurs initiales relativement proches des valeurs cherchées et pour des décalages modestes, n'aboutit pas souvent. Elle converge vers une solution parasite du système, ou bien elle fait intervenir une matrice singulière, ou encore elle diverge.

4. Attaques dans l'espace d'état

4.1 Estimation en une dimension

Le comportement statistique des suites chaotiques est lié aux paramètres. En estimant la densité de probabilité des points issus d'une suite, on montre que cette densité dépend essentiellement des valeurs de a et de b, et qu'elle ne dépend que faiblement des valeurs des conditions initiales de la récurrence ou des valeurs des décalages entre les points de la suite. Par exemple, pour a=2.2 et b=-0.91, un attracteur chaotique existe et peut être observé dans le plan de phase de la figure 2. Sa densité de probabilité est donnée par la figure 3.

La comparaison de la densité obtenue pour le message chiffré avec les densités de la base de donnée que le pirate aura pu constituer (pour différentes valeurs des paramètres) amène à une première estimation de a et b, relativement imprécise mais utile pour initialiser la méthode de Newton-Raphson.

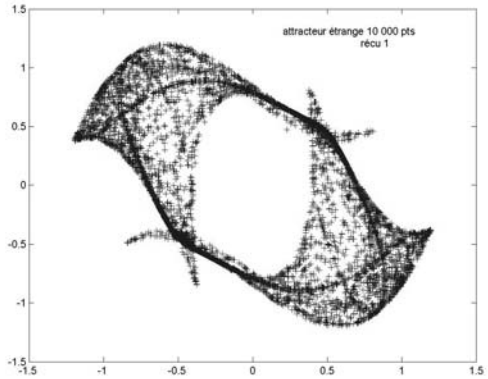


FIG. 2 : Attracteur étrange

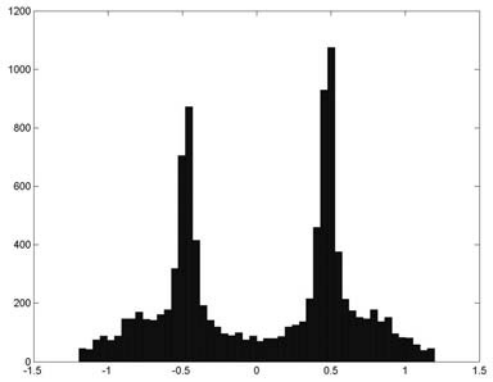


FIG. 3 : Densité de probabilité

4.2 Estimation en dimension supérieure

De manière similaire, le tracé de l'attracteur chaotique dans l'espace d'état (i.e. le tracé du nuage des points (s_n, s_{n+d}) si on note d le décalage utilisé lors du cryptage) peut lui aussi permettre une première estimation des paramètres a et b, et ce afin d'initialiser l'algorithme de Newton-Raphson. En effet, la forme de l'attracteur dans l'espace d'état ne dépend pas des conditions initiales de la suite $\{s_n\}$; elle est caractéristique de zones dans l'espace des paramètres. De plus, ces tracés donnent des informations sur la valeur du décalage d, ainsi que sur les positions où surviennent les changements de valeur binaire (de "0" à "1" ou de "1" à "0") dans la suite des bits du message en clair.

Dans le cas où le message en clair n'est pas choisi par le pirate mais correspond à un message réellement transmis, il est intéressant d'utiliser cette possibilité de détection des changements de bits pour essayer de décrypter directement le message. Par exemple, la figure 4 donne le tracé des points (c_n, c_{n+1}) pour un message chiffré avec un décalage constant égal à 2.

Toutefois, cette détection est trop imprécise pour permettre à elle seule un déchiffrement correct. Elle doit donc plutôt être utilisée en support à un algorithme d'estimation des paramètres *a* et *b*, quand l'attaque n'est pas « à clair choisi ».

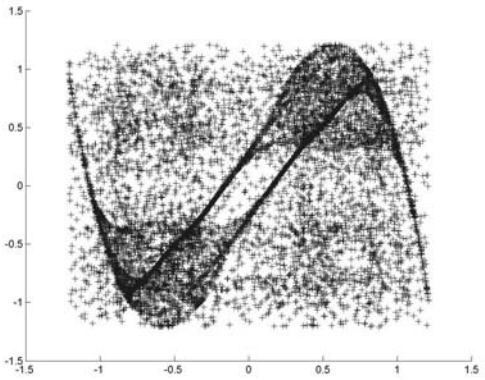


FIG. 4 : Points (c_n, c_{n+1}) d'un message crypté

5. Conclusions

Pour tester la sécurité d'un cryptosystème numérique réalisé dans notre laboratoire, nous effectuons des attaques « à clair choisi ». L'attaque s'interprète alors comme un problème d'estimation non linéaire des paramètres d'une récurrence chaotique, et dépasse ainsi le cadre de la cryptanalyse pour s'attacher à l'aspect pseudo-aléatoire du chaos. L'estimation des paramètres de la récurrence pouvant se ramener à la résolution de systèmes polynomiaux, nous développons un algorithme d'attaque basé sur une méthode numérique d'optimisation (la méthode de Newton-Raphson) et initialisé via l'étude des attracteurs chaotiques dans l'espace d'état. Un résultat principal apparaît : l'attaque réussit dès lors que l'on connaît 2 points d'indices consécutifs de la récurrence ; elle échoue la plupart du temps sinon. Ceci prouve la nécessité d'introduire des décalages dans la suite des points, et d'augmenter ainsi fortement la dimension de la récurrence utilisée. Ces décalages peuvent être constants ou mieux, varier dynamiquement en fonction du temps et des variables du cryptosystème, ce qui améliore la sécurité du chiffrement.

Remerciements

Ce travail a été supporté en partie par le Fonds National de la Science de la République Française (projet TRANSCHAOS de l'ACI 'Sécurité Informatique' 2003).

Références

[1] L. Kocarev, *Chaos-based cryptography: a brief overview*, IEEE Circuits and Systems Magazine, 2001, 1, (3), pp. 6-21.

[2] T. Stojanovski et L. Kocarev, *Chaos-based random number generators-part I: analysis [cryptography]*, IEEE Trans. on Circuits and Systems I, 2001, 48, (3), pp. 281-288.

[3] T. Stojanovski et L. Kocarev, *Chaos-based random number generators-part II: practical realization*, IEEE Trans. on Circuits and Systems I, 2001, 48, (3), pp. 382-385.

[4] L. Bénéteau, D. Fournier-Prunaret, V. Guglielmi, P. Pinel, S. Rouabhi et A.K. Taha, *Two encryption schemes using the chaotic dynamics of two-dimensional noninvertible maps*, Complements to Proceedings of International Conference on Nonlinear Dynamics of Electronic Systems (NDES'02), Juin 2002, Izmir, Turquie.

[5] V. Guglielmi, H. Poonith, D. Fournier-Prunaret, A.K. Taha, *Security performances of a chaotic cryptosystem*, IEEE International Symposium on Industrial Electronics (ISIE '04), Mai 2004, Ajaccio, France.

[6] C. Mira, D. Fournier-Prunaret, L. Gardini, H. Kawakami et J.C. Cathala, *Basin bifurcations of two-dimensional non invertible maps – Fractalization of basins*, International Journal of Bifurcation and Chaos, 1994, 4, (2), pp. 343-382.

[7] Y.-H. Chu et S. Chang, *Dynamical cryptography based on synchronised chaotic systems*, Electronics Letters, 1999, 35, (12), pp. 974-975.

[8] M.I. Sobhy et A.R. Shehata, *Methods of attacking chaotic encryption and countermeasures*, Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP'01), Mai 2001, Salt Lake City, USA.

Séparation de sources non-négatives par l'approche du maximum de vraisemblance

Saïd MOUSSAOUI[1], David BRIE[1], Cédric CARTERET[2]

[1]Centre de Recherche en Automatique de Nancy (CRAN), UMR 7039 CNRS-UHP-INPL
Université Henri Poincaré, Nancy 1, B.P. 239, Faculté des Sciences et Techniques, 54506 Vandœuvre-lès-Nancy, France

[2]Laboratoire de Chimie-Physique et Microbiologie pour l'Environnement (LCPME), UMR 7564 UHP-CNRS
Université Henri Poincaré, Nancy 1, 405, rue de Vandoeuvre, 54600, Villers-lès-Nancy, France

{said.moussaoui,david.brie}@cran.uhp-nancy.fr,carteret@lcpe.cnrs-nancy.fr

Résumé – Ce papier traite du problème de la séparation de sources non-négatives en utilisant l'approche du maximum de vraisemblance. Il est montré que cette approche peut être appliquée avec succès si on utilise une fonction non-linéaire dérivée de la modélisation des distributions des signaux sources par des densités à support positif. Dans le cas des signaux de spectroscopie, qui est notre application principale, une bonne loi candidate est la distribution gamma qui permet de prendre en compte conjointement la non-négativité et la parcimonie des signaux sources. Un exemple synthétique est présenté afin d'étudier les performances de l'algorithme résultant.

Abstract – This papers addresses the problem of non-negative source separation using the maximum likelihood approach. It is shown that this approach can be effective by considering that the sources are distributed according to a density having a non-negative support from which an adequate non-linear separating function can be derived. In the particular of spectroscopic data which is our main concern, a good candidate is the Gamma distribution which allows to encode both the non-negativity and the sparsity of the source signals. A synthetic example is used to assess the performances of the resulting algorithm.

1 Introduction

Ce travail est motivé par le besoin d'appliquer les méthodes de séparation de sources au traitement des signaux résultants de l'analyse de substances chimiques multi-composantes par spectroscopie. Les objectifs du traitement sont l'identification de la composition chimique de la substance et l'évaluation de la proportion de chaque composante. D'après la loi de Beer-Lambert, les signaux mesurés sont une combinaison linéaire des spectres des composantes pures inconnues. C'est donc un problème de séparation de sources où les signaux sources correspondent aux spectres pures et les concentrations de chaque espèces sont déduites à partir des coefficients de mélange. La contrainte principale de cette application est que les coefficients de mélange et les signaux sources sont non-négatifs.

Le modèle d'observation suppose que m signaux mesurés, notés $\boldsymbol{x}_t = [x_1(t), \ldots, x_m(t)]^T$, sont un mélange linéaire de p signaux sources non-mesurables $\boldsymbol{s}_t = [s_1(t), \ldots, s_p(t)]^T$

$$\boldsymbol{x}_t = \boldsymbol{A}\,\boldsymbol{s}_t + \boldsymbol{e}_t, \tag{1}$$

où $\boldsymbol{A}$ est la matrice de mélange et de dimension $(m \times p)$. Chaque colonne de cette matrice contient le profil d'évolution de la contribution de chaque source dans les mélanges observés. Le vecteur $\boldsymbol{e}_t$ représente les erreurs de mesure et les incertitudes du modèle d'observation. Ayant n échantillons, $(t = 1, \ldots, n)$, et en utilisant une notation matricielle, le modèle de mélange est exprimé par

$$\boldsymbol{X} = \boldsymbol{A}\,\boldsymbol{S} + \boldsymbol{E}, \tag{2}$$

où les matrices $\boldsymbol{X} \in \mathbb{R}^{m\times n}, \boldsymbol{S} \in \mathbb{R}_+^{m\times n}$ et $\boldsymbol{E} \in \mathbb{R}^{m\times n}$ contiennent, respectivement, les signaux mesurés, les signaux sources et les séquences de bruits sur chaque ligne. En supposant que le nombre de sources est connu *a priori*, le problème de la séparation de sources non-négatives consiste à estimer les signaux sources et les coefficients de mélanges qui reproduisent aux mieux les données tout en satisfaisant les contraintes de non-négativité.

Afin de résoudre ce problème, plusieurs approches sont possibles. La première est fondée sur une estimation par moindres carrés sous contrainte de non-négativité [1, 2]. Néanmoins, cette approche ne fournit une solution unique que sous certaines conditions restrictives [3, 4]. La sélection d'une solution particulière parmi toutes les solutions admissibles se fait alors par le rajout d'hypothèses supplémentaires. Nous utilisons une approche du maximum de vraisemblance pour tenir compte de l'hypothèse d'indépendance statistique des signaux sources. Sous cette hypothèse, c'est donc un problème d'analyse en composantes indépendantes non-négatives pour lequel une méthode existante permet de faire la séparation dans le cas de sources *well grounded* et orthogonales [5]. En revanche, comme cela est montré dans [6], l'utilisation des méthodes ICA qui imposent l'orthogonalité des sources, tels que JADE [7] ou SOBI [8], peut mener à des estimations négatives lorsque les échantillons disponibles des signaux sources présentent une corrélation spatiale significative. L'approche du maximum de vraisemblance n'impose pas cette contrainte d'orthogonalité mais requiert la spécification d'une fonction non-linéaire afin d'obtenir les sources les plus indépendantes possible et dont la distribution est proche de la densité correspondant à cette fonction non-linéaire. Nous utilisons des distributions gamma qui sont de bons modèles pour tenir compte simultanément de la non-négativité et de la parcimonie des signaux sources [9].

2 Séparation de sources non-négatives par maximum de vraisemblance

2.1 Formulation du problème

Sous l'hypothèse de signaux sources mutuellement statistiquement indépendants et temporellement i.i.d et en tenant compte du modèle de mélange $\boldsymbol{x}_t = \boldsymbol{A}\,\boldsymbol{s}_t$, dans le cas carré ($m = p$) et non bruité ($\boldsymbol{e}_t = 0$) et en introduisant la matrice séparante, notée $\boldsymbol{B}$, l'estimée de cette matrice par maximum de vraisemblance s'obtient par [10, 11]

$$\hat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \Big\{ -\log|\det(\boldsymbol{B})| - \frac{1}{n}\sum_{t=1}^{n}\sum_{j=1}^{p} \log p_{s_j}\big([\boldsymbol{B}\,\boldsymbol{x}_t]_j\big) \Big\}. \quad (3)$$

L'optimisation de ce critère en utilisant la méthode du gradient relatif/naturel [12] conduit à une mise à jour itérative selon

$$\hat{\boldsymbol{B}}^{(r+1)} = \hat{\boldsymbol{B}}^{(r)} - \mu^{(r)} \left[\frac{1}{n}\sum_{t=1}^{n} \boldsymbol{\psi}\left(\hat{\boldsymbol{y}}_t^{(r)}\right)\left(\hat{\boldsymbol{y}}_t^{(r)}\right)^T - \boldsymbol{I} \right] \hat{\boldsymbol{B}}^{(r)}, \quad (4)$$

où $\hat{\boldsymbol{y}}_t^{(r)} = \hat{\boldsymbol{B}}^{(r)}\boldsymbol{x}_t$ et $\boldsymbol{\psi}^T(\boldsymbol{z}) = [\psi_1(z_1), ..., \psi_n(z_n)]$ et les fonctions de séparation $\psi_j(z)$ sont définies par $\psi_j(z) = -\dfrac{\partial}{\partial z}\log p_{z_j}(z)$. Pour la cas où $m > p$, une réduction de dimension de la matrice des données par transformation aléatoire ($p \times m$) ou par analyse en composantes principales est nécessaire avant l'application de cet algorithme.

Idéalement, les fonctions de séparation se déduisent des densités de probabilité des sources recherchées [11, 13], qui dans la plupart des cas ne sont pas connues *a priori*. Il faudra alors estimer ces densités en même temps que la matrice de séparation. Pour cela, la distribution des sources est approchée soit en utilisant les statistiques d'ordre supérieur [10, 14] ou par un modèle paramétrique [11, 15]. Le choix d'un modèle de distribution permet de spécifier des informations préalables sur la structure de la densité de probabilité des sources. Quelques exemples de modèles sont la gaussienne généralisée, la distribution de Student et le mélange de gaussiennes. Par exemple, les deux premiers modèles correspondent à des distributions unimodales symétriques mais ne permettent pas de coder explicitement la contrainte de non-négativité des signaux sources. Pour le cas de sources asymétriques ou multi-modales l'utilisation d'un modèle de mélange de gaussiennes permet de s'affranchir de cette limitation, mais le cas de sources non-négatives reste toujours problématique.

2.2 Synthèse d'une fonction non-linéaire

Afin de prendre en compte explicitement la contrainte de non-négativité des sources, une modélisation de leurs distributions par des densités à support non-négatif est nécessaire. Dans le cadre de l'analyse des signaux de spectroscopie, un modèle possible est la distribution gamma dont la densité de probabilité s'exprime selon

$$p(s|\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\, s^{\alpha-1}\, e^{-\beta s}\, \mathbb{I}_{s\geq 0}, \quad (5)$$

où α et β sont les paramètres de forme et d'échelle. Le choix de cette densité est motivé par deux avantages ; celle-ci prend en compte explicitement la non-négativité car $p(s<0)=0$, et les deux paramètres α et β de cette densité peuvent être utilisés pour ajuster sa forme asymétrique. La fonction non-linéaire correspondante à la loi gamma est exprimée par

$$\psi(s;\alpha,\beta) = \beta - (\alpha-1)\,s^{-1}, \quad s>0. \quad (6)$$

Cette densité de probabilité est à support non-négatif, donc pour des estimées intermédiaires négatives, durant l'optimisation, il faudra redéfinir la fonction non-linéaire selon

$$\psi(s;\alpha,\beta) = \beta - \frac{(\alpha-1)}{\max(s,\epsilon)}, \quad \forall\, s\in\mathbb{R} \quad (7)$$

où ϵ est fixé à une petite valeur (de l'ordre de 10^{-3}). Un exemple de distribution gamma et de fonction non-linéaire qui en résulte sont illustrés par la figure 1.

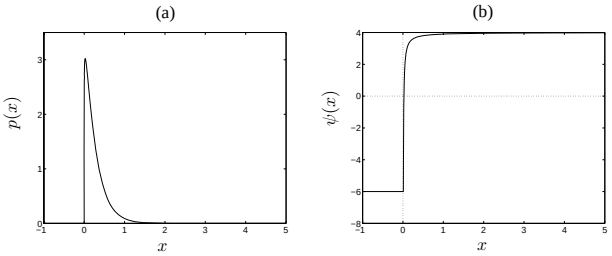


FIG. 1 – (a) Distribution gamma pour $\alpha = 1.1$ et $\beta = 4$ et (b) fonction non-linéaire associée pour $\epsilon = 0.01$.

2.3 Stabilité de l'algorithme

Des conditions suffisantes de stabilité de l'algorithme du gradient relatif/naturel sont exprimées par [16, 17]

$$\mathrm{E}[\psi'(\tilde{s}_j)\tilde{s}_j^2] - \mathrm{E}[\psi(\tilde{s}_j)\tilde{s}_j] > 0 \quad \forall\, j = 1,\ldots,p \quad (8)$$

où $\mathrm{E}[f(\tilde{s}_j)]$ sont les moments non-linéaires des sources centrées

$$\tilde{s}_j = s_j - \mathrm{E}[s_j] = s_j - \alpha_j/\beta_j. \quad (9)$$

Le calcul de ces moments donne

$$\mathrm{E}[\psi'(\tilde{s}_j)\tilde{s}_j^2] = \mathrm{E}\left[\frac{(\alpha_j-1)}{\tilde{s}_j^2}\cdot\tilde{s}_j^2\right] = (\alpha_j - 1); \quad (10)$$

et

$$\mathrm{E}[\psi(\tilde{s}_j)\tilde{s}_j] = \mathrm{E}\left[\beta_j\tilde{s}_j - (\alpha_j-1)\right] = (1-\alpha_j). \quad (11)$$

Par conséquent, une condition suffisante pour la stabilité de l'algorithme est

$$\alpha_j > 1, \quad j = 1,\ldots,p. \quad (12)$$

A partir de cette condition de stabilité, il apparaît que seulement des densités gamma ayant un paramètre de forme $\alpha > 1$ peuvent être utilisées, ce qui est un inconvénient de l'algorithme d'optimisation choisi. Nous revenons à ce point dans le paragraphe 3.2.

2.4 Estimation des paramètres des sources

Les paramètres des densités gamma $\boldsymbol{\theta} = \{\alpha_j, \beta_j\}_{j=1}^{p}$ sont estimés conjointement avec la matrice de séparation par la maximisation de la fonction de vraisemblance conjointe

$$\left(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\theta}}\right) = \arg\max_{\boldsymbol{B},\boldsymbol{\theta}} p(\boldsymbol{X}|\boldsymbol{B}, \boldsymbol{\theta}), \tag{13}$$

La minimisation du critère qui en résulte est réalisée par un algorithme de gradient alterné mettant à jour, à chaque itération, la matrice de séparation en utilisant la dernière valeur du vecteur de paramètres et en utilisant cette nouvelle estimation de la matrice de séparation pour mettre à jour ces paramètres. La mise à jour de la matrice de séparation étant réalisée selon l'équation (4), les paramètres des densités gamma sont mis à jour selon

$$\begin{cases} \alpha_j^{(r+1)} = \alpha_j^{(r)} - \rho_\alpha \nabla_\alpha \mathcal{J}(\boldsymbol{s}|\alpha_j^{(r)}, \beta_j^{(r)}), \\ \beta_j^{(r+1)} = \beta_j^{(r)} - \rho_\beta \nabla_\beta \mathcal{J}(\boldsymbol{s}|\alpha_j^{(r+1)}, \beta_j^{(r)}), \end{cases} \tag{14}$$

où $(\rho_\alpha, \rho_\beta)$ sont des pas positifs et

$$\begin{cases} \nabla_{\alpha_j} \mathcal{J}(\boldsymbol{s}|\alpha_j, \beta_j) = -\log \beta_j + \Psi(\alpha_j) - \dfrac{1}{n}\displaystyle\sum_{t=1}^{n} \log s_j(t) \\ \nabla_{\alpha_j} \mathcal{J}(\boldsymbol{s}|\alpha_j, \beta_j) = -\dfrac{\alpha_j}{\beta_j} + \dfrac{1}{n}\displaystyle\sum_{t=1}^{n} s_j(t), \end{cases}$$

où $\Psi(\alpha_j) = \dfrac{d}{d\alpha}\Gamma(\alpha_j)$ est la fonction digamma [18]. En tenant compte des conditions de stabilité (12) et de la définition de la densité gamma, les valeurs de α_j et β_j sont contraintes à $\alpha_j \geq (1+\epsilon)$ et $\beta_j \geq \epsilon$ où ϵ est fixé à une faible valeur (10^{-3}).

3 Exemple numérique

Les signaux sources et les coefficients de mélange sont simulés de sorte à ce qu'ils soient similaires aux données résultantes de l'analyse de substances chimiques multi-composantes par des techniques de spectroscopie. La figure 2 montre un exemple de $p = 3$ signaux sources normalisés (variance unité) de $n = 1000$ échantillons obtenus à partir d'un modèle simplifié dans lequel chaque signal résulte de la superposition de $K = 15$ motifs lorentziens et gaussiens dont les amplitudes et les paramètres de position et de dispersion sont choisis de façon pseudo-aléatoire. Les profils des coefficients de mélange sont simulés pour $m = 3$ observations.

Afin de mesurer la qualité de la reconstruction des signaux sources, nous utilisons l'erreur quadratique moyenne normalisée, notée $\mathcal{E}_{s_j}$ et définie par

$$\mathcal{E}_{s_j} = \sum_{t=1}^{n} \left(s_j(t) - \hat{s}_j(t)\right)^2 / \sum_{t=1}^{n} \left(s_j(t)\right)^2, \tag{15}$$

où $\hat{s}_j$ est l'estimée de la j-ème source s_j et les deux signaux sont réduits à une variance unité. Pour comparer les performances globales, nous utilisons l'indice

$$\mathcal{I}_{global} = \frac{1}{2p\,(p-1)} \sum_{i=1}^{p} \left\{ \left(\sum_{k=1}^{p} \frac{|g_{ik}|^2}{\max\limits_{\ell} |g_{i\ell}|^2} - 1 \right) + \left(\sum_{k=1}^{p} \frac{|g_{ki}|^2}{\max\limits_{\ell} |g_{\ell i}|^2} - 1 \right) \right\}, \tag{16}$$

où g_{ij} est (i,j)-ème élément de la matrice $\boldsymbol{G} = \hat{\boldsymbol{B}}\boldsymbol{A}$.

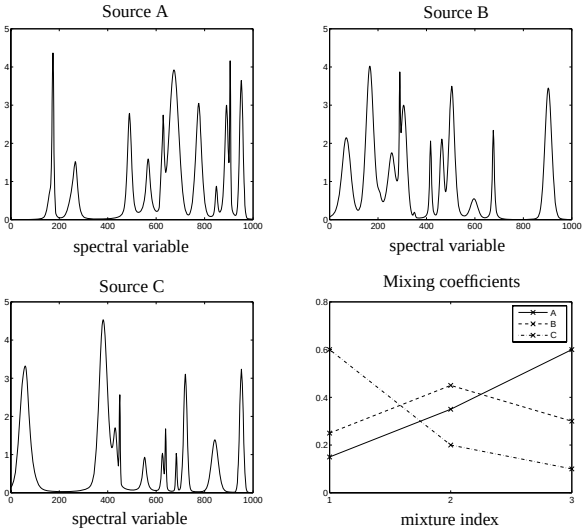


FIG. 2 – Signaux sources et coefficients de mélange simulés.

3.1 Cas sans bruit

La matrice de covariance empirique

$$\hat{\boldsymbol{R}}_s = \begin{bmatrix} 1.00 & 0.13 & -0.18 \\ 0.13 & 1.00 & -0.22 \\ -0.18 & -0.22 & 1.00 \end{bmatrix}, \tag{17}$$

n'est pas diagonale, ce qui montre que les échantillons disponibles des signaux sources présentent une corrélation spatiale non-négligeable. Ainsi, l'application d'une méthode qui force l'orthogonalité (SOBI ou JADE) va produire une solution incorrecte, de même la méthode NNICA fournit des estimations négatives, même si celle-ci considère la non-négativité des signaux sources. D'autre part, l'application d'une méthode fondée uniquement sur la contrainte de non-négativité (NMF ou ALS) ne fournit pas une solution unique car les conditions nécessaires pour l'unicité de la solution sous la contrainte de non-négativité uniquement ne sont pas satisfaites [4]. Par ailleurs, notons la supériorité de l'approche du maximum de vraisemblance en termes d'indice de séparation global ou de qualité de reconstruction des sources.

TAB. 1 – Comparaison des performances des différentes approches. MLPSS correspond à l'approche proposée et mes indices de performances sont exprimés en décibels (dB).

	JADE	NMF	NNICA	MLPSS
$\mathcal{E}_A$	-10.58	-16.20	-15.11	-35.88
$\mathcal{E}_B$	-10.96	-15.88	-14.19	-42.25
$\mathcal{E}_C$	-17.40	-14.14	-39.41	-40.30
$\mathcal{I}_{global}$	-12.06	-15.67	-15.90	-38.32

3.2 Cas avec bruit

Afin de discuter les performances de la méthode, un bruit gaussien est ajouté aux observations. Le niveau de bruit est exprimé en termes de rapport signal sur bruit (RSB) défini comme étant le rapport des variances des signaux non-bruités et des séquences de bruit rajoutées. La figure 3 montre l'évolution des performances des méthodes NNICA, MLPSS et BPSS [9] en fonction du niveau de bruit. Pour un RSB très fort, l'utili-

sation de la distribution gamma et l'estimation soit par maximum de vraisemblance ou par approche bayésienne donne de meilleures performances que NNICA. En dépit du fait que BPSS et MLPSS utilisent le même modèle *a priori* sur la distribution des sources, les performances de BPSS sont supérieures, ce qui s'explique par la restriction du paramètre de forme de la densité gamma à une valeur supérieure à 1, afin de garantir la stabilité de l'algorithme du gradient, et de la non-prise en compte de la positivité des coefficients de mélange. En augmentant le niveau de bruit, les performances de MLPSS diminuent puisque cette approche ne prend pas en compte la présence du bruit sur les observations. Afin de palier à ces limitations, une alternative est d'utiliser les algorithmes EM (Expectation-Maximization) afin d'estimer conjointement les coefficients de mélange et la covariance du bruit d'observation et d'envisager des distributions gamma avec des paramètres de forme inférieures à 1.

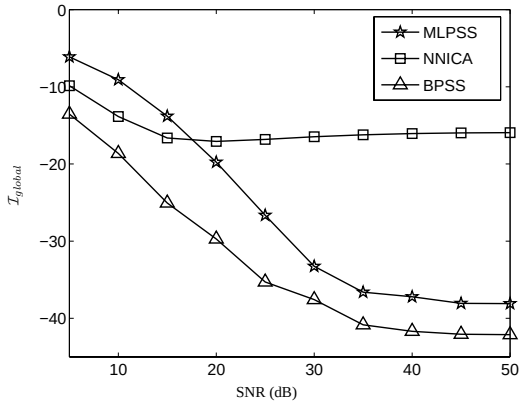


FIG. 3 – Influence du niveau de bruit sur les performances de NNICA, MLPSS et BPSS.

4 Conclusion

L'objectif de ce papier est de montrer que l'approche du maximum de vraisemblance peut être appliquée à la séparation de sources non-négatives à condition de choisir des fonctions non-linéaires adéquates. La non-négativité des sources est prise en compte via la synthèse d'une fonction non-linéaire déduite de la modélisation des sources par des densités à support non-négatif. Dans le cas de données spectrales, la non-négativité et la parcimonie des signaux sources sont incorporées dans la méthode de séparation en utilisant des distributions gamma pour la synthèse de fonctions non-linéaires appropriées. L'analyse des performances dans le cas de données bruitées a montré la nécessité de considérer la présence de bruit dans le modèle d'observation et de prendre en compte conjointement la non-négativité des signaux sources et des coefficients de mélange.

Remerciements

Ce travail est financé par la "Région Lorraine" et le CNRS.

Références

[1] R. Bro and S. De Jong, "A fast non-negativity constrained least squares algorithm," *Journal of Chemometrics*, vol. 11, pp. 393–401, 1997.

[2] D. Lee and H. Seung, "Learning the parts of objects by non–negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[3] D. Donoho and V. Stodden, "When does non-negative matrix factorization give a correct decomposition into parts ?" in *proceedings of NIPS*, 2003.

[4] S. Moussaoui, D. Brie, and J. Idier, "Non-negative source separation : Range of admissible solutions and conditions for the uniqueness of the solution," in *proceedings of ICASSP*, 2005, pp. 289–292.

[5] M. D. Plumbley, "Algorithms for non–negative independent component analysis," *IEEE Transactions on Neural Networks*, vol. 14, no. 3, pp. 534–543, 2003.

[6] D. Nuzillard and J. Nuzillard, "BSS applied to non-orthogonal signals," in *proceedings of ICA*, 1999.

[7] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non Gaussian signals," *IEE Proceedings-F*, vol. 140, no. 6, pp. 362–370, 1993.

[8] A. Belouchrani and K. Abed-Meraim, "Séparation aveugle au second ordre de sources corrélées," in *actes du Colloque GRETSI sur le Traitement du Signal et des Images (GRETSI'93)*, Juan-les-Pins, France, 1993.

[9] S. Moussaoui, D. Brie, A. Mohammad-Djafari, and C. Carteret, "Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling," à paraître dans IEEE Transactions on Signal Processing.

[10] M. Gaeta and J. Lacoume, "Estimateurs du maximum de vraisemblance étendus à la séparation de sources non-gaussiennes," *Traitement du signal*, vol. 7, no. 5, pp. 419–434, 1990.

[11] D.-T. Pham, P. Garrat, and C. Jutten, "Separation of a mixture of independent sources through a maximum likelihood approach," in *proceedings of EUSIPCO*, 1992.

[12] J.-F. Cardoso and B. H. Laheld, "Equivariant adaptive source separation," *IEEE Transactions on Signal Processing*, vol. 44, no. 12, pp. 3017–3030, 1996.

[13] J.-F. Cardoso, "Infomax and maximum likelihood for source separation," *IEEE Letters on Signal Processing*, vol. 4, pp. 112–114, 1997.

[14] P. Comon, "Independent component analysis – a new concept ?" *Signal Processing*, vol. 36, pp. 287–314, 1994.

[15] B. Pearlmutter and L. Parra, "A context-sensitive generalization of ICA," in *proceedings of NIPS*, 1996.

[16] S.-I. Amari, T.-P. Chen, and A. Cichocki, "Stability analysis of adaptive blind source separation," *Neural Networks*, vol. 10, no. 8, pp. 1345–1351, 1997.

[17] J.-F. Cardoso, "On the stability of source separation algorithms," *Journal of VLSI Signal Processing*, vol. 26, pp. 7–14, 2000.

[18] M. Abrahamovitz and I. Stegun, *Handbook of Mathematical Functions*. New York : Dover Publications, 1972.

Séparation de mélanges convolutifs de modulations CPM sur-échantillonnées

Pierre JALLON[1], Antoine CHEVREUIL[1], Philippe LOUBATON[1], Pascal CHEVALIER[2]

[1]Equipe signal et communications
Institut Gaspard Monge
Université de Marne la Vallée, 5, bd Descartes, Champs sur Marne, 77545 Marne La Vallée Cedex 2, France

[2]Thalès communications
EDS/SPM/SBP, 160, bd Valmy, 92700 Colombes, France

pierre.jallon@univ-mlv.fr, antoine.chevreuil@univ-mlv.fr
philippe.loubaton@univ-mlv.fr, pascal.chevalier@fr.thalesgroup.com

Résumé – Nous nous intéressons à la séparation de sources CPM mélangées par des canaux sélectifs en fréquences. Les débits des signaux transmis sont quelconques. En général, la connaissance des fréquences cycliques (inconnues du récepteur dans un contexte aveugle, et difficiles à estimer en pratique) n'est pas nécessaire à la mise en oeuvre d'algorithmes basés sur une approche par déflation; en effet, nous montrons que la minimisation du critère de Godard permet d'extraire une version filtrée d'une des sources.

Abstract – The separation of a convolutive mixture of CPM signals is addressed. The baud-rates of the sources are possibly different. In general, the knowledge of these baud-rates (unknown from the receiver in a blind context and very difficult to estimate in practice) is not required to process a deflation-based separation algorithm; indeed, we show that the minimization of the Godard criterion allows one to extract a filtered version of a source.

1 Introduction

Les problèmes de séparation de sources ont été l'objet de nombreux travaux scientifiques ces dernières années. Si ce thème de recherche a suscité autant d'intérêt, c'est principalement grâce à la diversité de ses applications: entre autres, applications audio, médicales ou encore améliorations des performances et détection de paramètres de systèmes de communications numériques. C'est ce dernier type d'application qui nous intéresse dans cette contribution:

Nous nous plaçons dans un contexte d'écoute passive où K émetteurs mal connus transmettent des signaux dans une même bande de fréquence. Ces signaux sont mélangés convolutivement par des canaux de propagation sélectifs en fréquence, et les signaux résultants sont reçus par un réseau de N capteurs ($N \geq K$) et échantillonnés à une cadence compatible avec le théorème de Shannon. On souhaite extraire des signaux disponibles le plus d'informations possibles sur les K émetteurs (débit, modulation utilisé, constellation des symboles transmis...). Pour cela, il est raisonnable dans un première étape de séparer du signal reçu les différents signaux transmis, et dans une seconde étape d'utiliser des algorithmes d'estimation et de classification adaptés au contexte mono-source. Dans cet article, nous nous intéressons uniquement à l'étape de séparation du mélange, et nous faisons l'hypothèse que les sources sont mutuellement indépendantes et résultent de modulations à phase continue (CPM) de débits éventuellement différents. L'un des points originaux de notre travail provient du fait que la fréquence d'échantillonnage des signaux reçus n'a a priori aucun lien avec les débits symboles des signaux sources; dans ces conditions, le mélange qu'il convient de séparer est constitué de signaux cyclostationnaires dont les fréquences cycliques sont inconnues.

Dans cet article, nous nous intéressons à l'utilisation d'une approche séquentielle par déflation, consistant à rechercher un filtre N entrées / 1 sortie, de fonction de transfert $\mathbf{g}(z)$, qui excité par le signal de dimension N reçu, produit en sortie un signal scalaire $r(n)$ coincidant avec une version filtrée d'une des sources. Pour ceci, il convient de mettre en évidence des fonctions, dites de contraste, définies à partir des statistiques de $r(n)$, dont les extremas globaux sont atteints si et seulement si $r(n)$ est une version filtrée de l'une des sources. Il est alors possible de retrancher du mélange observé la contribution de cette source et, de manière séquentielle, d'extraire une version filtrée de chacune des sources : voir [1, 5, 6, 7, 4].

Ainsi que cela est expliqué dans [2], il est relativement aisé de mettre en évidence des fonctions de contraste adaptées au contexte cyclostationnaire en introduisant des moyennages temporels dans les expressions des contrastes utilisés dans le cas stationnaire. A titre d'exemple, il est établi dans [2] que le maximum de la fonction $\frac{|<c_4(r(n))>|}{(<\mathbb{E}|r(n)|^2>)^2}$ ($< . >$ est l'opérateur de moyenne temporelle) est atteint si et seulement si $r(n)$ est une filtrée d'une des sources. Cependant, l'estimation consistante de cette fonction à partir d'un nombre fini d'observations nécessite, dans notre contexte aveugle, l'estimation des fréquences cycliques des signaux reçus. Comme ceci n'est pas toujours un problème simple, il est pertinent de tenter de mettre en évidence des fonctions de contraste dont l'estimation consistante ne nécessite pas la connaissance des fréquences cycliques du signal reçu. Dans [3], nous avons considéré le cas de mélanges de signaux modulés linéairement par des symboles indépendants identiquement distribués (i.i.d.), et avons étudié les propriétés

séparantes de la fonction de Godard définie par

$$J(r(n)) =< \mathbb{E}(|r(n)|^2 - 1)^2 >, \quad (1)$$

Le but de cet article est d'étudier dans quelle mesure les résultats de [3] s'étendent au cas de mélanges de signaux CPM. Dans la section 2, nous précisons la nature des signaux considérés. La section 3 est consacrée à la présentation de nos résultats. Enfin, la section 4 est consacrée à quelques résultats de simulation.

2 Modélisation du problème

Les K signaux sources sont supposés être des signaux modulés à phase continue: pour tout k, le k-ième signal s'écrit $\exp(i\phi_{a,k}(t))$ où la phase[1] $\phi_{a,k}(t)$ est donnée par:

$$\phi_{a,k}(t) = \pi h_k \int_{-\infty}^{t} \sum_{n\in\mathbb{Z}} a_{n,k} g_k(u - nT_k)du$$

Dans cette expression, T_k est la période-symbole, h_k est l'indice de modulation et $(a_{n,k})_{n\in\mathbb{Z}}$ est la suite de symboles transmis. Cette suite est une suite binaire ($a_{n,k} = \pm 1$ avec probabilité $1/2$) indépendante et identiquement distribuée. La fonction de mise en forme $g_k(t)$ vérifie $g_k(t) = 0$ si $t < 0$ ou $t > L_k$, et $\int_0^{L_k T_k} g_k(u)du = 1$.

Afin d'augmenter l'efficacité spectrale de ces signaux, un filtre passe-bas de fréquence de coupure $f_c = \frac{1+\gamma_k}{2T_k}$, $\gamma_k \in]0;1[$, est en général appliqué avant la transmission au k^{ieme} signal, et nous noterons $s_{a,k}(t)$ le signal effectivement émis.

Les K signaux sont émis dans une même bande de fréquences, et sont mélangés convolutivement par des canaux de propagation sélectifs en fréquence. Le récepteur dispose de N capteurs, et l'on désigne par $\mathbf{y}_a(t)$ le signal de dimension N formé des N enveloppes complexes $(y_{a,i}(t))_{i=1,\ldots,N}$ des signaux à temps continu reçus sur chacun des capteurs. $\mathbf{y}_a(t)$ est échantillonné à un rythme T_e qui satisfait la condition de Shannon, et on pose $y_i(n) = y_{a,i}(nT_e)$ et $\mathbf{y}(n) = \mathbf{y}_a(nT_e)$.

Puisque T_e vérifie la condition de Shannon, le signal $\mathbf{y}(n)$ coïncide avec la sortie d'un système linéaire K entrées / N sorties de fonction de transfert $\mathbf{H}(z) = [h_{i,j}(z)]_{(i,j)}$ excité par le signal de dimension K $\mathbf{s}(n) = (s_1(n), \ldots, s_K(n))^T$, où $s_k(n) = s_{a,k}(nT_e)$

Les signaux $\{s_k(n)\}$ sont cyclostastionnaires. En particulier, le moment d'ordre 2 de la k^{ieme} source est une fonction périodique du temps, de période $\frac{T_e}{T_k}$. Néanmoins, le filtre passe bas appliqué à l'émission assure, tant que $\gamma_k < 1$, que, en tant que fonction de n, les moments $\mathbb{E}(s_k(n)s_k(n-p)^*)$ possèdent au plus 3 coefficients de Fourier non nuls aux fréquences cycliques : $\alpha_k \in \left\{0, \frac{T_e}{T_k}, -\frac{T_e}{T_k}\right\}$. De plus, si γ_k est nul, alors le signal s_k n'a qu'une seule fréquence cyclique qui est 0, et est donc stationnaire à l'ordre 2.

Pour séparer le mélange, nous proposons d'utiliser une approche par déflation, consistant à déterminer un filtre N entrées / 1 sortie $\mathbf{g}(z)$ de façon à ce que $r(n) = [\mathbf{g}(z)]\mathbf{y}(n)$ coïncide avec une version de l'une des sources. Dans la suite, nous étudions les propriétés séparante du critère de Godard (1).

1. le suffixe "a" précise que la fonction en question est à temps continu

3 Résultats

La sortie du filtre $\mathbf{g}(z)$ excité par les observations s'écrit:

$$r(n) = \sum_{i=1}^{N} [g_i(z)]y_i(n) = \sum_{k=1}^{K} [f_k(z)]s_k(n)$$

où $\mathbf{f}(z) = [f_1(z), ..., f_K(z)]$ est le filtre global $\mathbf{f}(z) = \mathbf{g}(z)\mathbf{H}(z)$. Le système est linéaire, donc le signal $r(n)$ est un signal cyclostationnaire, dont le moment d'ordre 2 se décompose en série de Fourier avec des coefficient non nuls pour les fréquences cycliques décrites par l'ensemble $I = \{0\} \cup \left\{\pm\frac{Te}{T_k} \mid k = 1...K\right\}$:

$$\mathbb{E}\left(r(n)r^*(n-p)\right) = \sum_{\alpha\in I} R_r^{(\alpha)}(p)e^{-2i\pi\alpha n}$$

$R_r^{(\alpha)}(p)$ est appelée cyclo-corrélation de $r(n)$ à la fréquence cyclique α et nous notons $S_r^{(\alpha)}(e^{2i\pi\nu})$ sa transformée de Fourier.

Il est utile d'exprimer la moyenne des cumulants d'ordre 4 de $r(n)$, $< c_4(r(n)) >=< cum(r(n), r^*(n), r(n), r^*(n)) >$. Afin de simplifier la présentation de ce qui suit, nous allons supposer que les indices des modulations CPM utilisés ne sont ni entiers, ni demi-entiers. Les signaux $(s_k)_{k=1,\ldots,K}$ sont alors circulaires à l'ordre 2, de même que $r(n)$. Par conséquent, $\mathbb{E}(r(n)r(n-p) = 0$ pour tous n et p, et $< c_4(r(n)) >=< \mathbb{E}|r(n)|^4 > -2 < |\mathbb{E}|r(n)|^2|^2 >$. Appliquant l'égalité de Parseval et remarquant que $R_r^{(\alpha)}(0) = \left(R_r^{(-\alpha)}(0)\right)^*$ et $\mathbb{E}\left(r(n)^2\right) = 0$, nous pouvons écrire

$$< c_4(r(n)) >=< \mathbb{E}|r(n)|^4 > -2|R_r^{(0)}(0)|^2 - 4\sum_{\alpha\in I_+^*} |R_r^{(\alpha)}(0)|^2 \quad (2)$$

où I_+^* est l'ensemble des fréquences cycliques strictement positives. Toute la difficulté des problèmes de séparation de mélanges convolutifs de sources cyclostationnaires réside dans l'équation (2). En effet, le cumulant moyenné d'ordre 4 dépend de 3 termes : Les 2 premiers peuvent être simplement estimés, mais le troisième terme nécessite de connaître où d'estimer les fréquences cycliques pour être évalué. C'est la raison pour laquelle nous nous intéressons aux propriétés séparantes de la fonction de Godard (1) qui s'estime de façon consistante sans connaissance à priori sur les signaux par $\hat{J}(r(n)) = \frac{1}{N}\sum_{n=0}^{N-1}\left(|r(n)|^2 - 1\right)^2$.

Pour étudier la fonction de Godard, commençons par définir la norme de $f_k(z)$ par $\|f_k\|^2 = \int_{-1/2}^{1/2} |f_k(e^{2i\pi\nu})|^2 S_{s_k}^{(0)}(e^{2i\pi\nu})d\nu$. Notons que cette norme dépend de l'indice k, et qu'il serait en toute rigueur nécessaire de faire apparaitre cette dépendance dans la notation la symbolisant. Pour ne pas trop compliquer les notations, nous préférons nous contenter de la notation $\|f_k\|^2$. Comme dans [5], nous écrivons $r(n)$ sous la forme:

$$r(n) = \sum_{k=1}^{K} \|f_k\|\tilde{s}_k(n)$$

où $\tilde{s}_k(n) = \left[\frac{f_k(z)}{\|f_k\|}\right] s_k(n)$. Notons que $R_{\tilde{s}_k}^{(0)}(0) = 1$ et que si $\forall k \neq k_0$, $\|f_k\| = 0$ et $\|f_{k_0}\| \neq 0$ alors le signal $r(n)$ est égal à une version filtrée de $s_{k_0}(n)$. Dès lors, en posant $\beta_{\tilde{s}_k} =< \mathbb{E}|\tilde{s}_k(n)|^4 >$ et $\lambda(\tilde{s}_{k_1}, \tilde{s}_{k_2}) = \sum_{\alpha\in I_*^+} Re\left(R_{\tilde{s}_{k_1}}^{(\alpha)}(0)R_{\tilde{s}_{k_2}}^{(\alpha)*}(0)\right)$,

la fonction de Godard (1) admet comme expression :

$$J(r(n)) = \sum_{k=1}^{K} \|f_k\|^4 \beta_{\tilde{s}_k} - 2\sum_{k=1}^{K} \|f_k\|^2 + 1 + 4 \sum_{k_1 \neq k_2} \|f_{k_1}\|^2 \|f_{k_2}\|^2 (1 + 2\lambda(\tilde{s}_{k_1}, \tilde{s}_{k_2})) \quad (3)$$

La forme générale de cette expression rend très difficile l'étude directe du minimum global de $J(r(n))$. Dans la suite, nous allons donc utiliser le lemme suivant, et mettre en évidence des minorants bien choisis de J.

Lemme 3.1 *Soit $m(r)$ une fonction positive vérifiant:*

$$\forall r \ , \ J(r) \geq m(r)$$

Alors, si le minimum global de la fonction $m(r)$ est atteint si et seulement si le signal $r(n)$ ne dépend que d'une des sources constituant le mélange, et que $\min_r m(r) = \min_r J(r)$, alors le minimum global de la fonction $J(r)$ est atteint si et seulement si le signal $r(n)$ ne dépend que d'une des sources.

3.1 Cas n°1 : les signaux sources ont des fréquences cycliques différentes

Nous commençons par considérer le cas où les différents émetteurs utilisent des débits différents. Dans ce cas, les différents signaux sources ne partagent pas les mêmes fréquences cycliques, et les termes de corrélation cycliques croisées s'annulent:

$$\forall k_1 \neq k_2, R^{(\alpha)}_{\tilde{s}_{k_1}}(0) R^{(-\alpha)}_{\tilde{s}_{k_2}}(0) = \lambda(\tilde{s}_{k_1}, \tilde{s}_{k_2}) = 0$$

La fonction de Godard (3) se simplifie et on peut remarquer que son expression est la même que dans un contexte stationnaire. Afin d'étudier son minimum global, nous posons

$$\beta_m = \min_k \inf_{f_k} \beta_{\tilde{s}_k} = \min_k \inf_{f_k} < \mathbb{E}|\tilde{s}_k(n)|^4 > \quad (4)$$

et $m(r) = \beta_m \left(\sum_{k=1}^{K} \|f_k\|^4 \right) + 4 \sum_{k_1 \neq k_2} \|f_{k_1}\|^2 \|f_{k_2}\|^2 - 2\sum_{k=1}^{K} \|f_k\|^2 + 1$. Il s'agit d'une fonction beaucoup plus simple que J, puisqu'elle ne dépend en particulier que des $(\|f_k\|)_{k=1,\ldots,K}$. Il est clair que $J(r) \geq m(r)$. Par ailleurs, il est aisé d'établir que si $\beta_m < 2$, alors, le minimum de m, considéré en tant que fonction des paramètres $(\|f_k\|)_{k=1,\ldots,K}$, est égal à $1 - \frac{1}{\beta_m}$, et est atteint si et seulement si tous les $(\|f_k\|)_{k=1,\ldots,K}$ sont nuls sauf 1. Ceci permet de montrer que le minimum global de m est atteint si et seulement si $r(n)$ coïncide avec la version filtrée de la source réalisant le minimum de (4). Dans ces conditions, $\min_r m(r) = \min_r J(r)$, et le lemme (3.1) assure que J est un contraste dès que $\beta_m < 2$. Les signaux CPM sont des signaux de module constant, et donc leur moment d'ordre 4 vaut 1. Néanmoins, nous avons supposé qu'un filtre passe-bas est appliqué en émission pour contrôler la bande passante du signal transmis. On peut raisonnablement supposer que le filtre passe bas n'affecte pas trop la propriété de module constant et que le minimum des moments d'ordre 4 moyennés reste proche de 1, et en tous les cas, inférieur à 2. Cette affirmation est illustrée par quelques applications numériques : voir la courbe de gauche de la figure 3.1. Les filtres de mise en forme sont des fonctions portes (REC) ou des cosinus surélevés (RC) de durée T_s (1) ou $3T_s$ (3), et le facteur d'excès de bande des filtres d'émission varie entre 0 et 1.

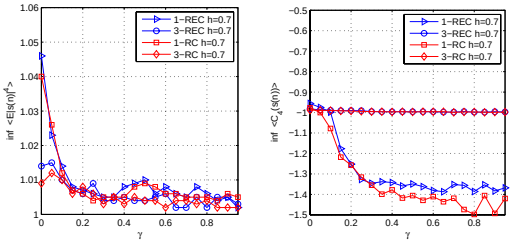


FIG. 1: *Statistiques d'ordre 4 pour un ensemble de modulation CPM*

3.2 Cas n°2: les signaux sources ont tous les mêmes fréquences cycliques

Nous considérons à présent le cas où les sources peuvent partager des fréquences cycliques. Certains termes $\{\lambda(\tilde{s}_{k_1}, \tilde{s}_{k_2})\}$ ne s'annulent pas, ce qui complique considérablement le problème. Pour simplifier notre étude, nous supposons dans un premier temps que toutes les sources ont le même débit. Le signal reçu ne possède donc qu'une seule fréquence cyclique strictement positive α. Afin d'étudier le minimum global de $J(r)$, nous allons de nouveau utiliser le lemme 3.1. La mise en évidence d'un minorant permettant de conclure est toutefois délicate, et nous nous contentons d'esquisser dans la suite les principaux éléments de notre démarche.

Pour construire un minorant de $J(r)$ (1), nous introduisons $\Lambda(x) = |R^{(\alpha)}_x(0)|$, et constatons que $\lambda(\tilde{s}_{k_1}, \tilde{s}_{k_2}) > -\Lambda(\tilde{s}_{k_1})\Lambda(\tilde{s}_{k_2})$. Par ailleurs, il est possible d'établir que pour tout k, et pour tout filtre SISO $\tilde{f}_k$ de norme 1, alors

$$\Lambda(\lceil \tilde{f}_k(z) \rceil s_k) \leq \frac{1}{2}$$

Puisque les signaux $(\tilde{s}_k)_{k=1,\ldots,K}$ sont construits à partir de filtres de norme 1, on peut constater que $J(r)$ est minorée par la fonction $\beta_m \left(\sum_{k=1}^{K} \|f_k\|^4 \right) + 2\sum_{k_1 \neq k_2} \|f_{k_1}\|^2 \|f_{k_2}\|^2 - 2\sum_{k=1}^{K} \|f_k\|^2 + 1$. Malheureusement, considérée en tant que fonction des $(\|f_k\|)_{k=1,\ldots,K}$, cette fonction n'atteint son minimum en un point du type $\|f_k\| = \delta(k - k_0)\|f_{k_0}\|$ que dans le cas où $\beta_m < 1$. Puisque β_m est bien sur plus grand que 1, il convient d'utiliser un minorant moins grossier. Pour ceci, il faut remarquer que si un filtre de norme 1 $\tilde{f}_k$ est tel que $\Lambda(\lceil \tilde{f}_k(z) \rceil s_k)$ est proche de sa valeur maximale $1/2$, alors, la moyenne temporelle du moment d'ordre 4 de $[\tilde{f}_k(z)]s_k(n)]$ s'écarte très notablement de 1. Les paramètres $\beta_{\tilde{s}_k}$ et $\Lambda(\tilde{s}_k)$ sont donc reliés, et ceci peut être utilisé pour construire un minorant bien adapté. Plus formellement, on peut montrer le lemme suivant.

Lemme 3.2 *Soit F_k l'ensemble des filtres SISO de norme 1, i.e. $\tilde{f}_k(z) \in F_k \Leftrightarrow \|\tilde{f}_k\| = 1$. Soit $\Lambda_* \in]0, 1/2[$, $F_k = F_{1,k}(\Lambda_*) \cup F_{2,k}(\Lambda_*)$ avec :*

$$F_{1,k}(\Lambda_*) = \left\{ \tilde{f}_k \in F / \Lambda(\lceil \tilde{f}_k(z) \rceil s_k) < \Lambda_* \right\}$$

$$F_{2,k}(\Lambda_*) = \left\{ \tilde{f}_k \in F / \Lambda(\lceil \tilde{f}_k(z) \rceil s_k) \geq \Lambda_* \right\}$$

Définissons δ par $\delta = \min_k \min_{\tilde{f}_k} < c_4(\tilde{s}_k(n)) >$ où $\tilde{s}_k(n) = [\tilde{f}_k(z)]s_k(n)$, et posons $\beta_* = \delta + 2 + 4\Lambda_*^2$. Alors, si $\tilde{f}_k \in F_{2,k}(\Lambda_*)$, $\beta_{\tilde{s}_k} \geq \beta_*$, où $\tilde{s}_k(n) = [\tilde{f}_k(z)]s_k(n)$.

A chaque valeur de $\Lambda_* \in]0, 1/2[$ va correspondre un minorant, et nous avons établi qu'il existe des valeurs de Λ_* pour

lesquelles le minorant correspondant satisfait les conditions du lemme 3.1 (voir plus loin). Il convient de noter que pour calculer β_*, il est nécessaire d'évaluer le terme δ. Son évaluation analytique est malheureusement impossible, et nous avons donc procédé numériquement en l'évaluant en fonction du facteur d'excès de bande γ pour diverses modulations CPM. Sur l'ensemble des cas traités (voir la courbe de droite de la figure 3.1), nous pouvons constater que $\delta \simeq -1.5$.

Soit $\mathbf{f}(z) = (f_1(z), \ldots, f_K(z))$ une fonction de transfert, et posons $\tilde{f}_k(z) = \frac{f_k(z)}{\|f_k\|}$. Si P composantes (les P premières sans restriction) de $(\tilde{f}_1(z), \ldots, \tilde{f}_K(z))$ appartiennent à $(F_{1,k}(\Lambda_*))_{k\leq P}$ et que les $K-P$ restantes (les $K-P$ dernières) appartiennent à $(F_{2,k}(\Lambda_*))_{k=P+1,\ldots,K}$, alors, il est facile de constater que $J(r) \geq m_P(r)$ où $m_P(r)$ est défini par

$$\begin{aligned} m_P(r(n)) &= \beta_m \sum_{k=1}^{P} \|f_k\|^4 + \beta_* \sum_{k=P+1}^{K} \|f_k\|^4 \\ &+ 4 \sum_{k_1\neq k_2\in[1\cdots P]^2} \|f_{k_1}\|^2\|f_{k_2}\|^2 \left(1-2\Lambda_*^2\right) \\ &+ 8 \sum_{k_1\neq k_2\in[1\cdots P]\cup[P+1\cdots K]} \|f_{k_1}\|^2\|f_{k_2}\|^2 \left(1-\Lambda_*\right) \\ &+ 4 \sum_{k_1\neq k_2\in[P+1\cdots K]^2} \|f_{k_1}\|^2\|f_{k_2}\|^2 \left(\frac{1}{2}\right) \\ &- 2\sum_{k=1}^{K} \|f_k\|^2 + 1 \end{aligned} \quad (5)$$

On peut alors constater que $m_P(r)$ ne dépend que des $(\|f_k\|)_{k=1,\ldots,K}$. Pour chaque valeur de P, on peut calculer ses points stationnaires, et les valeurs qu'elle y prend. Après quelques calculs, on peut alors montrer que si

$$\beta_m < 2(1-2\Lambda_*^2) \quad (6)$$

$$\beta_m < \frac{\beta_*-1}{K} + 1 \quad (7)$$

$$\beta_m < \frac{4(K-1)(1-\lambda_*)^2}{K-2+\beta_*} \quad (8)$$

alors, pour tout $0 \leq P \leq K$, $m_P(r) \geq (1-\frac{1}{\beta_m})$, l'égalité ayant lieu si et seulement si tous les $\|f_k\|$ sont nuls, sauf l'un des P premiers. On rappelle que $\beta_* = \delta_* + 2 + 4\Lambda_*^2$, et que ces 3 inégalités expriment donc des contraintes sur $(\beta_m, \delta_*, \Lambda_*)$. Il convient de noter que les inégalités mettent également en jeu le nombre de sources K, et que plus K est grand, plus les conditions (6, 7, 8) sont sévères. A titre d'exemple, sur la "base de données" de modulations CPM utilisées dans les figures 1 et 2, $\beta_m = 1.05, \delta_* = -1.5$, et la quantité $\Lambda_* = 0.477$ permet aux 3 conditions d'être vérifiées jusqu'à $K = 8$. Le lemme 3.1 permet alors de montrer que la minimisation de J permet de séparer jusqu'à 8 sources de la base de données.

3.3 Cas n°3: Si chaque émetteur a le même débit qu'au plus un autre émetteur

Ce cas peut se traiter comme le précédent. Pour ceci, on partitionne les sources par groupe, un groupe étant défini comme un ensemble de sources partageant le même débit. Les termes de corrélation croisés inter-groupes s'annulent (cas n°1), et les termes de corrélation croisés intra-groupes dépendent de la répartition des sources dans $F_1(\Lambda_*)$ et $F_2(\Lambda_*)$ (cas n°2). On obtient comme dans le cas n°2 un certain nombre de fonctions similaires à $m_P(r)$, dont on peut étudier les points stationnaires. De nouveau, on peut mettre en évidence des conditions sous lesquelles la minimisation du critère de Godard permet la séparation.

4 Illustrations

Nous avons simulé un mélange de 4 signaux CPM: chaque signal a un débit choisi aléatoirement qui vaut $T_s = 3$ ou $T_s = 4$, un filtre de mise en forme choisi parmis les suivants: 1-REC, 3-REC, 1-RC, 3-RC et un indice de modulation h qui vaut 0.7. Le canal de propagation simulé est un canal multi-trajets avec au plus 3 trajets présentant chacun une atténuation de Rayleigh et les observations sont échantillonnées au rythme $T_e = 1$.

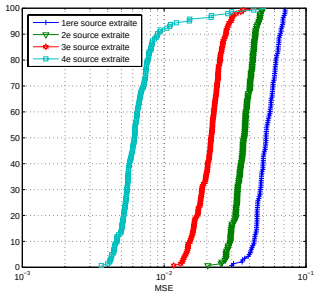


FIG. 2: *Performance de l'algorithme de séparation pour un mélange de 4 CPM*

Pour mesurer la qualité de la séparation de la source k, nous utilisons le critère suivant:

$$MSE(k) = \sum_{l=1}^{N} \frac{< |c_{k,l}(n) - \hat{c}_{k,l}(n)|^2 >}{< |y_l(n)|^2 >}$$

où $c_{k,l}(n)$ est la contribution de la source k sur le capteur l, et $\hat{c}_{k,l}(n)$ sa version estimée calculée par l'algorithme de séparation. Les résultats obtenus, très concluants, sont illustrés par la figure 1, où nous avons représenté la fonction de répartition des MSE en fonction du numéro de la source extraite.

Références

[1] P. Comon. Contrasts for multichannel blind deconvolution. *IEEE Trans. Signal Processing*, 3:209–211, July 1996.

[2] Pierre Jallon, Antoine Chevreuil, Philippe Loubaton, and Pascal Chevalier. Separation of convolutive mixtures of cyclostationary sources: a contrast function based approach. In *proc. ICA'04*, Grenade, Spain, September 2004.

[3] Pierre Jallon, Antoine Chevreuil, Philippe Loubaton, and Pascal Chevalier. Separation of convolutive mixtures of linear modulated signals using constant modulus algorithm. In *ICASSP 2005*, Philadelphia, USA, 2005.

[4] P.A. Regalia. On the equivalence vetween the godard and shalvi-weinstein schemes of blind equalization. *Signal Processing*, 73:185–190, 1999.

[5] C. Simon, Ph Loubaton, and C. Jutten. Separation of a class of convolutive mixtures: a contrast function approach. *Signal Processing*, 81:883–887, 2001.

[6] J.K. Tugnait. Blind spatio-temporal equalization and impulse response estimation for mimo channels using a godard cost function. *IEEE Trans. Signal Processing*, 1997.

[7] J.K. Tugnait. Adaptive blind separation of convolutive mixtures of independant linear signals. *Signal Processing*, 1999.

Approximation en champ moyen pour la séparation de sources appliquée aux images hyperspectrales

Nadia Bali et Ali Mohammad–Djafari

Laboratoire des Signaux et Systmes (cnrs – Supélec – ups)
Supélec , Plateau de Moulon, 91192 Gif-sur-Yvette Cedex, France
{bali, djafari}@lss.supelec.fr

Résumé – Nous considérons le problème de séparation aveugle de sources que l'on trouve dans le traitement d'images hyperspectrales avec des distributions a priori modélisées par des champ de Markov composites où les variables cachées reprsentent soit les contours soit les étiquettes des régions. La mise en oeuvre des estimateurs bayésiens des sources, des variables cachées, de la matrice de mélange et les hyperparamètres du modèle de séparation de source peut se faire par des méthodes d'échantillonnage de Gibbs. Cette algorithme présente une complexité et un coût excessive ce qui limite son utilisation dans des applications pratiques. Dans ce papier, nous proposons une approche d'approximation en champ moyen qui consiste à approcher la loi a posteriori jointe par une loi séparable. Les performances relatives de ces modèles avec ou sans approximation en champ moyen sont comparées.

Abstract – We consider the Blind Source Separation (BSS) of images whose prior distributions are modelled through a compound Gauss-Markov modeling with either a hidden contour process, or a hidden classification labels modelled with a Potts distribution.The joint estimation of sources, hidden variables, mixing matrix and all the hyperparameters of the model (noise covariance, means and variances of the pixels in each region) can be done through the Gibbs sampling of the joint posterior probability law of these unknowns, given the observed mixed data. In previous works, we had implemented this general algorithm. However the huge complexity and cost of this posterior law limit its use in practical applications. In this paper, we propose a Mean Field Approximation (MFA) approach which consists in approximating this joint posterior by a separable one whose sampling will be much easier and less expensive. We then compare the relative performances of the models with regards of the MFA.

1 Introduction

Une image de télédétection est une image échantillonnée spatialement acquise graçe à un capteur survolant une scène. Contrairement aux images monospectrales, ou de niveaux de gris, qui fournissent pour chaque pixel, une valeur scalaire correspondant à l'énergie du rayonnement sur une bande de longueur d'onde mesurée [1], les images hyperspectrales fournissent, pour chaque pixel, des mesures sur plusieurs bandes.
Les données d'une image hyperspectrale peuvent être considérées de plusieurs façons :

- une collection d'observations scalaires sur un espace 3D $X(\omega, \boldsymbol{r})$, où $\omega \in \Omega$ représente la longueur d'onde et $\boldsymbol{r} \in \mathcal{R}^2$ la position d'un pixel;
- une collection d'images $X_\omega(\boldsymbol{r})$;
- une collection de spectres $X_{\boldsymbol{r}}(\omega)$.

La réduction de données constitue une étape importante dans le processus de traitement d'images hyperspectrales si on considère l'énorme quantité d'informations à traiter. Les modèles de réduction de données opèrent :

— soit sur les spectres :

$$\begin{aligned} X_{\boldsymbol{r}}(\omega) &= \sum_{z=1}^{K} A_{\boldsymbol{r},z} S_z(\omega) \\ \boldsymbol{X}(\omega) &= \{X_{\boldsymbol{r}}(\omega), \boldsymbol{r} \in \mathcal{R}\} \\ \boldsymbol{S}(\omega) &= \{S_z(\omega), z = 1, \cdots, K\} \\ \boldsymbol{X}(\omega) &= \boldsymbol{A}\boldsymbol{S}(\omega) \end{aligned} \tag{1}$$

où $\{S_z(\omega), z = 1, \cdots, K\}$ sont les spectres principaux sur lesquels on décompose tous les autres spectres.
— soit sur les images :

$$\begin{aligned} X_\omega(\boldsymbol{r}) &= \sum_{z=1}^{K} A_{\omega,z} S_z(\boldsymbol{r}) \\ \boldsymbol{X}(\boldsymbol{r}) &= \{X_\omega(\boldsymbol{r}), \omega \in \Omega\} \\ \boldsymbol{S}(\boldsymbol{r}) &= \{S_z(\boldsymbol{r}), z = 1, \cdots, K\} \\ \boldsymbol{X}(\boldsymbol{r}) &= \boldsymbol{A}\boldsymbol{S}(\boldsymbol{r}) \end{aligned} \tag{2}$$

où $\{S_z(\boldsymbol{r}), z = 1, \cdots, K\}$ sont les images principales sur lesquelles on décompose toutes les autres images.

Dans les deux cas la matrice $\boldsymbol{A}$ est la matrice de décomposition qu'il faut estimer en même temps que les sources $\boldsymbol{S}$. On retrouve alors les problèmes de l'analyse en composantes principales (ACP), de l'analyse en composantes indépendantes (ACI) et de la séparation de sources. Dans le cadre d'une approche bayésienne, nous proposons deux modèles *a priori* différents sur les sources : Le premier

est un modèle markovien composite avec un processus de ligne [2]. Le deuxième est un modèle markovien composite qui introduit un processus d'étiquettes pour les régions [3].

2 Méthodes proposées

Les deux modèles de séparation de sources (1) et (2) peuvent s'écrire de la manière générique suivante :

$$\boldsymbol{x}(\boldsymbol{r}) = \boldsymbol{A}\boldsymbol{s}(\boldsymbol{r}) + \boldsymbol{\epsilon}(\boldsymbol{r}) \tag{3}$$

où $\boldsymbol{x}(\boldsymbol{r})$ représente le vecteur d'observations, $\boldsymbol{s}(\boldsymbol{r})$ représente le vecteur des sources inconnues, $\boldsymbol{A}$ est une matrice de mélange de dimension $(m \times n)$ et $\boldsymbol{\epsilon}(\boldsymbol{r})$ représente le bruit et les incertitudes de modèle. Les bruits ϵ_j sont supposés indépendants centrés, blancs et gaussiens de variance $\left\{\sigma^2_{\epsilon_j},\ j = 1, \cdots, m\right\}$. Cette hypothèse nous donne :

$$p(\boldsymbol{x}|\boldsymbol{s}, A) \propto \mathcal{N}(\boldsymbol{A}\boldsymbol{s}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$$

où $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = diag(\sigma^2_{\epsilon_1}, \cdots, \sigma^2_{\epsilon_m})$.

Dans l'approche bayésienne, une étape importante est la modélisation des lois *a priori* pour les sources. Dans notre travail nous considérons les deux modèles *a priori* suivants :

Modèle 1 : Un champ de Markov composite (intensité - contours) où le champ d'intensité $s_\omega(\boldsymbol{r})$ est un champ de Gauss-Markov et quatre champs de contours $l_\omega(\boldsymbol{r}, \boldsymbol{r}'), \boldsymbol{r}' \in \mathcal{V}(\boldsymbol{r})$ qui sont supposés séparables et indépendants :

$$\begin{aligned} &p(s_\omega(\boldsymbol{r}), l_\omega(\boldsymbol{r}, \boldsymbol{r}'), \boldsymbol{r} \in \mathcal{R}) \propto \exp\left[-U(s_\omega(\boldsymbol{r}), l_\omega(\boldsymbol{r}, \boldsymbol{r}'))\right] \\ &U = \sum_{\boldsymbol{r}} \sum_{\boldsymbol{r}' \in \mathcal{V}(\boldsymbol{r})} \lambda_\omega(s_\omega(\boldsymbol{r}) - s_\omega(\boldsymbol{r}'))^2(1 - l_\omega(\boldsymbol{r}, \boldsymbol{r}')) \\ &\qquad + \alpha_\omega l_\omega(\boldsymbol{r}, \boldsymbol{r}') \end{aligned} \tag{4}$$

Ce modèle avait été déjà proposé par [4] dans le cadre de séparation de sources appliqué au traitement des images de documents anciens.

Modèle 2 : Un champ de Markov composite (intensité - régions) où le champ d'intensité est gaussien mais séparable, par contre le champ des étiquettes suit un modèle markovien (Potts) :

$$\begin{aligned} &p(s_\omega(\boldsymbol{r}), z_\omega(\boldsymbol{r})\boldsymbol{r} \in \mathcal{R}) \propto \exp\left[-U(s_\omega(\boldsymbol{r}), z_\omega(\boldsymbol{r}))\right] \\ &U = \frac{1}{2}\sum_{z=1}^{K} \sum_{\boldsymbol{r} \in \mathcal{R}_z} \left[\frac{(s_\omega(\boldsymbol{r}) - m_{\omega z})^2}{\sigma^2_{\omega z}}\right] \\ &\qquad + \sum_{\boldsymbol{r} \in \mathcal{R}} \sum_{\boldsymbol{r}' \in \mathcal{V}(\boldsymbol{r})} \alpha_\omega(z(\boldsymbol{r}) - z(\boldsymbol{r}')) \end{aligned} \tag{5}$$

où $\mathcal{R}_z = \{\boldsymbol{r} \ : Z(\boldsymbol{r}) = z\}$, $\mathcal{R} = \bigcup_{z=1}^{K} \mathcal{R}_z$ et $\sum_{\boldsymbol{r}} = \sum_z \sum_{\boldsymbol{r} \in \mathcal{R}_z}$.

Dans les deux cas notre objectif est d'obtenir la loi *a posteriori* :

$$p(\boldsymbol{s}, \boldsymbol{c}, \boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{x}|\boldsymbol{s}, \boldsymbol{c}, \boldsymbol{\theta})\ p(\boldsymbol{s}, \boldsymbol{c}|\boldsymbol{\theta})\ p(\boldsymbol{\theta})$$

où $\boldsymbol{c}$ représente soit $\boldsymbol{l}$ ou $\boldsymbol{z}$ et $\boldsymbol{\theta}$ soit $(\mathcal{A}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ ou $(\boldsymbol{A}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, \{m_{\omega z}, \sigma^2_{\omega z}\})$ qui regroupe la matrice de mélange, les variances de bruit et les autres paramètres du modèle *a priori*. C'est à partir de cette loi que nous pouvons définir des estimateurs pour toutes nos inconnues. L'approche classique est d'utiliser l'échantillonneur de Gibbs [3] pour générer itérativement les échantillons $(\boldsymbol{s}, \boldsymbol{c}, \boldsymbol{\theta})$ suivant les lois a *posteriori* appropriés :

$$\boldsymbol{s} \sim p(\boldsymbol{s}|\boldsymbol{c}, \boldsymbol{\theta}, \boldsymbol{x}) \longrightarrow \boldsymbol{c} \sim p(\boldsymbol{c}|\boldsymbol{s}, \boldsymbol{\theta}, \boldsymbol{x}) \longrightarrow \boldsymbol{\theta} \sim p(\boldsymbol{\theta}|\boldsymbol{s}, \boldsymbol{c}, \boldsymbol{x})$$

Cette méthode est très coûteuse en temps de calcul. Afin d'aboutir à une méthode plus approchée mais moins coûteuse, nous avons choisie d'utiliser l'approximation en champ moyen [5, 2]. Le champ moyen est à l'origine une méthode d'approximation pour le calcul de l'espérance d'un champ de Markov. L'approximation en champ moyen fournit également une approximation de la distribution d'un champ de Markov. Plus précisément, la méthode permet d'approcher un système markovien, avec des interactions complexes, par un système de variables indépendantes, plus facile à manipuler.

2.1 Approximation champ moyen pour les contours

La théorie de l'approximation en champ moyen suggère que dans le calcul des statistiques d'un élément singulier dans un champ aléatoire l'effet de tous les autres éléments peut être approcher par l'effet de leur moyenne. En appliquant cette approximation à notre modèle *a priori* (4) on obtient :

$$p(\underline{\boldsymbol{s}}, \underline{\boldsymbol{l}}|\boldsymbol{\theta}, \underline{\boldsymbol{x}}) \simeq \prod_{\boldsymbol{r}} p_{\boldsymbol{r}}(\boldsymbol{s}(\boldsymbol{r}), \boldsymbol{l}(\boldsymbol{r})|\bar{\boldsymbol{s}}(\boldsymbol{r}), \bar{\boldsymbol{l}}(\boldsymbol{r}), \boldsymbol{\theta}, \boldsymbol{x}(\boldsymbol{r}))$$

où $\bar{\boldsymbol{s}}(\boldsymbol{r})$ est la moyenne de la loi $p_r(\boldsymbol{s}(\boldsymbol{r})|\bar{\boldsymbol{s}}(\boldsymbol{r}'), \bar{\boldsymbol{l}}(\boldsymbol{r}), \boldsymbol{\theta}, \boldsymbol{x}(\boldsymbol{r}))$ et $\bar{\boldsymbol{l}}(\boldsymbol{r}) = \sum_{\boldsymbol{r}' \in \mathcal{V}(\boldsymbol{r})} \bar{\boldsymbol{l}}(\boldsymbol{r}, \boldsymbol{r}')$ est la moyenne de la loi $p(\boldsymbol{l}(\boldsymbol{r}))|\bar{\boldsymbol{s}}(\boldsymbol{r}), \boldsymbol{\theta}, \boldsymbol{x}(\boldsymbol{r}))$. Pour la première nous avons :

$$p_{\boldsymbol{r}}(\boldsymbol{s}(\boldsymbol{r})|\bar{\boldsymbol{s}}(\boldsymbol{r}'), \bar{\boldsymbol{l}}(\boldsymbol{r}), \boldsymbol{\theta}, \boldsymbol{x}(\boldsymbol{r})) \propto \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{r}), \boldsymbol{B}(\boldsymbol{r}))$$

$$\begin{cases} \boldsymbol{B}(\boldsymbol{r}) &= [\boldsymbol{A}^t\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{A} + 2\sum_{\boldsymbol{r}' \in \mathcal{V}(\boldsymbol{r})} \Lambda_{\boldsymbol{r}}(\boldsymbol{r}')]^{-1} \\ \boldsymbol{\mu}(\boldsymbol{r}) &= \boldsymbol{B}(\boldsymbol{r})[\boldsymbol{A}\boldsymbol{\Sigma}_\epsilon^{-1}\boldsymbol{x}(\boldsymbol{r}) + 2\sum_{\boldsymbol{r}' \in \mathcal{V}(\boldsymbol{r})} \Lambda_{\boldsymbol{r}}(\boldsymbol{r}')\bar{s}(\boldsymbol{r}')] \end{cases}$$

$\Lambda_{\boldsymbol{r}}(\boldsymbol{r}') = \text{diag}\left[\boldsymbol{\lambda}. * (\boldsymbol{1} - \bar{\boldsymbol{l}}(\boldsymbol{r}, \boldsymbol{r}'))\right]$ avec $\boldsymbol{\lambda} = \{\lambda_\omega, \omega \in \Omega\}$. La deuxième est une loi exponentielle avec la moyenne :

$$\bar{l}_\omega(\boldsymbol{r}) = \sum_{\boldsymbol{r}' \in \mathcal{V}(\boldsymbol{r})} \frac{\exp\left[-\alpha_\omega\right]}{\exp\left[-\lambda_\omega(\bar{s}_\omega(\boldsymbol{r}) - \bar{s}_\omega(\boldsymbol{r}'))^2\right] + \exp\left[-\alpha_\omega\right]}$$

Ainsi l'algorithme revient à calculer itérativement :

$$\begin{cases} \bar{\boldsymbol{s}} &= E_{\boldsymbol{s}}(\boldsymbol{s}|\bar{\boldsymbol{l}}, \widehat{\boldsymbol{\theta}}, \boldsymbol{x}) \longrightarrow \bar{\boldsymbol{s}}(\boldsymbol{r}) = \boldsymbol{\mu}(\boldsymbol{r}) \\ \bar{\boldsymbol{l}} &= E_{\boldsymbol{l}}(\boldsymbol{l}|\bar{\boldsymbol{s}}, \widehat{\boldsymbol{\theta}}, \boldsymbol{x}) \\ \widehat{\boldsymbol{A}} &= [\sum_{\boldsymbol{r}} \boldsymbol{x}(\boldsymbol{r})\bar{\boldsymbol{s}}'(\boldsymbol{r})][\sum_{\boldsymbol{r}} \bar{\boldsymbol{s}}(\boldsymbol{r})\bar{\boldsymbol{s}}'(\boldsymbol{r}) + \boldsymbol{B}(\boldsymbol{r})]^{-1} \\ \widehat{\boldsymbol{\Sigma}}_\epsilon &= [\sum_{\boldsymbol{r}} \boldsymbol{x}(\boldsymbol{r})\boldsymbol{x}'(\boldsymbol{r})] - [\sum_{\boldsymbol{r}} \boldsymbol{x}(\boldsymbol{r})\bar{\boldsymbol{s}}'(\boldsymbol{r})] \\ &\quad [\sum_{\boldsymbol{r}} \bar{\boldsymbol{s}}(\boldsymbol{r})\bar{\boldsymbol{s}}'(\boldsymbol{r}) + \boldsymbol{B}(\boldsymbol{r})]^{-1}[\sum_{\boldsymbol{r}} \bar{\boldsymbol{s}}(\boldsymbol{r})\boldsymbol{x}'(\boldsymbol{r})] \end{cases}$$

2.2 Approximation champ moyen pour les régions

Avec le modèle *a priori* (5) pour les sources on a :

$$p(\underline{\boldsymbol{s}}, \underline{\boldsymbol{z}}|\boldsymbol{\theta}, \underline{\boldsymbol{x}}) = p(\underline{\boldsymbol{s}}|\underline{\boldsymbol{z}}, \underline{\boldsymbol{x}}, \boldsymbol{\theta})\ p(\underline{\boldsymbol{z}}|\underline{\boldsymbol{x}}, \boldsymbol{\theta})$$

où $p(\underline{\boldsymbol{s}}|\underline{\boldsymbol{z}}, \underline{\boldsymbol{x}}, \boldsymbol{\theta})$ est une gaussienne et $p(\underline{\boldsymbol{z}}|\underline{\boldsymbol{x}}, \boldsymbol{\theta})$ est la distribution d'un champ de Potts. Appliquant le même schéma

de l'approximation en champ moyen, l'objectif est d'approcher $p(\underline{\boldsymbol{s}}, \underline{\boldsymbol{z}}|\boldsymbol{\theta}, \boldsymbol{x})$ par $\prod_{\boldsymbol{r}} p_{\boldsymbol{r}}(\boldsymbol{s}(\boldsymbol{r}), \boldsymbol{z}(\boldsymbol{r})|\bar{\boldsymbol{s}}(\boldsymbol{r}), \bar{\boldsymbol{z}}(\boldsymbol{r}), \boldsymbol{\theta}, \boldsymbol{x}(\boldsymbol{r}))$ ce qui donne :

$$p_{\boldsymbol{r}}(\boldsymbol{s}(\boldsymbol{r})|\bar{\boldsymbol{z}}(\boldsymbol{r}), \boldsymbol{\theta}, \boldsymbol{x}(\boldsymbol{r})) \propto \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{r}), \boldsymbol{B}(\boldsymbol{r}))$$

avec

$$\begin{cases} \boldsymbol{B}(\boldsymbol{r}) = \left[\boldsymbol{A}^t \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{A} + \sum_{\bar{\boldsymbol{z}}(\boldsymbol{r})} \boldsymbol{\Sigma}_{z(\boldsymbol{r})}^{-1}\right]^{-1} \\ \boldsymbol{\mu}(\boldsymbol{r}) = \boldsymbol{B}(\boldsymbol{r})[\boldsymbol{A}^t \boldsymbol{\Sigma}_\epsilon^{-1} \boldsymbol{x}(\boldsymbol{r}) + \sum_{\bar{\boldsymbol{z}}(\boldsymbol{r})} \boldsymbol{\Sigma}_{z(\boldsymbol{r})}^{-1} \boldsymbol{m}_{z(\boldsymbol{r})}] \end{cases}$$

et l'approximation en champ moyen pour $\boldsymbol{z}$ donne :

$$\begin{aligned} p_{\bar{\boldsymbol{z}}}(\boldsymbol{z}(\boldsymbol{r})|\boldsymbol{x}(\boldsymbol{r}), \boldsymbol{\theta}) &\propto p(\boldsymbol{z}(\boldsymbol{r})|\boldsymbol{x}(\boldsymbol{r}), \bar{\boldsymbol{z}}(\boldsymbol{r}') \in \mathcal{V}(\boldsymbol{r}), \boldsymbol{\theta}) \\ \bar{\boldsymbol{z}}(\boldsymbol{r}) &= \frac{\sum_{\boldsymbol{z}(\boldsymbol{r})} \boldsymbol{z}(\boldsymbol{r}) p_{\bar{\boldsymbol{z}}}(\boldsymbol{z}(\boldsymbol{r})|\boldsymbol{x}(\boldsymbol{r}), \boldsymbol{\theta})}{\sum_{\boldsymbol{z}(\boldsymbol{r})} p_{\bar{\boldsymbol{z}}}(\boldsymbol{z}(\boldsymbol{r})|\boldsymbol{x}(\boldsymbol{r}), \boldsymbol{\theta})} \end{aligned}$$

Cette approche revient donc à calculer itérativement :

$$\begin{cases} \bar{\boldsymbol{s}}(\boldsymbol{r}) = \mathrm{E}\left\{\boldsymbol{s}(\boldsymbol{r})|\boldsymbol{x}(\boldsymbol{r}), \bar{\boldsymbol{z}}(\boldsymbol{r}), m_{\omega z}, \sigma_{\omega z}^2, \boldsymbol{\theta}\right\} = \boldsymbol{\mu}(\boldsymbol{r}) \\ \bar{\boldsymbol{z}}(\boldsymbol{r}) = \mathrm{E}\left\{\boldsymbol{z}(\boldsymbol{r})|\bar{\boldsymbol{z}}(\boldsymbol{r}'), \boldsymbol{x}(\boldsymbol{r}), \boldsymbol{\theta}\right\} \\ \widehat{\boldsymbol{A}} = [\sum_{\boldsymbol{r}} \boldsymbol{x}(\boldsymbol{r})\bar{\boldsymbol{s}}'(\boldsymbol{r})][\sum_{\boldsymbol{r}} \bar{\boldsymbol{s}}(\boldsymbol{r})\bar{\boldsymbol{s}}'(\boldsymbol{r}) + \boldsymbol{B}(\boldsymbol{r})]^{-1} \\ \widehat{\boldsymbol{\Sigma}}_\epsilon = [\sum_{\boldsymbol{r}} \boldsymbol{x}(\boldsymbol{r})\boldsymbol{x}'(\boldsymbol{r})] - [\sum_{\boldsymbol{r}} \boldsymbol{x}(\boldsymbol{r})\bar{\boldsymbol{s}}'(\boldsymbol{r})] \\ \qquad [\sum_{\boldsymbol{r}} \bar{\boldsymbol{s}}(\boldsymbol{r})\bar{\boldsymbol{s}}'(\boldsymbol{r}) + \boldsymbol{B}(\boldsymbol{r})]^{-1}[\sum_{\boldsymbol{r}} \bar{\boldsymbol{s}}(\boldsymbol{r})\boldsymbol{x}'(\boldsymbol{r})] \\ m_{\omega z} = \frac{\sum_r \bar{s}_\omega(\boldsymbol{r}) p(z_\omega(\boldsymbol{r})|\bar{z}_\omega(\boldsymbol{r}'), x_\omega(\boldsymbol{r}), \boldsymbol{\theta})}{\sum_r P(z_\omega(\boldsymbol{r})|\bar{z}_\omega(\boldsymbol{r}'), x_\omega(\boldsymbol{r}), \boldsymbol{\theta})} \\ \sigma_{\omega z}^2 = \frac{\sum_r (\bar{s}_\omega(\boldsymbol{r}) - m_{\omega z})^2 p(z_\omega(\boldsymbol{r})|\bar{z}_\omega(\boldsymbol{r}'), x_\omega(\boldsymbol{r}), \boldsymbol{\theta})}{\sum_r P(z_\omega(\boldsymbol{r})|\bar{z}_\omega(\boldsymbol{r}'), x_\omega(\boldsymbol{r}), \boldsymbol{\theta})} \end{cases}$$

3 Résultats

Avec l'objective de comparer les performances relatives de ces modèles, nous commençons par génèrer artificiellement les étiquettes $\{z_j; j = 1 \cdots, n\}$ en utilisant un modèle de Potts avec K_j niveaux correspondant au nombre de classes dans une source j et α_j comme paramètre du modèle de Potts. Par la suite on génére les sources suivent les probabilités conditionnelles $p(s_j|z_j)$, et finalement on génére les mélanges $\{x_j; j = 1 \cdots, n\}$ en utilisant une matrice de mélange $\boldsymbol{A}$ et en rajoutant un bruit de covariance $\boldsymbol{\Sigma}_\epsilon$ suivant la probabilité conditionnelle $p(\boldsymbol{x}|\boldsymbol{s}, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{A})$.

Figure 1 montre un exemple avec les parmètres suivents :
$m = 2, n = 2, K = [3, 2], \alpha = [2, 1.5]$ pour la génération des étiquettes $\{Z_j; j = 1 \cdots, n\}$ les moyennes et le variances pour les sources sont gńérer de façon aléatoire pour générer les sources $\{S_j, j = 1, \cdots, n\}$,
$\boldsymbol{\Sigma}_\epsilon = \mathrm{diag}\,[.01, .01]$ et $\boldsymbol{A} = \begin{bmatrix} .7143 & .3333 \\ .2857 & .6667 \end{bmatrix}$ pour la génération des mélanges
$\{X_i, i = 1, \cdots, m\}$.

Figure 2 montre les résultats que l'on obtient avec notre méthode d'approximation en champ moyen utilisant le modèle intensité–région. La matrice de mélange estimée est $\bar{\boldsymbol{A}} = \begin{bmatrix} .6693 & .3527 \\ .2743 & .6957 \end{bmatrix}$ et d'où une bonne séparation des deux sources.

Figure 3 montre les résultats obtenu par la méthode d'approximation en champ moyen utilisant le modèle intensité–contour on obtient $\bar{\boldsymbol{A}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$. On peut constater que cette méthode n'éstime pas bien la matrice de mélange.

Figure 4 montre les résultats obtenu par la méthode d'échantillonnage de Gibbs intensité–région où on obtient $\bar{\boldsymbol{A}} = \begin{bmatrix} .8794 & -.9254 \\ .4761 & -.3789 \end{bmatrix}$. Cette méthode nécessite plus d'itérations pour converger sachant qu'on s'est arrêté à 10000 itérations.
Nous avons utilisé une distance quadratique $d_s = \frac{\|\boldsymbol{S} - \widehat{\boldsymbol{S}}\|^2}{\|\boldsymbol{S}\|^2}$ entre les sources originales et les sources estimées pour mesurer la qualité de l'estimation des sources.

Une deuxième série d'expériences a été faite sur un ensemble de 32 images hyperspectrales générées synthétiquement à l'aide de 3 sources. Cette expérience a montré la performance de la méthode d'approximation en champ moyen par région. Il faux noter que la méthode utilisant les contours échoue au niveau algorithmique et qu'on n'a pas pu montrer de résultats.

Figure 5 montre ces 32 images sous forme d'un cube.

Figure 6 montre le résultat de séparation avec le modèle *a priori* (5) et l'algorithme de Gibbs.

Figure 7 montre le résultat séparation avec le modèle *a priori* (5) et l'approximation en champ moyen.

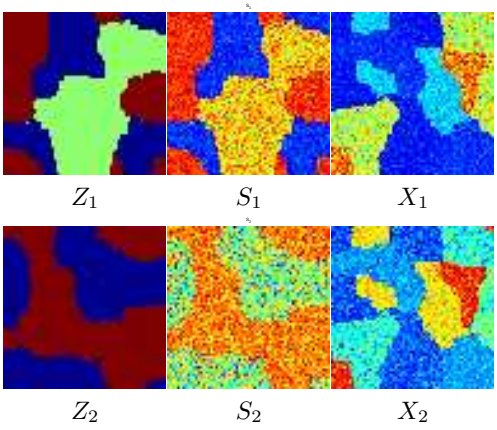


FIG. 1: *Schéma de Simulations des données : les étiquettes Z_1, Z_2 sont générées suivent une distribution de Potts, Sources S_1, S_2 sont générées conditionnellement aux étiquettes associées et les observations X_1, X_2 qui sont obtenues selon ule modèle de mélange 3)*

4 Conclusion et perspectives

Nous proposons une méthode de séparation de sources pour le problème de réduction de données en imagerie hyperspectrale à l'aide de deux modèles markoviens pour les images sources. Le premier modèle est assez courant en traitement d'image alors que le deuxième modèle que nous proposons est nouveau et permet non seulement la séparation mais aussi une segmentation commune des images hyperspectrales.

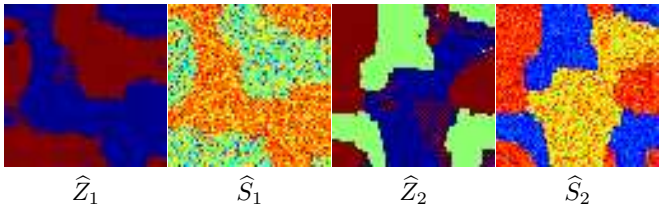


FIG. 2: *Résultat de séparation par la méthode d'approximation en champ moyen avec région.*$\{d_{s1} = .0065, d_{s2} = .0029\}$

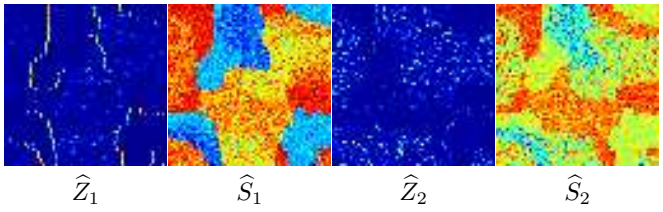


FIG. 3: *Résultats de séparation par la méthode d'approximation en champ moyen avec processus contour.* $\{d_{s1} = .04, d_{s2} = 0.05\}$

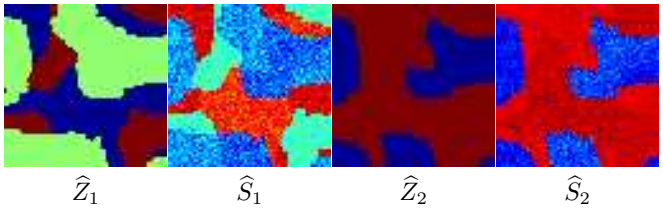


FIG. 4: *Résultats de séparation par la méthode d'échantillonnage de Gibbs avec variables cachées région.* $\{d_{s1} = 12.1221, d_{s2} = .4.6623\}$

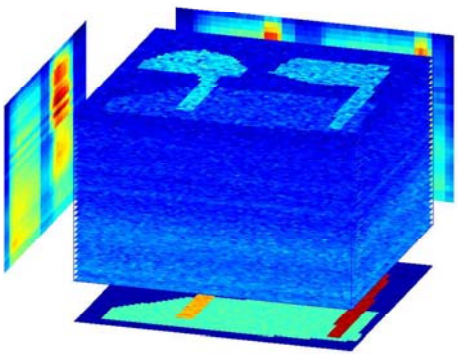


FIG. 5: *Un ensemble de 32 images hyperspectrales simulée artificiellement avec la présence de 3 sources*

Les principales contributions de ce travail sont l'utilisation d'une modélisation intensité région dans un problème de séparation de source et l'approximation en champ moyen dans un problème de séparation de sources utilisant ce modèle.

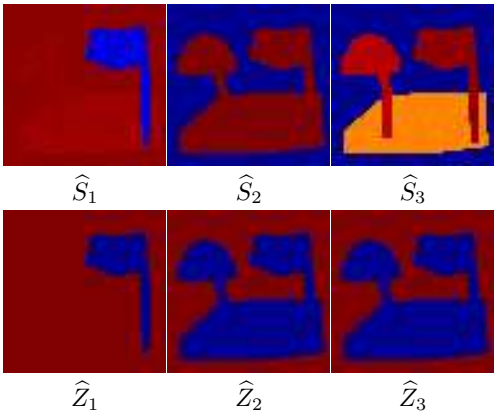


FIG. 6: *Résultats de séparation par la méthode d'échantillonnage de Gibbs*

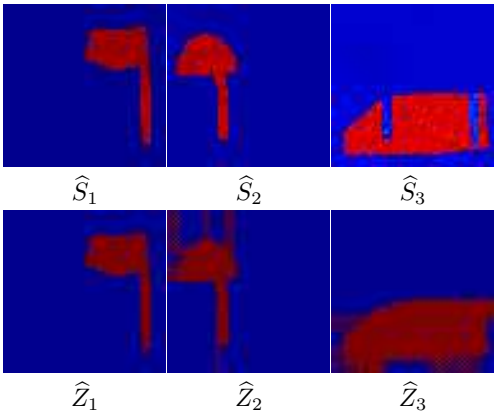


FIG. 7: *Résultats de séparation par la méthode d'approximation en champ moyen avec région*

Références

[1] G. Rellier, *Analyse de textures dans l'espace hyperspectral par des méthodes probabilistes*, Thèse de Doctorat, Université de Nice-Sophia Antipolis, Nice, 2002.

[2] J. Zhang, « The Mean Field Theory in EM Procedures for blind Markov random field image restoration », in *IEEE Transactions on Image Processing*. IEEE, 1993, vol. 2, pp. 27–40.

[3] H. Snoussi et A. Mohammad-Djafari, « Fast joint separation and segmentation of mixed images », *Journal of Electronic Imaging*, vol. 13, n° 2, pp. 349–361, 2004.

[4] A. Tonazzini, L. Bedini et E. Salerno, « Blind Separation of Auto-Correlated images from noisy mixtures using MRF models », in *4th International Symposition on Independent Component Analysis and Blind Signal Separation*, 2003, pp. 675–680.

[5] F. Forbes et N. Peyrard, « Hidden Markov random field model selection criteria based on mean field-like approximations », in *Pattern Analysis and Machine Intelligence*. IEEE Transactions, 2003, vol. 25, pp. 1089–1101.

Solution analytique et généralisation d'un contraste à référence

Atman JBARI[1], Abdellah ADIB[1,2], Driss ABOUTAJDINE[1]

[1]GSCM-Faculté des Sciences de Rabat, Rabat, Maroc

[2]Institut Scientifique, av. Ibn Battouta, B.P. 703, Rabat, Maroc

ajbari2@yahoo.fr, adib@israbat.ac.ma, aboutaj@fsr.ac.ma

Résumé − Cet article traite du problème de la séparation aveugle de sources par la considération d'un système à référence. À partir d'un contraste à référence, on construit une famille de contrastes puis on développe une solution analytique. Les résultats de simulation présentent une comparaison de notre proposition avec le contraste de base et en tenant compte de plusieurs valeurs du paramètre de généralisation.

Abstract − This paper deals with the problem of blind source separation by consideration of a reference system. From a referenced contrast, we construct a family of contrasts then we propose an analytical solution to optimize it. The simulation results present a comparison study between our proposal and the basic contrast in considering several values of the generalization parameter.

1 Introduction

La séparation de sources vise à maximiser l'indépendance statistique des signaux à partir d'un mélange. Les méthodes classiques reposent sur l'utilisation directe des statistiques des signaux. Dans notre article on propose de résoudre le problème à l'aide d'un système à référence [1] [2]. La figure 1 montre la problèmatique étudiée.

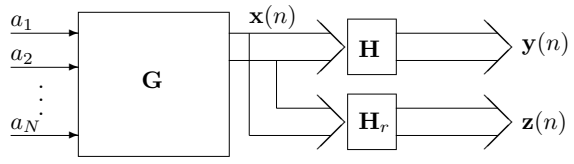


FIG. 1: Système mélange/séparation.

Avec : $\mathbf{a}$: le vecteur source; $\mathbf{x}$: vecteur observaion; $\mathbf{y}$: vecteur estimation; $\mathbf{z}$: vecteur référence; G : matrice de mélange; H : matrice de séparation et H_r : matrice du système à référence (H_rG est unitaire).

La séparation consiste donc à estimer la matrice H tel que : $HG = DP$ avec : D matrice diagonale et P matrice de permutation. Dans [2], un contraste à référence, sous contrainte de blanchiment préalable des observations, est proposé:

$$\mathcal{I}_{R,z}(y) = \sum_{i,j=1}^{N} \mid cum\{y_i, y_i^*, \underbrace{z_j, z_j^* \ldots, z_j^*}_{R\times}\} \mid^2 \qquad (1)$$

Ce contraste peut être exprimé sous la forme d'un critère de diagonalisation conjointe [3]. Notre but est de généraliser ce contraste et de lui proposer une solution analytique.

2 Contraste généralisé

On propose une nouvelle famille de contraste comme combinaison linéaire de celui défini en (1) et des cumulants croisés des observations. On pose:

$$\mathcal{C}^1_{R,z}(\mathbf{y}) = \sum_{i1>i2,j=1}^{N} \mid cum\{y_{i1}, y_{i2}^*, \underbrace{z_j, z_j^* \ldots, z_j^*}_{R\times}\} \mid^2 \qquad (2)$$

$$\mathcal{C}^2_{R,z}(\mathbf{y}) = \sum_{i1<i2,j=1}^{N} \mid cum\{y_{i1}, y_{i2}^*, \underbrace{z_j, z_j^* \ldots, z_j^*}_{R\times}\} \mid^2 \qquad (3)$$

Proposition : 1 *La fonction*

$$\mathcal{J}_{R,z}(\mathbf{y}) = \mathcal{I}_{R,z}(\mathbf{y}) + \alpha_1\mathcal{C}^1_{R,z}(\mathbf{y}) + \alpha_2\mathcal{C}^2_{R,z}(\mathbf{y}) \qquad (4)$$

avec : $\alpha_1 \leq 0$ et $\alpha_2 \leq 0$ est un contraste *pour les signaux normalisés.*

Preuve:

Soit :

$$\sum_{i1,i2,j=1}^{N} \mid cum\{y_{i1}, y_{i2}^*, \underbrace{z_j, z_j^* \ldots, z_j^*}_{R\times}\} \mid^2 =$$
$$\sum_{i,j=1}^{N} \mid cum\{y_i, y_i^*, \underbrace{z_j, z_j^* \ldots, z_j^*}_{R\times}\} \mid^2 + \mathcal{C}^1_{R,z}(\mathbf{y}) \qquad (5)$$
$$+\mathcal{C}^2_{R,z}(\mathbf{y})$$

En pondérant les cumulants croisés par des coefficients $\alpha_1 \leq 0$ et $\alpha_2 \leq 0$ on obtient la fonction $\mathcal{J}_{R,z}(\mathbf{y})$ qui vérifie :

$$\begin{array}{c} \mathcal{J}_{R,z}(\mathbf{y}) \leq \mathcal{I}_{R,z}(\mathbf{y}),\ \mathcal{C}^1_{R,z}(\mathbf{a}) = 0, \\ \mathcal{C}^2_{R,z}(\mathbf{a}) = 0,\ \mathcal{J}_{R,z}(\mathbf{a}) = \mathcal{I}_{R,z}(\mathbf{a}) \end{array} \qquad (6)$$

On en déduit alors que la fonction $\mathcal{J}_{R,z}(\mathbf{y})$ est un contraste [4].

3 Solution analytique

En vue d'une maximisation du contraste (1) on cherche une solution analytique. Pour simplifier, on considère des signaux réels. La matrice H est paramétrée ainsi :

$$H = \begin{pmatrix} c & s \\ -s & c \end{pmatrix} \tag{7}$$

Avec : $c = \cos\theta$ et $c = \sin\theta$ On note :

$$Y_{s,t,u} = cum\{y_s, y_t, z_u, z_u\}, \tag{8}$$

$$X_{s,t,u} = cum\{x_s, x_t, z_u, z_u\} \tag{9}$$

Dans le cas de deux sources, le contraste (1) peut être développé ainsi :

$$\begin{aligned} \mathcal{I}_{2,z}(y) = & \quad Y_{1,1,1}^2 + Y_{1,1,2}^2 + Y_{2,2,1}^2 + Y_{2,2,2}^2 \\ = & \quad (c^4 + s^4)\, F_1 + 2\, c^2\, s^2\, F_2 \\ & +4\, c^2\, s^2 (c^2 - s^2)\, F_3 \end{aligned} \tag{10}$$

Avec :

$$F_1 = X_{1,1,1}^2 + X_{2,2,1}^2 + X_{1,1,2}^2 + X_{2,2,2}^2 \tag{11}$$

$$F_2 = X_{1,2,1}^2 + X_{1,2,2}^2 + \frac{1}{2}(X_{1,1,1}X_{2,2,1} + X_{1,1,2}X_{2,2,2}) \tag{12}$$

$$F_3 = X_{1,2,1}(X_{1,1,1} - X_{2,2,1}) + X_{1,2,2}(X_{1,1,2} - X_{2,2,2}) \tag{13}$$

Par un calcul trigonométrique, on obtient la forme suivante:

$$\mathcal{I}_{2,z}(y) = \Gamma + \Lambda \cos 4\theta + \Omega \sin 4\theta \tag{14}$$

Avec:

$$\Gamma = \frac{3F_1 + 4F_2}{4},\ \Lambda = \frac{F_1 - 4F_2}{4},\ \text{et}\ \Omega = F_3 \tag{15}$$

La valeur de θ est donc l'argument du nombre complexe $t = \Lambda + j\Omega$ avec $(j^2 = -1)$.
Dans le cas réel, le contraste généralisé peut s'écrire :

$$\mathcal{J}_{2,z}(y) = \mathcal{I}_{2,z}(y) + (\alpha_1 + \alpha_2)(Y_{1,2,1}^2 + Y_{1,2,2}^2) \tag{16}$$

En posant : $\alpha = \alpha_1 + \alpha_2$ et en suivant la même démarche de calcul, on obtient :

$$\begin{aligned} \mathcal{J}_{2,z}(y) = & \quad (c^4 + s^4)F_1^{'} + 2c^2 s^2 F_2^{'} + 4c^2 s^2 (c^2 - s^2) \\ & F_3^{'} + 2sc(c^2 - s^2)F_4^{'} \end{aligned} \tag{17}$$

Avec :

$$F_1^{'} = F_1 + \alpha(X_{1,2,1}^2 + X_{1,2,2}^2) \tag{18}$$

$$\begin{aligned} F_2^{'} = & \quad F^2 + \tfrac{\alpha}{2}(X_{1,1,1}^2 + X_{1,1,2}^2 + X_{2,2,1}^2 + X_{2,2,2}^2 \\ & -2(X_{1,1,1}X_{2,2,1} + X_{1,2,1}^2 + X_{2,1,2}^2 \\ & +X_{2,2,2}X_{1,1,2}) \end{aligned} \tag{19}$$

$$F_3^{'} = F_3 \tag{20}$$

$$F_4^{'} = -\alpha F_3 \tag{21}$$

Puis après : $\mathcal{J}_{2,z}(y) = \Gamma^{'} + \Lambda^{'} \cos 4\theta + \Omega^{'} \sin 4\theta$, avec

$$\Gamma^{'} = \frac{3F_1^{'} + 4F_2^{'}}{4},$$

$$\Lambda^{'} = \frac{F_1^{'} - 4F_2^{'}}{4}$$

et

$$\Omega^{'} = \frac{2F_3^{'} + F_4^{'}}{2}.$$

La valeur de θ est donc l'argument du nombre complexe $t^{'} = \Lambda^{'} + j\Omega^{'}$ avec $(j^2 = -1)$.

Dans le cas d'un nombre quelconque de sources, la matrice H est un produit de $\frac{N(N-1)}{2}$ rotations de Givens.

4 Résultats de simulation

Pour évaluer la qualité de séparation on utilise l'indice suivant défini dans [5] appliqué sur la matrice globale $S = HG$:

$$\begin{aligned} PI \quad = & \frac{1}{2(N-1)} \left[\sum_{i=1}^{N} \left(\frac{\sum_{j=1}^{N} |S_{ij}|^2}{\max_k |S_{ik}|^2} - 1 \right) \right. \\ & \left. + \sum_{j=1}^{N} \left(\frac{\sum_{i=1}^{N} |S_{ij}|^2}{\max_k |S_{kj}|^2} - 1 \right) \right] \end{aligned} \tag{22}$$

Cet indice mesure la somme des contributions des éléments de S par rapport au maximum en lignes et en colonnes. Dans le cas d'une séparation parfaite, PI est voisin de 0.

Le vecteur de référence $\mathbf{z}$ est construit à partir des données blanchies des observations. La validation de la solution analytique du contraste de base et de la nouvelle famille de contrastes est effectuée par les simulations suivantes :

1. Influence de la gaussianité (fig 2) : On utilise des sources paramétrées [6] par τ pour avoir un kurtosis négatif, nul ou positif. Chaque signal peut prendre les trois valeurs suivantes $[-1, 0, \tau]$ avec les probabilités respectives suivantes : $\frac{1}{\tau+1}$, $\frac{\tau-1}{\tau}$, $\frac{1}{\tau(\tau+1)}$. Pour chaque valeur de τ, On caclule la moyenne de l'indice PI en dB sur 100 réalisations Monte-Carlo de la matrice de mélange. La séparation est assurée dans le cas sur-gaussien et sous-gaussien.
2. Signaux de communications : signal à module constant et signal QAM-4 (fig 3) : On restitue les signaux sources avec $PI = 0.0023$.
3. Signaux sonores (fig 4) : Les signaux représentent des sons échantillonnés avec une fréquence de 22.05Khz. La séparation est effectuée par le contraste généralisé avec $\alpha = -1$, et on trouve un $PI = 0.0054$.
4. Influence du rapport S/B (fig 5) : La puissance des sources reste constante. On fait varier la puissance du bruit et on évalue la moyenne de PI sur 100 réalisations Monte-Carlo.

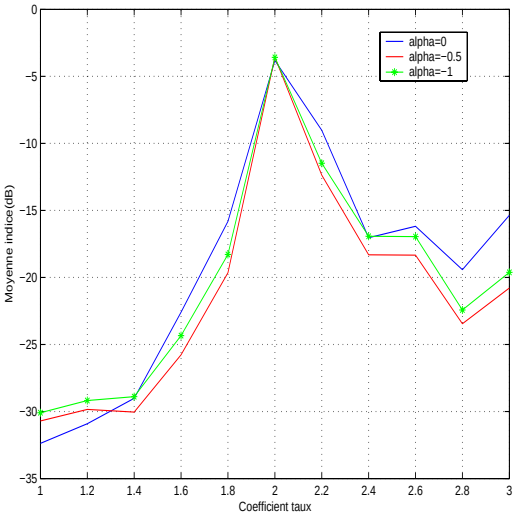


FIG. 2: Influence de la gaussianité.

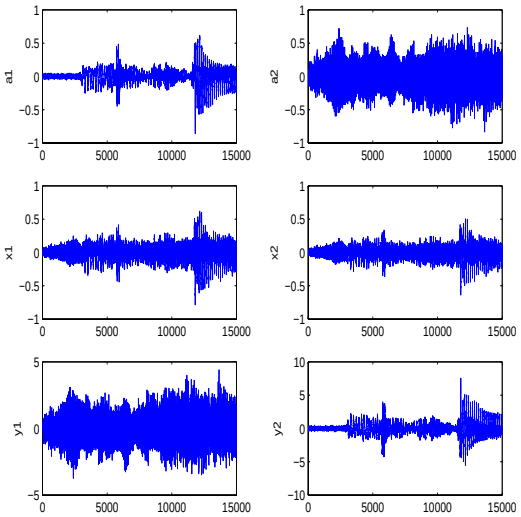


FIG. 4: Cas de mélange sonore.

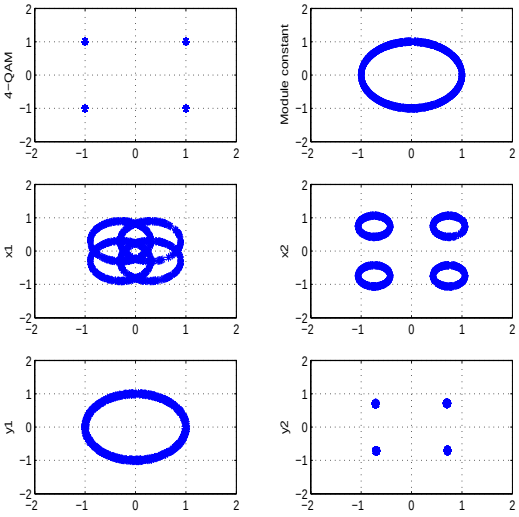


FIG. 3: Cas d'un signal à module constant et signal QAM-4.

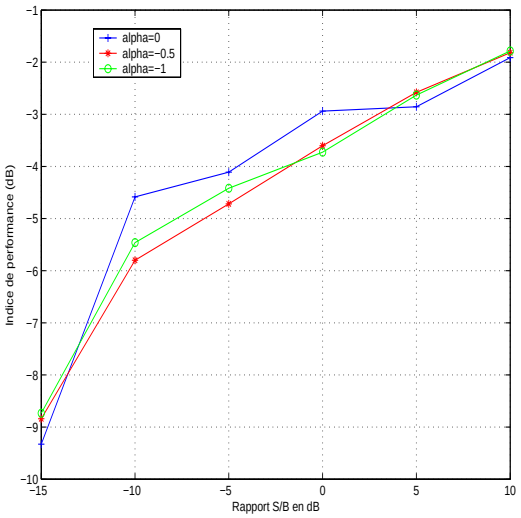


FIG. 5: Influence de la puissance du bruit.

5 Conclusion

Nous avons développé la solution analytique d'un contraste à référence. Ce dernier a été généralisé pour donner une nouvelle famille à deux paramètres dans le cas complexe, ce qui constitue deux degrés de liberté pour améliorer la qualité de séparation. Enfin nous avons testé notre solution par des simulations sur des signaux de natures différentes. L'influence de la puissance du bruit sur les performances a été également étudiée.

References

[1] M. Feng and K. D. Kammeyer. *"Blind source separation for communication siganls using antenna arrays"*. in Proc. ICUPC'98, pp. 665–669, Florence, Oct 1998.

[2] A. Adib, E. Moreau and D. Aboutajdine. *"Blind sources separation by simultaneous generalized referenced contrasts diagonalization"*. Proc. of the Fourth Int. Conf. on Independent Component Analysis (ICA'03), pp. *657–661*, Nara, Japan, April, *2003*.

[3] A. Adib, *"Séparation aveugle de sources. Fonctions de contrastes et systèmes à référence"*, Université Mohammed V-Agdal. Maroc Thèse d'état. 24 Juillet 2004.

[4] E. Moreau, J.-C. Pesquet, and N. Thirion-Moreau, *"An equivalence between non symetrical contrasts and cumulant matchiong for blind signal separation"*, ICA'99, Ausois, France, pp. 301-306, January 1999.

[5] S. Choi and A. Cichocki *"Blind Separation of non stationary sources in noisy mixture"*, Electronics Letters,vol. 36,no. 9, pp. 848–849, April 2000.

[6] E. Moreau *"A generalization of joint diagonalization criteria for source separation"*, IEEE Trans. Signal Processing, March 1999.

Déflation parallèle avec des contrastes APF pour l'extraction aveugle de sources

Vicente Zarzoso[1,*], Ludwig Rota[2], Pierre Comon[2]

[1]Department of Electrical Engineering and Electronics, The University of Liverpool
Brownlow Hill, Liverpool L69 3GJ, UK

[2]Laboratoire I3S, Les Algorithmes – Euclide-B, 2000 route des Lucioles, BP 121
06903 Sophia Antipolis Cedex, France

vicente@liv.ac.uk, rota@i3s.unice.fr, comon@i3s.unice.fr

Résumé – Le présent article aborde le problème de la séparation de sources discrètes, notamment dans le cas sous-déterminé où il existe moins de capteurs que de sources. Les contrastes APF (*alphabet polynomial fitting*) exploitent la connaissance de l'alphabet des modulations numériques pour accomplir l'égalisation spatio-temporelle. Des méthodes d'extraction sont donc susceptibles d'être développées à partir de ces contrastes. Le fait qu'un critère APF vise une modulation particulière permet le fonctionnement en parallèle des algorithmes à déflation pour l'extraction des signaux de chaque modulation. La déflation parallèle permet de réduire la diversité nécessaire à l'extraction de toutes les sources du mélange. Ainsi, cette nouvelle approche serait capable d'améliorer les performances d'estimation et les coûts de calcul par rapport à la déflation classique.

Abstract – The present paper addresses the problem of discrete source separation, particularly in the scenario where there exist less sensors than sources (underdetermined case). Alphabet polynomial fitting (APF) contrasts exploit the prior knowledge of digital modulation alphabets to perform the space-time equalization. Source extraction methods can be developed from such contrasts. The feature that each APF criterion targets a specific modulation enables the parallel operation of deflation algorithms to extract the signals of each modulation in parallel. Parallel deflation allows the reduction of the diversity necessary for the extraction of all sources from the mixture. Hence, this new approach can improve the estimation performance and the computational complexity of classical deflation.

1 Introduction

L'égalisation de canal a pour objectif de reconstruire les signaux transmis qui ont été distordus par le milieu de propagation. L'égalisation aveugle a attiré une importante attention depuis les travaux pionniers de Sato et de Godard [1]. L'avantage principal des techniques aveugles est le fait que les séquences pilote ne sont pas requises, ce qui permet d'augmenter le débit effectif de transmission. Dans les scénarios à entrées et sorties multiples (*multiple-input multiple-output, MIMO*), le mélange spatial de différentes sources transmises s'ajoute à l'interférence entre symboles introduite par le canal dispersif dans le temps. L'extraction aveugle de signaux peut être accomplie à travers une approche à déflation, selon laquelle les sources sont estimées l'une après l'autre [2]. Comme principale limitation de la déflation classique, les erreurs d'estimation s'accumulent au long des étapes d'extraction successives. Par ailleurs, la diversité doit être suffisamment élevée. Le nombre de capteurs doit être typiquement supérieur au nombre de sources pour assurer une égalisation satisfaisante.

La présente contribution porte sur le problème de l'extraction aveugle de signaux discrets, notamment dans le cas sous-determiné où il existe moins de capteurs que de sources. L'originalité de ce travail réside dans l'utilisation d'un critère qui exploite la connaissance a priori de l'alphabet fini de la modulations numérique [3, 4]. Ce critère est basé sur l'ajustement de la sortie de l'égaliseur au polynôme définissant l'alphabet, d'où l'appellation *alphabet polynomial fitting (APF)*. Par opposition aux principes traditionnels indépendants de la distribution des sources tels que le module constant (CM) [1] ou la maximisation du kurtosis (KM) [5], le critère APF vise une modulation spécifique. Cette particularité mène au nouveau concept de la déflation parallèle : un critère polynomial peut être utilisé dans un processus déflationnaire pour extraire les signaux de chaque modulation. La déflation parallèle permettrait ainsi de réduire la diversité requise pour l'extraction de toutes les sources du mélange tout en extrayant les différents modulations simultanément. Par conséquent, cette nouvelle approche serait capable d'augmenter la performance d'extraction en réduisant le coût de calcul par rapport à la déflation classique.

D'autre part, les critères APF sont susceptibles d'être optimisés par des procédures à descente par gradient où à descente Newton basées sur un pas optimal calculé algébriquement à chaque itération. À coût de calcul abordable, la technique à pas optimal est capable d'éviter la plupart des minima locaux de la fonction de contraste.

* Royal Academy of Engineering Research Fellow.

2 Extraction aveugle de sources

2.1 Problème et modèle

Nous considérons un système MIMO linéaire dispersif et invariant dans le temps à réponse impulsionnelle finie (RIF) d'ordre M, avec la relation entrée-sortie suivant :

$$\mathbf{x}(n) = \sum_{k=0}^{M} \mathbf{H}_k \mathbf{s}(n-k) + \mathbf{v}(n), \qquad n \in \mathbb{N} \tag{1}$$

où $\mathbf{s}(n) \in \mathbb{C}^K$, $\mathbf{x}(n) \in \mathbb{C}^P$, $\mathbf{v}(n) \in \mathbb{C}^P$ et $\mathbf{H} = [\mathbf{H}_0, \mathbf{H}_1, \ldots, \mathbf{H}_M]$, $\{\mathbf{H}_k\}_{k=1}^{M} \in \mathbb{C}^{P\times K}$, représentent les sources, la sortie du canal, le bruit, et la réponse impulsionnelle matricielle du canal, respectivement.

Un égaliseur RIF décrit par la réponse impulsionnelle matricielle $\mathbf{F} = [\mathbf{F}_0^{\mathrm{T}}, \mathbf{F}_1^{\mathrm{T}}, \ldots, \mathbf{F}_{N-1}^{\mathrm{T}}]^{\mathrm{T}}$, $\{\mathbf{F}_k\}_{k=0}^{N-1} \in \mathbb{C}^{P\times K}$, de longueur N vise à extraire les sources en traitant les signaux à la sortie du canal. Le signal vectoriel de sortie est donné par

$$\hat{\mathbf{s}}(n) = \sum_{k=0}^{N-1} \mathbf{F}_k^{\mathrm{H}} \mathbf{x}(n-k), \qquad n \in \mathbb{N}$$

où le symbole $^{\mathrm{H}}$ représente l'opérateur complexe conjugué. Alternativement, l'extraction de la p-ème composante de $\hat{\mathbf{s}}(n)$ peut être exprimée de la forme :

$$\hat{s}_p(n) = \mathbf{f}_p^{\mathrm{H}} \mathbf{x}_N(n) \tag{2}$$

avec $\mathbf{x}_N(n) = [\mathbf{x}(n)^{\mathrm{T}}, \mathbf{x}(n-1)^{\mathrm{T}}, \ldots, \mathbf{x}(n-N+1))^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{C}^{PN}$ (le symbole $^{\mathrm{T}}$ désigne l'opérateur transposition) et $\mathbf{f}_p$ indique la p-ème colonne de l'égaliseur matriciel $\mathbf{F}$.

2.2 Déflation classique

La déflation classique vise à extraire les K sources l'une après l'autre. Cette technique peut être employée avec un critère indépendant de la distribution des sources, tel que les principes CM ou le KM ; par exemple, ce dernier [5] est utilisé dans l'article original [2]. Un seul critère est appliqué pour extraire chaque source des observations. Afin d'éviter d'extraire la même source plusieurs fois, la contribution de la source extraite doit être estimée (p.ex., via des techniques de corrélation) et soustraite de la sortie des capteurs. Cette procédure est répétée jusqu'à ce que les K sources aient été obtenues. La diversité requise pour l'extraction des K sources est limitée par un nombre de capteurs de l'ordre de $P \approx K$. Par ailleurs, les erreurs d'estimation s'accumulent avec le nombre d'extractions, donc la qualité de l'extraction diminue graduellement. La déflation classique est illustrée par la Fig. 1.

3 Extraction basée sur l'alphabet

3.1 Critères basés sur l'alphabet

Par la suite, $K = \sum_{i=1}^{R} K_i$ désigne le nombre total de sources emises, où K_i est le nombre de signaux ayant le même alphabet $\mathcal{A}_i$ et R est le nombre d'alphabets différents présents parmi les sources. Cela correspond à l'hypothèse suivante sur les signaux d'entrée :

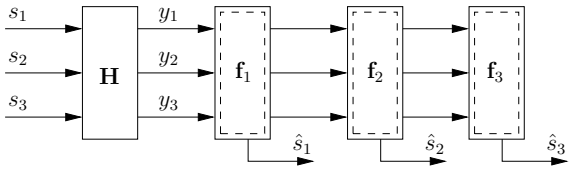


FIG. 1 – *Déflation classique. Extraction de 3 signaux $\{s_p\}_{p=1}^{3}$ ayant typiquement (mais pas forcément) la même modulation. La déflation conventionnelle estime les signaux d'entrée un par un.*

TAB. 1 – *Alphabets et polynômes associés à quelques modulations discrètes.*

Modulation	$\mathcal{A}$	$Q(z)$
BPSK	$\{-1, +1\}$	$z^2 - 1$
q-PSK	$\{\mathrm{e}^{\jmath 2k\pi/q}\}\vert_{k\in 0,\ldots,q-1}$	$z^q - 1$
QAM-16	$\{\pm 1, \pm 3\} + \{\pm\jmath, \pm 3\jmath\}$	$\sum_{k=0}^{4} \alpha_k z^{4k}$

$\alpha_0 = 50625/256$,$\alpha_1 = 12529/16$,$\alpha_2 = -221/8$, $\alpha_3 = 17$,$\alpha_4 = 1$.

H1. Les sources $\mathbf{s}^{(i)} = [s_1^{(i)}, \ldots, s_{K_i}^{(i)}]^{\mathrm{T}}$ appartiennent à un alphabet fini $\mathcal{A}_i$, caractérisé par les d_i racines complexes du polynôme $Q_i(z) = 0$, où d_i correspond au nombre total de symboles possibles dans la constellation.

Cette hypothèse est essentielle pour l'application des critères basés sur l'alphabet fini. Par exemple, un signal à modulation q-PSK est caractérisé par le polynôme $Q(z) = z^q - 1$. Ainsi, chaque modulation discrète peut être associée à un critère APF, comme illustré par les exemples de la Table 1.

L'hypothèse **H1** sur les entrées discrètes du canal MIMO permet de réaliser l'extraction des sources en minimisant le critère polynomial suivant [3] :

Theorem 1 *Soit $\mathcal{S}_i$ l'ensemble de processus stochastiques prenant des valeurs dans l'alphabet $\mathcal{A}_i$, et $\mathcal{F}$ l'ensemble de filtres RIF. Alors, le critère :*

$$\mathcal{J}_{\mathrm{APF}}^{(i)}(\mathbf{F}^{(i)}) = \sum_{p=1}^{K_i} \mathrm{E}\{\vert Q_i(\hat{s}_p)\vert^2\}, \qquad \mathbf{F}^{(i)} \in \mathcal{F} \tag{3}$$

*est une fonction de contraste sous l'hypothèse **H1**, où $\mathbf{F}^{(i)} = [\mathbf{f}_1^{(i)}, \mathbf{f}_2^{(i)}, \ldots, \mathbf{f}_{K_i}^{(i)}]$ contient les filtres égaliseurs (colonnes de la matrice $\mathbf{F}$) visant à extraire les composantes $\{s_p^{(i)}\}_{p=1}^{K_i} \in \mathcal{S}_i$ des sources $\mathbf{s}$ à partir du mélange observé dans l'équation* (1).

Un critère APF peut être utilisé pour effectuer la déflation classique lorsque tous les signaux émis ont le même alphabet, c'est à dire, $K = K_1$ et $K_i = 0$, $\forall i > 1$. Cependant, de nouvelles approches sont possibles grâce au caractère discriminatoire du critère APF, plus fort que celui des principes traditionnels comme le CM et le KM. Les nouvelles approches consistent à extraire les sources avec des alphabets différents en parallèle, d'où les termes de « extraction parallèle » et « déflation parallèle », qui sont présentés par la suite.

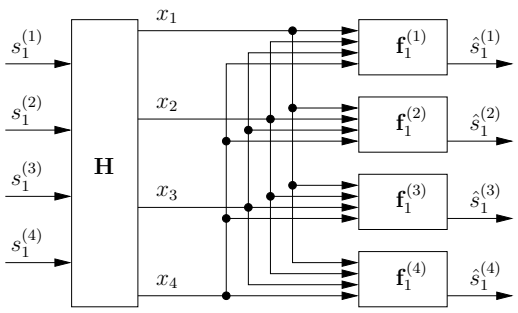


FIG. 2 – *Extraction parallèle. À partir des signaux observés à la sortie des capteurs, l'extraction parallèle permet la séparation simultanée des sources ayant des modulations différentes.*

3.2 Extraction parallèle

L'extraction parallèle peut être effectuée lorsque les K sources emises ont des modulations différentes, c'est à dire, $K_i = 1, \forall i$. Chaque égaliseur est calculé à partir du critère APF correspondant à l'alphabet $\mathcal{A}_i$. Ainsi, les égaliseurs de chaque modulation sont susceptibles d'être déterminés en parallèle à partir de la même sortie des capteurs. La Fig. 2 montre un exemple de l'extraction parallèle des signaux $\{s_1^{(i)}\}_{i=1}^4$ avec des alphabets respectifs $\{\mathcal{A}_i\}_{i=1}^4$. L'extraction parallèle peut être considérée comme un cas particulier du concept plus général de la déflation parallèle.

3.3 Déflation parallèle

Dans le cas général, les capteurs reçoivent un mélange de R groupes de sources, où le i-ème groupe est composé de K_i signaux avec la même modulation $\mathcal{A}_i$. Il est donc possible d'extraire les sources du même groupe à travers un algorithme à déflation avec le critère adapté à la modulation correspondante. Pour les autres groupes ayant des modulations différentes et donc leurs propres critères APF, le processus de déflation peut être effectué en parallèle pourvu que les alphabets ne soient pas sous-ensembles les uns des autres. Par conséquent, la propriété discriminatoire des critère APF est capable de découpler un problème de séparation de K signaux en R problèmes d'extraction de K_i sources à modulation $\mathcal{A}_i$, $i = 1, \ldots, R$. Par opposition à la déflation classique, la diversité requise par la déflation parallèle est réduite à $P \approx \max(K_i)$. Cette amélioration de diversité offre d'autres avantages en termes de performance (notamment une accumulation d'erreur plus basse) et de coût de calcul par rapport à la déflation classique. La déflation parallèle se réduit à l'extraction parallèle lorsque $R = K$, où la déflation n'est plus nécessaire.

4 Optimisation des critères APF

Afin d'estimer la j-ème source à l'alphabet $\mathcal{A}_i$, la fonction de contraste (3) doit être minimisée par rapport à un vecteur égaliseur $\mathbf{f}$ utilisé pour extraire une seule composante, comme décrit par l'équation (2). Après convergence,

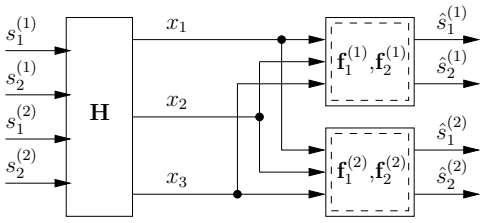


FIG. 3 – *Déflation parallèle dans le cas sous-déterminé. L'extraction de plus de sources que de capteurs est rendue possible par la déflation parallèle, pourvu qu'il y ait la diversité suffisante pour extraire les sources de chaque alphabet.*

le vecteur $\mathbf{f}$ correspondra à la colonne $\mathbf{f}_j^{(i)}$ de l'égaliseur matriciel $\mathbf{F}$. Après une initialisation adéquate (p.ex., via le filtre à coefficient central unité), le vecteur égaliseur est actualisé itérativement dans la direction du gradient $\mathbf{g}$:

$$\mathbf{f}' = \mathbf{f} - \mu\mathbf{g}$$

où μ est le pas de l'itération. Dans un algorithme à descente par gradient, $\mathbf{g} = \nabla\mathcal{J}_{\text{APF}}^{(i)}(\mathbf{f})$, tandis qu'un algorithme à descente Newton implique aussi le Hessien de $\mathcal{J}_{\text{APF}}^{(i)}$.

L'une des caractéristiques les plus importantes des critères APF est que la fonction $\mathcal{J}_{\text{APF}}^{(i)}(\mathbf{f}')$ est un polynôme de degré $2q$ dans le pas μ, pour des constellations constituées par q symboles. Cette caractéristique n'est pas exclusive des contrastes APF, mais est aussi partagée par d'autres critères d'égalisation tels que le CM et le KM [3]. Ainsi, la minimisation à la plus forte descente du contraste (3) peut être accomplie en trouvant le pas optimal

$$\mu_{\text{opt}} = \min_{\mu} \arg\, \mathcal{J}_{\text{APF}}^{(i)}(\mathbf{f} - \mu\mathbf{g})$$

parmi les racines de $\partial\mathcal{J}_{\text{APF}}^{(i)}(\mathbf{f} - \mu\mathbf{g})/\partial\mu$, un polynôme de degré $(2q-1)$. Dans certains cas, la recherche des racines peut être réalisée algébriquement : le crìtère APF adapté aux signaux BPSK et le critère CM sont associés à des polynômes respectifs de degré 3, dont la résolution est donnée par la formule de Cardano ; le critère KM normalisé implique un polynôme de degré 4 dont les racines sont obtenues par la formule de Ferrari. Les coefficients de ces polynômes sont fonctions polynomiales simples des données observées ainsi que des vecteurs égaliseur et gradient de la dernière itération [4, 6]. Par conséquent, l'incorporation de la technique à pas optimal n'entraîne qu'une augmentation modérée du coût de calcul. En échange, puisque μ_{opt} fournit le minimum global de $\mathcal{J}_{\text{APF}}^{(i)}$ dans la direction $\mathbf{g}$, la technique à pas optimal présente une robustesse améliorée contre les minima locaux en relation à la minimisation conventionnelle à descente par gradient [7].

Après convergence du vecteur égaliseur, la contribution de la source estimée aux observations est calculée et soustraite de la sortie des capteurs afin d'éviter d'extraire la même source de nouveau, comme dans la déflation classique. Cette contribution est obtenue facilement à partir de la corrélation croisée entre la source estimée et la sortie des capteurs. Pour extraire la source suivante, le critère

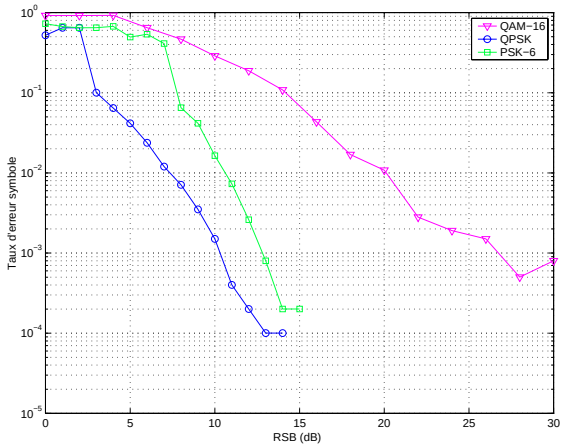


FIG. 4 – *Extraction parallèle de 3 sources à modulation différente pour divers SNRs.*

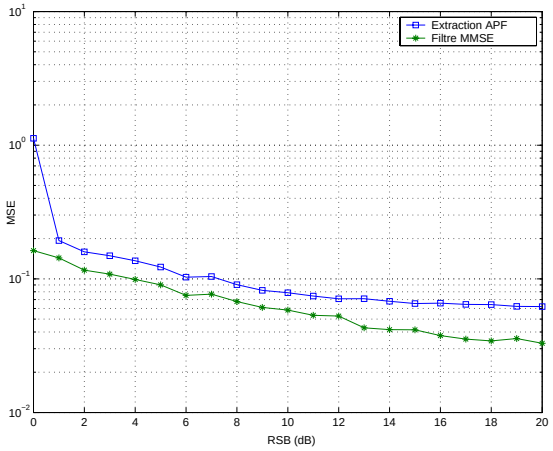


FIG. 5 – *Extraction APF d'un signal QPSK à partir d'un mélange sous-déterminé.*

APF est à minimiser de nouveau, mais en utilisant la sortie des capteurs sans la contribution de la source estimée précédemment. Ce processus est répété jusqu'à ce que tous les sources à la même modulation aient ête obtenues. Dans la déflation parallèle, les algorithmes de déflation des critères APF différents peuvent être exécutés en parallèle.

5 Résultats préliminaires

Extraction parallèle Dans une première expérience, $K = 3$ sources à des modulations différentes (QAM-16, QPSK, PSK-6) sont mélangées par un canal d'ordre $M = 2$. Les signaux observées à la sortie de $P = 3$ capteurs bruités sont traités par un algorithme d'extraction parallèle constitué par les critères APF associés à chaque modulation et par des filtres égaliseurs de longueur $N = 14$. Les coefficients du canal ainsi que le bruit additif sont tirés aléatoirement selon des distributions gaussiennes à chaque itération de Monte Carlo. La Fig. 4 résume la performance de l'extraction parallèle en termes de taux d'erreur symbole pour divers rapports signal à bruit (RSB).

Déflation parallèle La deuxième expérience teste un canal introduisant une dispersion temporelle d'une période de symbole ($M = 1$) et mélangant $K = 4$ sources ($R = 2$, avec 2 signaux QPSK et 2 signaux QAM-16) à la sortie de $P = 3$ capteurs :

$$\mathbf{H} = [\mathbf{H}_0, \mathbf{H}_1]$$

avec

$$\mathbf{H}_0 = \begin{bmatrix} -0.66 & -0.19 & 0.65 & 0.92 \\ 0.22 & -0.96 & 0.43 & -0.85 \\ -0.30 & -0.76 & 0.95 & 0.85 \end{bmatrix}$$

$$\mathbf{H}_1 = \begin{bmatrix} 0.75 & -0.98 & -0.75 & -0.38 \\ -0.97 & 0.27 & 0.90 & 0.53 \\ 0.95 & 0.65 & 0.30 & -0.52 \end{bmatrix}$$

Cette situation décrit un mélange sous-déterminé. L'extraction d'une des sources QPSK à partir d'un bloc de 600 symboles est illustrée par la Fig. 5. Il faut remarquer que, malgré la difficulté du scénario sous-déterminé, la performance de l'extraction APF reste très proche de la borne de l'erreur quadratique moyenne minimale (*minimum mean square error, MMSE*).

6 Conclusions

L'utilisation de fonctions de contraste adaptées à la modulation du signal permet la définition d'une nouvelle approche pour l'extraction aveugle de sources. Cette approche, qui a été introduite dans la présente contribution, est capable d'extraire en parallèle les sources à modulation différente, pourvu qu'aucun alphabet ne soit sous-ensemble d'un autre. La déflation parallèle peut s'avérer particulièrement utile lorsque des modulations différentes cohabitent dans le même environnement de transmission. Un tel scénario est vraisemblable dans des réseaux de commmunications sans fil de future génération, où des modulations seront attribuées dynamiquement en fonction du service requis et des conditions du canal, similairement aux techniques de chargement de bits utilisées dans les communications multi-porteuses. Les expériences préliminaries rapportées dans cet article sont encourageantes. Des résultats expérimentaux plus détaillés illustrant la performance de l'approche à déflation parallèle seront présentés à la conférence.

Références

[1] D. N. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Trans. on Communications*, vol. 28, no. 11, pp. 1867–1875, Nov. 1980.

[2] N. Delfosse and P. Loubaton, "Adaptive blind separation of independent sources: a deflation approach," *Signal Processing*, vol. 45, no. 1, pp. 59–83, July 1995.

[3] P. Comon, "Contrasts, independent component analysis, and blind deconvolution," *International Journal of Adaptive Control and Signal Processing (Special Issue on Blind Signal Separation)*, vol. 18, no. 3, pp. 225–243, Apr. 2004.

[4] L. Rota and P. Comon, "Blind equalizers based on polynomial criteria," in *Proc. ICASSP-2004*, Montreal, Canada, May 17–21, 2004, vol. IV, pp. 441–444.

[5] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems (channels)," *IEEE Transactions on Information Theory*, vol. 36, no. 2, pp. 312–321, Mar. 1990.

[6] V. Zarzoso and P. Comon, "Blind channel equalization with algebraic optimal step size," in *EUSIPCO-2005*, Antalya, Turkey, Sept. 4–8, 2005, à paraître.

[7] V. Zarzoso and P. Comon, "Optimal step-size constant modulus algorithm," *IEEE Trans. on Comms.*, Oct. 2004, soumis.

Méthodes séquentielles de séparation aveugle de sources en présence de bruit

Marc CASTELLA[1], Jean-Christophe PESQUET[2]

[1]Institut National des Télécommunications (INT)
Département CITI (Communications, Images et Traitement de l'Information)
9 rue Charles Fourier 91011 Évry Cedex, France

[2]Institut Gaspard Monge, CNRS - UMR 8049 et Université de Marne-la-Vallée
Equipe Signal et Communications
5, Boulevard Descartes, Champs sur Marne 77454 Marne la Vallée CEDEX 2, France
marc.castella@int-evry.fr, jean-christophe.pesquet@univ-mlv.fr

Résumé – Le problème traité est celui de la séparation aveugle d'un mélange convolutif multi-entrées/multi-sorties de sources mutuellement indépendantes. Le cadre général considéré est celui de sources non nécessairement i.i.d., l'accent étant mis sur les méthodes séquentielles de séparation.

Plus précisément, ce travail s'intéresse à la prise en compte, dans le modèle de mélange et dans les critères de séparation, de la présence d'un bruit additif sur les capteurs. L'étude que nous réalisons justifie et illustre le fait qu'une simple soustraction des cumulants du bruit, ici supposés connus, permet d'améliorer la qualité d'estimation du système séparant.

Abstract – The topic which is addressed here concerns the blind separation of a multi-input/multi-output convolutive mixture of mutually independent sources. The considered framework includes the case of non necessarily i.i.d. sources, emphasis being put on iterative source separation methods (*i.e.* the sources are extracted one by one).

More precisely, this work shows how to take into account the presence of an additive noise, both in the model and in the design of separation criteria. Our study justifies and illustrates that the estimation of the inverse of the mixing system can be improved by simply subtracting the cumulants of the noise.

1 Introduction

Depuis plusieurs années, la séparation de sources est un sujet de recherche particulièrement actif. Cet engouement peut s'expliquer par le large spectre d'applications envisageables ; on peut notamment citer les télécommunications, l'acoustique, la sismique, le domaine biomédical, ... Le cadre aveugle est ici considéré, lequel offre l'avantage de ne requérir que des hypothèses faibles sur le mélange : en dehors de sa structure générale, souvent supposée linéaire, aucun paramètre de ce dernier n'est supposé connu.

Compte tenu de l'intérêt applicatif, on peut s'étonner que le cas de données bruitées n'ait donné lieu qu'à des contributions peu nombreuses mais récentes [3, 10], témoignant d'une attention nouvelle et accrue dédiée à ce cas. Nous souhaitons ici faire le point sur la situation en mettant l'accent sur les mélanges convolutifs multi-entrées / multi-sorties altérés par un bruit additif sur les capteurs.

Depuis plusieurs années, des méthodes efficaces existent pour la séparation des mélanges linéaires instantanés de sources [5, 2]. Le cas des mélanges convolutifs demeure en revanche plus difficile. Les directions principalement explorées ont été les suivantes. Généralisant les contrastes introduits pour les mélanges instantanés, des méthodes de séparation conjointes ont été proposées [6, 9]. Leur inconvénient réside à la fois dans l'existence de maxima locaux parasites et dans la nécessité d'un préblanchiment temporel qui est délicat. D'autre part, des approches fréquentielles ont également prouvé leur validité dans le cas des sources i.i.d. [1], malgré certaines difficultés d'estimation statistique. Enfin, des approches séquentielles de reconstruction des sources ont été proposées. La plus populaire d'entre elles, la déflation [11, 13], consiste à extraire les sources une par une ; après chaque reconstruction, la contribution sur les observations de la source récupérée est soustraite par une méthode de moindres carrés. Le procédé est répété jusqu'à l'extraction de la dernière source Il existe par ailleurs des méthodes séquentielles qu'on peut qualifier de hiérarchiques [12] où les sources précédemment extraites sont utilisées afin de contraindre l'espace de recherche des sources suivantes [8, 4].

Nous souhaitons, dans le cas bruité, étudier plus particulièrement les méthodes séquentielles de séparation, dont l'efficacité est reconnue quand le bruit est négligé. En contrepartie, ces techniques présentent des phénomènes d'accumulation d'erreurs au fur et à mesure des extractions, ce qui fournit une motivation supplémentaire à l'étude de l'influence du bruit.

2 Formulation du problème

2.1 Modèle

Considérons un signal vectoriel à temps discret $(\mathbf{s}(n))_{n\in\mathbb{Z}}$, dont les composantes $(s_1(n))_{n\in\mathbb{Z}}, \ldots, (s_N(n))_{n\in\mathbb{Z}}$ sont appe-

lées signaux sources. Le modèle d'un mélange convolutif de ces N sources avec ajout d'un bruit additif sur un réseau de Q capteurs s'écrit :

$$\mathbf{x}(n) = \sum_{k\in\mathbb{Z}} \mathbf{M}(k)\mathbf{s}(n-k) + \mathbf{b}(n) \triangleq \mathbf{M}[z]\mathbf{s}(n) + \mathbf{b}(n) \quad (1)$$

où $(\mathbf{b}(n))_{n\in\mathbb{Z}} \triangleq ((b_1(n))_{n\in\mathbb{Z}}, \ldots, (b_Q(n)_{n\in\mathbb{Z}}))^T$ correspond à un bruit additif et $(\mathbf{M}(k))_{k\in\mathbb{Z}}$ est la réponse impulsionnelle d'un filtre multivarié appelé filtre de mélange. Conformément à l'équation (1), nous désignons par $\mathbf{M}[z] = \sum_{n\in\mathbb{Z}} \mathbf{M}(n)z^{-n}$ la transformée en z d'un filtre (ici, filtre multivarié de mélange) et par $\mathbf{M}[z]\mathbf{s}(n)$ la sortie du filtre excité par le signal $(\mathbf{s}(n))_{n\in\mathbb{Z}}$.

Le but de la séparation de source consiste en la détermination d'un filtre séparateur $\mathbf{W}[z]$ dont la sortie

$$\mathbf{y}(n) \triangleq \mathbf{W}[z]\mathbf{x}(n) \triangleq \sum_{k\in\mathbb{Z}} \mathbf{W}(k)\mathbf{x}(n-k) \quad (2)$$

reconstitue au mieux les sources. Dans le cadre aveugle dans lesquel nous nous plaçons, les sources sont *inobservables* et aucune connaissance supplémentaire n'est disponible sur le filtre de mélange. Les hypothèses suivantes sont supposées satisfaites :

- A.1 Stationnarité des sources et non nullité de leurs auto-cumulants d'ordre quatre. La fonction d'autocorrélation de chacune des sources est définie positive et notée respectivement $(\gamma_i(k))_{k\in\mathbb{Z}}, i \in \{1, \ldots, N\}$.
- A.2 Indépendance mutuelle des sources $(s_1(n))_{n\in\mathbb{Z}}$, …, $(s_N(n))_{n\in\mathbb{Z}}$.
- A.3 Inversibilité à gauche du filtre $\mathbf{M}[z]$ (ce qui implique en particulier $Q \geq N$).

Dans le cas non bruité, les hypothèses A.1 à A.3 garantissent qu'il est possible de reconstituer les sources à une permutation et un filtrage scalaire près, càd qu'il est possible de déterminer $\mathbf{W}[z]$ tel que le filtre global $\mathbf{G}[z] \triangleq \mathbf{W}[z]\mathbf{M}[z]$ s'écrive $\mathbf{P}\mathrm{diag}(h_1[z], \ldots, h_N[z])$, où $\mathbf{P}$ est une matrice de permutation et les $h_i[z], i \in \{1, \ldots, N\}$ sont des filtres scalaires. Ces derniers se réduisent à des retards dans le cas de sources temporellement indépendantes et identiquement distribuées (i.i.d.).

2.2 Hypothèses sur le bruit

Les méthodes de séparation que nous abordons sont basées sur des cumulants d'ordre supérieur et l'hypothèse gaussienne est souvent prétextée pour éliminer ou négliger l'influence du bruit : il convient cependant de ne pas oublier que, même dans ce cas favorable, les statistiques d'ordre deux interviennent sous forme de contraintes et peuvent être perturbées par le bruit. Nous supposerons ici :

- A.4 Le bruit est stationnaire, centré et indépendant des sources. De plus, ses statistiques d'ordre 2 et 4 sont connues.

2.3 But et notations

Compte tenu de la présence de bruit, une reconstruction parfaite des sources est impossible malgré l'hypothèse A.3. Aussi, nous nous concentrons sur le problème de l'estimation d'un inverse $\mathbf{W}[z]$ du système de mélange.

Les méthodes séquentielles de séparation que nous considérons extraient les sources une par une, càd qu'elles déterminent successivement les sorties globales $(y_i(n))_{n\in\mathbb{Z}}$ pour i variant de 1 à N. Ceci revient à chercher successivement les lignes correspondantes du filtre séparateur, ou encore les lignes du filtre global. Pour alléger les notations, à l'étape i, nous notons simplement $\mathbf{g}[z]$ et $\mathbf{w}[z]$ les $i^{\text{ème}}$ lignes en cours de recherche des filtres $\mathbf{G}[z]$ et $\mathbf{W}[z]$. Le signal scalaire $y(n)$ est également la $i^{\text{ème}}$ sortie correspondante. Par ailleurs, nous aurons besoin, pour tout filtre ligne $\mathbf{g}[z]$ de définir une norme par :

$$\|\mathbf{g}\| \triangleq \Big(\sum_{j=1}^{N} \|g_j\|_j^2\Big)^{\frac{1}{2}} \quad (3)$$

où la norme j sur la $j^{\text{ème}}$ composante est donnée par :

$$\|g_j\|_j \triangleq \Big(\sum_{(k_1,k_2)\in\mathbb{Z}^2} g_j(k_1)g_j^*(k_2)\gamma_j(k_2-k_2)\Big)^{\frac{1}{2}}. \quad (4)$$

3 Extraction d'une source

Nous considérons d'abord l'obtention de la sortie $y(n)$ du séparateur, par exemple lors de la première étape d'extraction. Dans le cas non bruité, il a été démontré que le critère

$$\mathbf{g}[z] \mapsto \left|\frac{\mathrm{C}^4\{y(n)\}}{\mathrm{E}\{|y(n)|^2\}^2}\right|^2 \quad (5)$$

où $\mathrm{C}^4\{.\}$ désigne l'auto-cumulant d'ordre quatre, est une fonction de contraste [11]. Plus précisément, sa maximisation mène à un filtre ligne global $\mathbf{g}[z]$ n'ayant qu'une unique composante non identiquement nulle.

Des résultats théoriques similaires s'appliquent dans le cas bruité à condition d'introduire les termes correctifs nécessaires. En effet, sous les hypothèses A.1 à A.3, la maximisation du critère :

$$\mathbf{g}[z] \mapsto \left|\frac{\mathrm{C}^4\{\mathbf{g}[z]\mathbf{s}(n)\}}{\|\mathbf{g}\|^4}\right|^2 \quad (6)$$

fournit un filtre n'ayant qu'une composante non nulle. Or, en utilisant l'hypothèse A.4, un calcul immédiat donne :

$$\mathrm{E}\{|y(n)|^2\} = \|\mathbf{g}\|^2 + \mathrm{E}\{|\mathbf{w}[z]\mathbf{b}(n)|^2\} \quad (7)$$

$$\mathrm{C}^4\{y(n)\} = \mathrm{C}^4\{\mathbf{g}[z]\mathbf{s}(n)\} + \mathrm{C}^4\{\mathbf{w}[z]\mathbf{b}(n)\}. \quad (8)$$

D'autre part, en écrivant (7) et (8) sous la forme

$$\|\mathbf{g}\|^2 = \sum_{j_1,j_2=1}^{Q} \sum_{(k_1,k_2)\in\mathbb{Z}^2} w_{j_1}(k_1)w_{j_2}^*(k_2)$$
$$\Big(\mathrm{E}\{x_{j_1}(n-k_1)x_{j_2}^*(n-k_2)\} - \mathrm{E}\{b_{j_1}(n-k_1)b_{j_2}^*(n-k_2)\}\Big) \quad (9)$$

$$\mathrm{C}^4\{\mathbf{g}[z]\mathbf{s}(n)\} =$$
$$\sum_{j_1,j_2,j_3,j_4=1}^{Q} \sum_{(k_1,k_2,k_3,k_4)\in\mathbb{Z}^4} w_{j_1}(k_1)w_{j_2}^*(k_2)w_{j_3}(k_3)w_{j_4}^*(k_4)$$
$$\Big(\mathrm{Cum}\{x_{j_1}(n-k_1), x_{j_2}^*(n-k_2), x_{j_3}(n-k_3), x_{j_4}^*(n-k_4)\}$$
$$- \mathrm{Cum}\{b_{j_1}(n-k_1), b_{j_2}^*(n-k_2), b_{j_3}(n-k_3), b_{j_4}^*(n-k_4)\}\Big) \quad (10)$$

on constate que la solution proposée revient à soustraire la covariance (resp. quadricovariance) du bruit de la matrice de covariance (resp. quadricovariance) du vecteur des observations $(\mathbf{x}(n))_{n\in\mathbb{Z}}$. Evidemment, pour un bruit gaussien, seule la correction (9) est nécessaire.

TAB. 1 – Valeur médiane sur 1000 réalisations du critère ISI_1 (mélange de 3 sources sur 4 capteurs, filtre de longeur 3, 10000 échantillons).

σ_b	0.6	0.5	0.4	0.3	0.2	0.1
ISI_1, bruit «soustrait»	0.0088	0.0074	0.0057	0.0050	0.0050	0.0047
ISI_1, bruit ignoré	0.0242	0.0193	0.0144	0.0105	0.0076	0.0051

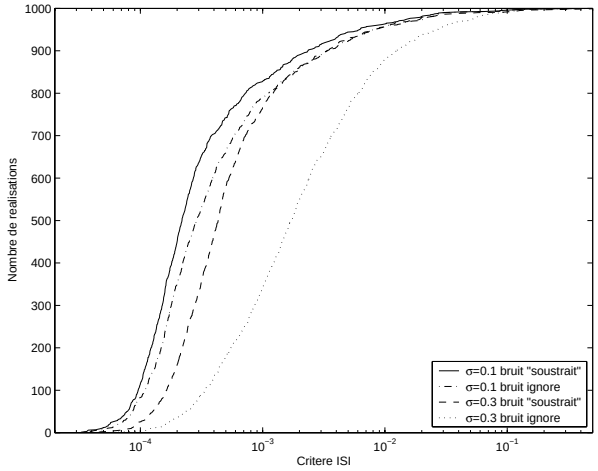


FIG. 1 – Répartition des valeurs du critère ISI_1 obtenues sur 1000 réalisations (mélange de 2 sources sur 4 capteurs, filtre de longueur 3, 10000 échantillons).

4 Extraction séquentielle des sources

4.1 Déflation

L'extraction séquentielle des sources se fonde explicitement sur l'exploitation de statistiques d'ordre deux. Ainsi, l'approche par déflation utilise la sortie $(y(n))_{n\in\mathbb{Z}}$ du séparateur obtenue à l'étape précédente pour soustraire la contribution de la source correspondante dans le mélange. Plus précisément, il s'agit de déterminer par moindres carrés un filtre ajusté $\mathbf{t}[z] = (t_1[z], \ldots, t_Q[z])^T$ de taille $Q \times 1$. Sous réserve que l'étape précédente ait permis la détermination d'un filtre ligne $\mathbf{w}[z]$ extrayant exactement une source, par exemple la première, on a $\mathbf{w}[z]\mathbf{M}[z] = (h_1[z], 0, \ldots, 0)$). Il est alors possible d'écrire, en utilisant (1) ainsi que l'indépendance mutuelle des sources et celle du bruit :

$$\begin{aligned}&\mathrm{E}\{|\big(x_i(n) - b_i(n)\big) - t_i[z]\mathbf{w}[z]\big(\mathbf{x}(n) - \mathbf{b}(n)\big)|^2\}\\ =&\mathrm{E}\{|\big(M_{i1}[z] - t_i[z]h_1[z]\big)s_1(n)|^2\} + \sum_{j=2}^{N}\mathrm{E}\{|M_{ij}[z]s_j(n)|^2\}.\end{aligned} \tag{11}$$

L'annulation de la contribution de la première source sur les capteur est obtenue à condition que, pour tout $i \in \{1, \ldots, Q\}$, le filtre $M_{i1}[z] - t_i[z]h_1[z]$ soit identiquement nul. Par conséquent, le filtre ajusté $\mathbf{t}[z]$ peut être obtenu comme résultant de la minimisation du membre de droite de l'équation (11). Dans le cas non bruité, ceci se réduit à la minimisation de $\mathrm{E}\{|x_i(n) - t_i[z]y(n)|^2\}$. En présence de bruit, le critère d'optimisation reste du même type, mais où la corrélation du bruit doit être soustraite de celle des observations. Ceci revient à minimiser, pour tout $i \in \{1, \ldots, Q\}$, le critère quadratique :

$$\begin{aligned}&\mathrm{E}\{|x_i(n) - t_i[z]y(n)|^2\} - \mathrm{E}\{|t_i[z]\mathbf{w}[z]\mathbf{b}(n)|^2\}\\ &\qquad + 2\Re\{\mathrm{E}\{b_i^*(n)t_i[z]\mathbf{w}[z]\mathbf{b}(n)\}\}.\end{aligned} \tag{12}$$

Enfin, il est important de noter que le retour de prédiction de l'étape de déflation introduit un filtrage du bruit sur les capteurs par le filtre $\mathbf{I} - \mathbf{t}[z]\mathbf{w}[z]$. Dans la pratique, il convient, à chaque étape, de prendre en compte les modifications des caractéristiques statistiques du bruit que cela entraîne.

TAB. 2 – Valeur médiane sur 500 réalisations du critère ISI obtenu par une méthode de déflation (3 sources, 4 capteurs, 10000 échantillons, $\sigma_b = 0.3$).

	$\mathbf{b}(n)$ "soustrait"	$\mathbf{b}(n)$ ignoré
ISI_1	0.0058	0.0115
ISI_2	0.0120	0.0164
ISI_3	0.0196	0.0210

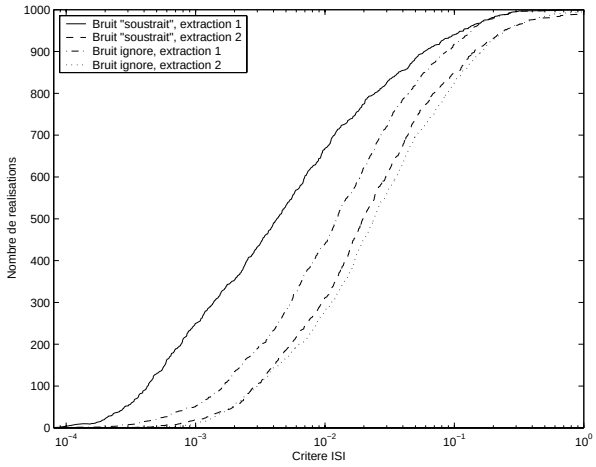


FIG. 2 – Répartition des valeurs des critères ISI ($i = 1, 2$) obtenues sur 1000 réalisations (mélange de 2 sources sur 3 capteurs, filtre de longueur 3, 10000 échantillons, $\sigma_b = 0.3$).

4.2 Méthode hiérarchique par décorrélation

Les résultats précédents s'adaptent aux approches hiérarchiques qui visent la décorrélation spatiale des sorties du séparateur. Il a été démontré [4] qu'il est possible, au lieu d'effectuer une soustraction de déflation, de contraindre la recherche des lignes successives du séparateur de sorte que les sorties soient mutuellement décorrélées. Cette décorrélation se traduit sur la ligne du filtre séparateur que l'on cherche en une contrainte quadratique dans laquelle intervient l'autocorrélation du vecteur d'observation $(\mathbf{x}(n))_{n\in\mathbb{Z}}$. L'insertion du bruit dans le mo-

dèle permet, au travers de calculs analogues au cas précédent, de constater que la contrainte idéale souhaitée est obtenue en soustrayant la corrélation du bruit à celle du vecteur d'observation.

5 Résultats

Les performances de l'estimation du système inverse ont été évaluées sur la base d'un critère d'interférences entre symboles, qui s'écrit pour la $i^{\text{ème}}$ ($1 \leq i \leq N$) ligne du filtre global :

$$\mathrm{ISI}_i \triangleq \frac{\sum_{j,k} |G_{ij}(k)|^2 - \max_{j,k} |G_{ij}(k)|^2}{\max_{j,k} |G_{ij}(k)|^2} \qquad (13)$$

Les sources considérées étaient réelles du type PAM4 et le bruit additif était gaussien, décorrélé, de variance σ_b^2 sur chaque capteur. Le tableau 1 compare les résultats obtenus pour l'extraction de la première source lorsque le bruit est pris en compte ou ignoré. La répartition des valeurs du critère ISI_1 obtenues au cours des réalisations est donnée pour deux niveaux de bruit différents sur la figure 1. On constate que la méthode proposée permet d'améliorer l'estimation de l'inverse du filtre de mélange, ce qui s'avère d'autant plus utile que la puissance de bruit est importante.

Enfin, l'extraction successive des sources avec prise en compte du bruit est illustrée dans le cadre de la déflation. Le tableau 2 fournit les valeurs du critère de séparation lors de l'extraction successive de trois sources. La figure 2 indique la répartition sur l'ensemble des réalisations des valeurs des critères de séparation dans le cas d'un mélange de deux sources sur trois capteurs. Nous constatons à nouveau l'intérêt de la prise en compte du bruit. En parallèle, nous pouvons noter une dégradation des résultats au fur et à mesure des extractions, due à l'accumulation d'erreurs propre aux méthodes séquentielles.

6 Conclusion

Nous avons dans cet article illustré l'effet du bruit dans les méthodes aveugles de séparation de sources basées sur l'emploi du kurtosis. La plupart du temps négligé, l'influence du bruit dans les statistiques d'ordre 2 et 4 a ici été prise en compte dans le but d'une amélioration de l'estimation de l'inverse du système de mélange.

Un autre problème serait de minimiser l'erreur quadratique moyenne entre chaque source et son estimée respective. Il est intéressant, à ce propos, de noter que l'utilisation du filtre inverse obtenu par la méthode présentée n'est *a priori* pas la solution la plus efficace. En effet, le bruit se répercute sur chacune des estimées avec une variance $\mathrm{E}\{|\mathbf{W}_{i:}[z]\mathbf{b}(n)|^2\}$ qui n'est pas explicitement contrôlée. Néanmoins, il a été observé que l'égalisation par des critères de module constant présente de bonnes performances en terme d'estimation linéaire en moyenne quadratique. Des résultats théoriques [15, 14, 7] concernant le comportement des algorithmes de ce type invitent à penser qu'il conviendrait d'utiliser des méthodes d'estimation non-linéaire afin d'obtenir de meilleurs résultats.

Références

[1] V. Capdevielle, C. Servière, and J.-L. Lacoume. Blind separation of wide-band sources in the frequency domain. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2080–2083, Detroit, 1995.

[2] J.-F. Cardoso. Blind signal separation : statistical principles. *Proceedings of the IEEE. Special issue on blind identification and estimation*, 9(10) :2009–2026, October 1998.

[3] J.-F. Cardoso and D.-T. Pham. Optimization issues in noisy gaussian ICA. In *Proc. of ICA'04*, pages 41–48, Granada, Spain, 2004.

[4] M. Castella, A. Chevreuil, and J.-C. Pesquet. Séparation aveugle d'un mélange convolutif de sources non linéaires par une approche hiérarchique. In *GRETSI*, Paris, France, September 2003.

[5] P. Comon. Independent component analysis, a new concept. *Signal Processing*, 36(3) :287–314, April 1994.

[6] P. Comon. Contrasts for multichannel blind deconvolution. *IEEE Signal Processing Letters*, 3(7) :209–211, July 1996.

[7] I. Fijalkow, A. Touzni, and J. R. Treichler. Fractionally spaced equalization using CMA : Robustness to channel noise and lack of disparity. *IEEE Trans. Signal Processing*, 45(1) :56–66, January 1997.

[8] Y. Inouye and T. Sato. Iterative algorithms based on multistage criteria for multichannel blind deconvolution. *IEEE Trans. Signal Processing*, 47(6) :1759–1764, June 1999.

[9] E. Moreau and J.-C. Pesquet. Generalized contrasts for multichanel blind deconvolution of linear systems. *IEEE Signal Processing Letters*, 4(6) :182–183, June 1997.

[10] P. A. Regalia and M. Mboup. Properties of some blind equalization criteria in noisy multiuser environments. *IEEE Trans. Signal Processing*, 49(12) :3112–3122, December 2001.

[11] C. Simon, P. Loubaton, and C. Jutten. Separation of a class of convolutive mixtures : a contrast function approach. *Signal Processing*, (81) :883–887, 2001.

[12] A. Touzni, I. Fijalkow, M. G. Larimore, and J. R. Treichler. A globally convergent approach for blind MIMO adaptive deconvolution. *IEEE Trans. Signal Processing*, 49(6) :1166–1178, June 2001.

[13] J. K. Tugnait. Identification and deconvolution of multichannel linear non-gaussian processes using higher order statistics and inverse filter criteria. *IEEE Trans. Signal Processing*, 45(3) :658–672, March 1997.

[14] H. H. Zeng, L. Tong, and C. R. Johnson, Jr. Relationships between the constant modulus and wiener receivers. *IEEE Trans. on Information Theory*, 44(4) :1523–1538, July 1998.

[15] H. H. Zeng, L. Tong, and C. R. Johnson, Jr. An analysis of constant modulus receivers. *IEEE Trans. Signal Processing*, 47(11) :2990–2999, November 1999.

Simuler une distribution normale à support positif à partir de plusieurs lois candidates

Vincent MAZET[1], David BRIE[1], Jérôme IDIER[2]

[1]Université Henri Poincaré, Nancy 1, CRAN, CNRS UMR 7039,
BP 239, F-54506 Vandœuvre-lès-Nancy Cedex

[2]École Centrale de Nantes, IRCCyN, CNRS UMR 6597,
BP 92101, F-44321 Nantes Cedex 3

{vincent.mazet,david.brie}@cran.uhp-nancy.fr, jerome.idier@irccyn.ec-nantes.fr

Résumé – Dans cet article, nous présentons une méthodologie pour générer des échantillons suivant une loi normale à support positif dans le cas scalaire grâce à l'algorithme d'acceptation-rejet. Nous proposons plusieurs lois candidates, et choisissons celle qui donnera le meilleur taux d'acceptation moyen pour des paramètres de la loi cible donnés. Nous obtenons alors une méthode très rapide puisque générant très peu de rejet.

Abstract – In this article, we provide an accept-reject algorithm to simulate positive normal variables in the univariate case. The main idea consists in determining among different proposal distributions a priori chosen the one which gets the highest average probability of acceptation, depending on the shape of the target distribution. It yields a fast method since it generates low reject.

1 Introduction

En déconvolution impulsionnelle, le signal à restaurer est souvent modélisé par un processus Bernoulli-gaussien. Or, dans certaines applications (telles que la spectroscopie [6, 9], ou la génétique [7]), les pics sont tous positifs. Ils peuvent donc être modélisés par un a priori Bernoulli-gaussien à support dans $\mathbb{R}^+$, la loi a posteriori devient alors le mélange d'un pic de Dirac et d'une gaussienne de moyenne et de variance quelconques tronquée en zéro (c'est-à-dire à support positif). Dans le cas où l'on souhaite résoudre le problème en utilisant une méthode MCMC (notamment lorsque le nombre de variables est très important), il est nécessaire de générer des variables gaussiennes positives. Un autre type d'application est la séparation myope de sources positives dont les coefficients de mélanges sont également positifs en analyse de mélanges chimiques [8].

Nous présentons donc dans cet article une approche pour simuler la loi normale à support positif qui a pour expression :

$$f(x) = \frac{1}{C}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\mathbb{1}_{\mathbb{R}^+} \qquad (1)$$

où

$$C = \sqrt{\frac{\pi\sigma^2}{2}}\left[1+\operatorname{erf}\left(\frac{\mu}{\sqrt{2\sigma^2}}\right)\right].$$

Notons que la valeur de C n'est pas nécessaire pour la mise en œuvre de l'algorithme. On peut calculer facilement la moyenne et la variance de cette loi :

$$\mathbb{E}[x] = \mu + \sqrt{\frac{2\sigma^2}{\pi}}\frac{\exp\left(-\mu^2/2\sigma^2\right)}{1+\operatorname{erf}\left(\mu/\sqrt{2\sigma^2}\right)},$$

$$\operatorname{Var}[x] = \sigma^2 + \frac{\mu^2}{4} - \left[\frac{\mu}{2} + \sqrt{\frac{2\sigma^2}{\pi}}\frac{\exp\left(-\mu^2/2\sigma^2\right)}{1+\operatorname{erf}\left(\mu/\sqrt{2\sigma^2}\right)}\right]^2.$$

La technique d'inversion de la fonction de répartition [3, 4] consiste à générer $u \sim \mathcal{U}_{[0,1]}$, puis à calculer :

$$x = \mu + \sqrt{2\sigma^2}\operatorname{erf}^{-1}\left(u + \operatorname{erf}(\mu/\sqrt{2\sigma^2})(u-1)\right)$$

où erf est la fonction erreur. Cette méthode a l'avantage de fournir une solution explicite mais suppose que les fonctions erf et erf^{-1} sont parfaitement connues. En pratique, l'utilisation de ces fonctions peut se révéler inefficace car elles ne peuvent être qu'approchées numériquement et l'erreur d'approximation devient importante lorsque $|\mu|$ est trop grand (cf. section 4) [5, 10].

Une autre approche consiste à utiliser un algorithme d'acceptation-rejet [3, 10, 11]. La loi candidate la plus simple est évidemment la loi normale, mais elle n'est adaptée que dans le cas où μ est suffisamment grand (cf. figure 2). Au contraire, la loi exponentielle proposée par Robert [10] n'est bien adaptée que lorsque μ est suffisamment petit (cf. figure 2). En effet, lorsque $\mu \to -\infty$, la loi normale tronquée tend à ressembler à une loi exponentielle [5].

Comme la forme de la loi varie en fonction de μ et σ^2, nous proposons d'utiliser une méthode basée sur l'algorithme d'acceptation-rejet en utilisant plusieurs lois candidates g, l'idée étant que chacune soit adaptée aux formes particulières de la loi cible. Cette idée d'un *algorithme d'acceptation-rejet mixte* a déjà été utilisée par Geweke [5], qui propose deux lois candidates (les lois normale et exponentielle) pour simuler une loi normale tronquée unilatéralement et détermine les intervalles d'utilisation de ces lois de manière empirique. Or, une détermination empirique donne des valeurs d'intervalles dépendant du code utilisé. Pour éviter cela, nous déterminons les valeurs des intervalles en fonction du taux d'acceptation théorique (lié au temps de calcul), ce qui permet d'obtenir des valeurs indépendantes du codage. De plus, afin d'améliorer les performances

de l'algorithme, nous proposons d'utiliser quatre lois candidates. Cette idée d'algorithme d'acceptation-rejet mixte peut s'adapter sur toute loi cible qui n'est pas directement échantillonnable par les techniques classiques, et dont la forme varie en fonction de ses paramètres.

Dans la section 2, nous présentons la méthodologie de manière générale puis nous l'appliquons au cas particulier de la simulation d'une loi normale à support positif dans la section 3. La section 4 propose des comparaisons avec des méthodes existantes et montre que l'approche proposée simule correctement une loi normale tronquée. Enfin, la section 5 conclue ce travail en y présentant quelques perspectives de recherche.

2 Méthodologie générale

L'algorithme d'acceptation-rejet requiert la détermination d'une loi candidate g et d'une constante M telles que

$$\forall x \in S, \quad M \geq f(x)/g(x) \tag{2}$$

où S est le support de f. La loi candidate $g(x)$ est nécessairement non nulle sur S afin que M soit fini et donc que le taux d'acceptation moyen soit non nul (cf. équation (4)). L'algorithme d'acceptation-rejet découle du lemme suivant :

Lemme 1 [3, 10, 11] La variable aléatoire x résultante de l'algorithme suivant est distribuée suivant f :

1. générer $z \sim g(z)$ et $u \sim \mathcal{U}_{[0,1]}$,
2. calculer $\rho(z) = f(z)/Mg(z)$,
3. si $u \leq \rho(z) : x = z$ (acceptation),
   sinon : retour en 1 (rejet).

Le choix des lois candidates est déterminant pour les performances de la méthode. Tout d'abord, elles doivent être facilement échantillonnables, sinon l'algorithme d'acceptation-rejet perd son intérêt. En particulier, elles doivent être simulables avec un taux d'acceptation de 1, sinon le taux d'acceptation moyen correspondant doit être pris en compte dans le calcul du taux d'acceptation moyen global. Précisons également que le choix des lois candidates doit prendre en compte la complexité de l'algorithme : certaines lois intéressantes a priori s'avéreront inadéquates car la détermination de M ou le calcul de ses paramètres peut être difficile, gourmand en temps de calcul, voire impossible ! Ainsi, on peut choisir des lois courantes (dans notre exemple : les lois normale et exponentielle) ou construire des lois particulières afin d'obtenir des lois candidates convenables (une loi normale couplée à la loi uniforme pour notre exemple).

Pour le choix de M, n'importe quelle constante vérifiant l'équation (2) convient ; cependant, il est nécessaire que M soit la plus petite possible pour avoir un taux d'acceptation moyen élevé (équation (4)). La valeur optimale de M est donc :

$$M = \max_{x \in S} f(x)/g(x).$$

Malheureusement, cette valeur n'est pas toujours calculable. Cependant, si elle l'est, alors on peut déterminer le taux d'acceptation ρ :

$$\rho(x) = f(x)/Mg(x). \tag{3}$$

Enfin, le taux d'acceptation moyen $\overline{\rho} = \mathbb{E}[\rho(X)]$ nous permettra de déterminer les zones d'utilisation des lois candidates :

$$\overline{\rho} \triangleq \int \rho(x)g(x)dx = \frac{1}{M}\int f(x)dx = \frac{1}{M}. \tag{4}$$

Remarquons que si $g(x)$ est proche de zéro, alors M devient très grand et le taux d'acceptation moyen diminue. L'efficacité de l'algorithme dépend donc de l'adéquation entre f et g ; en particulier, pour que M reste borné, il faut que g ait des queues plus lourdes que celle de f, sans toutefois que leur rapport soit trop grand, au risque d'augmenter M [11].

L'expression du taux d'acceptation dépend dans certains cas des paramètres de la loi candidate (par exemple, le paramètre α de la loi exponentielle dans notre cas particulier : cf. section 3). Il faut alors trouver les valeurs de ces paramètres qui maximisent le taux d'acceptation.

Parmi l'ensemble des lois candidates utilisées, une seule fournit le meilleur taux d'acceptation moyen pour des paramètres de la loi cible donnés. Il faut donc déterminer les intervalles sur ces paramètres pour déterminer quelle est la meilleure loi candidate à utiliser sur l'intervalle considéré. L'algorithme proposé est donc le même que l'algorithme d'acceptation-rejet avec une étape préliminaire qui consiste à choisir la meilleure loi candidate g_i en fonction des paramètres de f, puis à calculer la constante M correspondante :

1. sélectionner la loi candidate g_i la plus appropriée en fonction des paramètres de f,
2. calculer M_i,
3. générer $z \sim g_i(z)$ et $u \sim \mathcal{U}_{[0,1]}$,
4. calculer $\rho(z) = f(z)/M_i g_i(z)$,
5. si $u \leq \rho(z) : x = z$ (acceptation),
   sinon : retour en 3 (rejet).

Notons que si f est tronquée à $t \neq 0$, la méthode peut être adaptée par simple translation de la variable aléatoire générée $X : Y = X + t$. De même, sans perdre en généralité, on peut se restreindre au cas $\sigma^2 = 1$, puisque tous les autres cas se déduisent par changement d'échelle $Y = X\sigma$: nous supposerons dans la suite que $\sigma^2 = 1$.

3 Simulation d'une loi normale tronquée

Quatre lois candidates (représentées figure 1) sont proposées pour simuler la loi cible f :

① La loi normale :

$$g_1(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) ;$$

② La loi normale couplée à la loi uniforme (c'est une distribution définie sur $\mathbb{R}^+$, uniforme sur $[0, \mu[$ et qui suit la loi normale $\mathcal{N}(\mu, \sigma^2)$ sur $[\mu, +\infty[$, $\mu \geq 0$) :

$$g_2(x) = \frac{\mathbb{1}_{\mathbb{R}^+}}{\mu + \sqrt{\frac{\pi\sigma^2}{2}}} \begin{cases} 1 & \text{si } 0 \leq x < \mu, \\ \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) & \text{si } x \geq \mu ; \end{cases}$$

③ La loi normale tronquée en la moyenne ($\mu \leq 0$) :

$$g_3(x) = \frac{2}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathbb{1}_{[\mu,+\infty]} ;$$

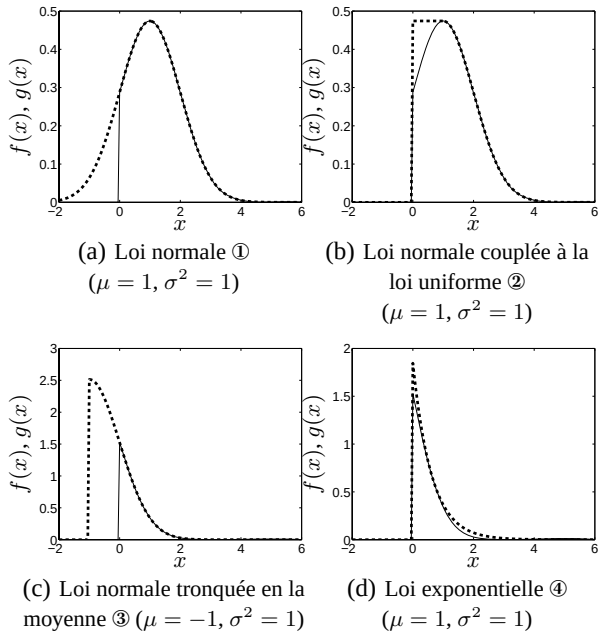
④ La loi exponentielle [10] :

$$g_4(x) = \alpha \exp(-\alpha x)\mathbb{1}_{\mathbb{R}^+}$$

où la valeur de α correspond à la valeur qui maximise le taux d'acceptation moyen :

$$\alpha = \left(\sqrt{\mu^2 + 4\sigma^2} - \mu\right)/2\sigma^2. \tag{5}$$

Des techniques de génération d'échantillons distribués suivant ces lois candidates sont présentées dans l'annexe A.



(a) Loi normale ① ($\mu = 1, \sigma^2 = 1$)
(b) Loi normale couplée à la loi uniforme ② ($\mu = 1, \sigma^2 = 1$)
(c) Loi normale tronquée en la moyenne ③ ($\mu = -1, \sigma^2 = 1$)
(d) Loi exponentielle ④ ($\mu = 1, \sigma^2 = 1$)

FIG. 1 – Lois candidates proposées

Notre choix est motivé par le fait que les lois ① et ④ permettent d'obtenir un taux d'acceptation moyen très important lorsque $|\mu| \gg 0$, et les lois ② et ③ permettent d'améliorer le taux d'acceptation moyen autour de zéro (cf. figure 2). Pour chacune de ces lois, nous pouvons calculer l'expression des constantes M (et ainsi les taux d'acceptation moyens) :

$$M_1 = \sqrt{2\pi\sigma^2}/C$$
$$M_2 = \left(\mu + \sqrt{\pi\sigma^2/2}\right)/C$$
$$M_3 = \sqrt{2\pi\sigma^2}/2C$$
$$M_4 = \exp\left(\frac{\alpha}{2}(2\mu + \alpha\sigma^2)\right)/\alpha C$$

Le calcul des taux d'acceptation $\rho(x)$ est alors direct.

Les taux d'acceptation moyens pour chacune des quatre lois candidates ont été représentés en fonction de μ figure 2. Nous pouvons alors trouver les expression des trois points d'intersections qui délimitent les zones. Les calculs sont directs pour μ_B et μ_C, mais on ne peut que déterminer une valeur approchée de μ_A. En effet, d'après l'équation (5), on a :

$$\mu = (1 - \alpha^2\sigma^2)/\alpha,$$

et, en égalant les taux d'acceptation moyen des lois ③ et ④, on obtient :

$$\alpha\sigma \exp(\alpha^2\sigma^2/2) = e\sqrt{2/\pi}.$$

Cette dernière équation définit $\alpha\sigma$ sous une forme implicite. On peut en déduire que $\alpha\sigma$ s'exprime sous la forme

$$\alpha\sigma = \sqrt{W(2e^2/\pi)} \approx 1,1367,$$

où W est la fonction de Lambert définie comme l'inverse multivaluée de la fonction $f(w) = we^w$ [2]. Ainsi :

$$\mu_A \approx -0,257\sigma, \qquad \mu_B = 0, \qquad \mu_C = \sigma\sqrt{\pi/2}.$$

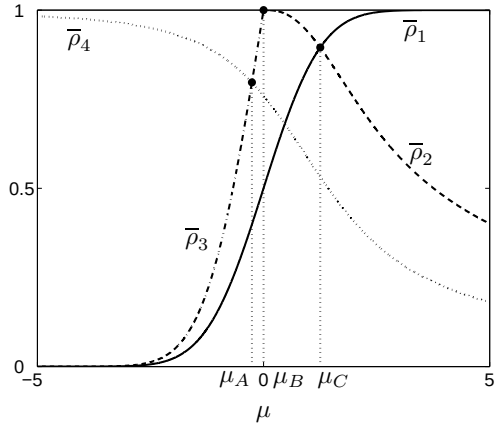


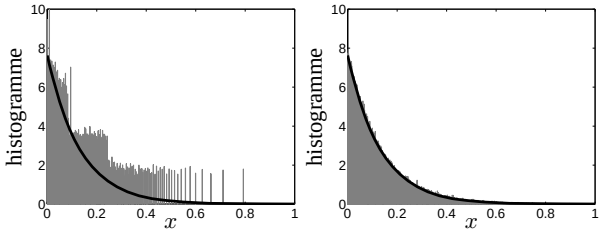
FIG. 2 – Taux d'acceptation moyens pour les quatre lois candidates en fonction de μ ($\sigma^2 = 1$)

Le taux d'acceptation moyen le plus faible correspond à $\mu = \mu_A$; il est égal à 0,797 environ, ce qui est un très bon taux d'acceptation. L'algorithme obtenu est donc très rapide car il génère très peu de rejet.

4 Simulations numériques

Toutes les simulations ont été faites avec Matlab. La fonction erf est définie dans [1], et les variables uniformes et normales sont générées avec `rand` et `randn` respectivement.

La méthode d'inversion [3, 4], bien qu'elle soit explicite et que le taux d'acceptation soit de 1, s'avère inefficace dans certains cas pratiques ou $-\mu$ est trop grand [5]. Cela est dû à la fonction erf dont le calcul engendre des approximations créant alors des problèmes numériques. Pour illustrer ce propos, considérons la génération de 100 000 variables normales positives avec le paramètre $\mu = -7.5$ grâce à la méthode d'inversion et l'approche proposée. Les histogrammes correspondants sont représentés figure 3[1]. Il est clair que la méthode



(a) Méthode d'inversion (b) Approche proposée

FIG. 3 – Histogrammes de 100 000 variables ($\mu = -7.5$, $\sigma^2 = 1$), et la distribution normale tronquée correspondante.

d'inversion s'avère inefficace alors que l'approche proposée fonctionne très bien. Évidemment, les performances de la méthode d'inversion pourraient être améliorées en utilisant une meilleure approximation de la fonction erf, mais cela entraînerait un coup calculatoire conséquent. De plus, pour certaines

[1] Les éventuelles valeurs infinies générées par la méthode d'inversion n'ont pas été représentées sur l'histogramme.

	AP	①	②	④
$\mu = -2$	0,36 s	11,094 s	($\mu < 0$!)	0,39 s
$\mu = 0$	0,406 s	0,485 s	0,422 s	0,454 s
$\mu = 0.5$	0,406 s	0,375 s	0,422 s	0,515 s
$\mu = 2$	0,281 s	0,297 s	0,531 s	0,813 s

TAB. 1 – Comparaison des temps de calcul pour l'approche proposée (AP) et les algorithmes d'acceptation-rejet utilisant les lois candidates ①, ② et ④.

valeurs des paramètres, la méthode d'inversion peut générer des valeurs négatives ou infinies. Si elles restent relativement peu nombreuses, le problème peut être résolu en générant de nouvelles variables, mais, dans certains cas, la méthode peut ne donner que des valeurs inexploitables (par exemple avec $\mu = -8.5$) : la méthode d'inversion s'avère alors inappropriée.

Comparons maintenant l'approche proposée avec trois algorithmes d'acceptation-rejet classiques utilisant les lois candidates ①, ② et ④. Nous comparons le temps de calcul pour différents μ, ce qui, en général, est relativement proportionnel au taux d'acceptation moyen. Les temps de simulation de 10 000 variables sont présentés dans le tableau 1. Dans tous les cas, les temps de calcul sont cohérents avec les taux d'acceptation moyens de la figure 2, sauf pour $\mu = 0,5$ où la loi normale ① est plus rapide que la loi normale couplée à la loi uniforme ②. Cela est dû au fait que générer une variable normale est beaucoup plus rapide que générer une variable suivant la loi ②. Mais comme cette différence dépend du programme utilisé, il est difficile de déterminer les différents intervalles à partir des temps de calcul réels. C'est pourquoi nous préférons le faire à partir des taux d'acceptation moyens théoriques, ce qui permet tout de même d'obtenir de bonnes valeurs pour les intervalles. De toute manière, la différence est faible et particulière au cas $\mu = 0,5$.

5 Conclusion

Nous avons présenté une approche pour simuler des variables suivant une loi dont la forme varie avec ses paramètres. L'idée développée utilise un algorithme d'acceptation-rejet dont la loi candidate est sélectionnée parmi d'autres lois choisies a priori. L'approche est ensuite appliquée au cas particulier de la génération de variables aléatoires normales à support positif. Quatre lois candidates ont été présentées, et les intervalles sur lesquelles elles sont le plus performantes en terme de taux d'acceptation ont été déterminés. Des simulations ont illustré le fait que la méthode fonctionne bien et rapidement dans toutes les situations possibles.

Le travaux futurs se concentrerons sur la recherche de lois candidates plus performantes, telle que la loi normale couplée à un polynôme, ou la loi normale tronquée en la moyenne dont les deux paramètres sont déterminés pour maximiser le taux d'acceptation moyen. De plus, les recherches pourront également être tournées vers le cas de la simulation d'une loi normale tronquée des deux côtés ce qui peut être utile dans le traitement de données censurées, ou pour simuler des lois normales multidimensionnelles à support positif.

A Simulation des lois candidates

Un échantillon x distribué suivant la loi ② est situé soit dans la partie uniforme (d'aire $\mathcal{A}_u$), soit dans la partie gaussienne (d'aire $\mathcal{A}_g$). Comme ces aires sont égales à :

$$\mathcal{A}_u = \frac{\mu}{\mu + \sqrt{\pi\sigma^2/2}} \qquad \text{et} \qquad \mathcal{A}_g = \frac{\sqrt{\pi\sigma^2/2}}{\mu + \sqrt{\pi\sigma^2/2}},$$

elles sont de somme unité. Ainsi, l'algorithme suivant permet de générer un échantillon x suivant la loi ② :

1. générer $u \sim \mathcal{U}_{[0,1]}$,
2. si $u < \mathcal{A}_u$, générer $v \sim \mathcal{U}_{[0,1]}$ puis calculer $x = \mu v$,
   sinon, générer $v \sim \mathcal{N}(0, \sigma^2)$ puis calculer $x = |v| + \mu$.

Pour générer un échantillon x suivant la loi ③, il suffit de décaler de μ la valeur absolue d'une variable gaussienne centrée :

$$x = |y| + \mu \qquad \text{où} \qquad y \sim \mathcal{N}(0, \sigma^2).$$

Enfin, la méthode d'inversion de la fonction de répartition permet de simuler la loi exponentielle en générant $u \sim \mathcal{U}_{[0,1]}$, puis en calculant $x = -\ln(1-u)/\alpha$.

Références

[1] W.J. Cody, Rational Chebyshev approximations for the error function, *Math. Comp.* (1969) 631–638.

[2] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey et D.E. Knuth, On the Lambert W Function, *Adv. Comp. Math.* **5** (1996) 329–359.

[3] L. Devroye, *Non-uniform random variate generation*. Springer-Verlag, New York (1986). Disponible sur http ://jeff.cs.mcgill.ca/˜luc/rnbookindex.html.

[4] A.E. Gelfand, A.F.M. Smith et T.-M. Lee, Bayesian Analysis of constrained parameter and truncated problems using Gibbs sampling, *J. Amer. Stat. Assoc.* **87** (1992) 523–532.

[5] J. Geweke, Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints, *Computing Science and Statistics : Proceedings of the 23rd Symposium on the Interface*, (1991) 571–578.

[6] S. Gulam Razul, W.J. Fitzgerald et C. Andrieu, Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC, *Nucl. Inst. Meth. A* **497** (2003) 492–510.

[7] N.M. Haan et S.J. Godsill, Bayesian model for DNA sequencing, *ICASSP'02*, Orlando, Floride, USA, (2002) **4** 4020–4023.

[8] S. Moussaoui, D. Brie, A. Mohammad-Djafari et C. Carteret, Separation of non-negative mixture of non-negative sources using a Bayesian approach and MCMC sampling, soumis à *IEEE Trans. Signal Process.* (2005)

[9] V. Mazet, J. Idier et D. Brie, Déconvolution impulsionnelle positive myope, *20e GRETSI*, Louvain-la-Neuve, Belgique, 6-9 septembre 2005.

[10] C.P. Robert, Simulation of truncated normal variables, *Stat. Comp.* **5** (1995) 121–125.

[11] C.P. Robert, *Méthodes de Monte Carlo par chaînes de Markov*, Economica (1996).

La programme Matlab est librement téléchargeable sur Internet à l'adresse suivante : http ://mtde.cran.uhp-nancy.fr/Personnes/Perso_Mazet/rpnorm-fr.htm.

Déconvolution impulsionnelle positive myope

Vincent MAZET[1], Jérôme IDIER[2], David BRIE[1]

[1]Université Henri Poincaré, Nancy 1, CRAN, CNRS UMR 7039,
BP 239, F-54506 Vandœuvre-lès-Nancy Cedex

[2]École Centrale de Nantes, IRCCyN, CNRS UMR 6597,
BP 92101, F-44321 Nantes Cedex 3

{vincent.mazet,david.brie}@cran.uhp-nancy.fr, jerome.idier@irccyn.ec-nantes.fr

Résumé – Dans cet article, nous considérons le problème de la déconvolution impulsionnelle positive myope, c'est-à-dire que nous souhaitons restaurer un signal impulsionnel dans le cas particulier où tous les pics sont positifs, en estimant également les paramètres de la réponse impulsionnelle et les hyperparamètres du problème. Du fait de la complexité du problème, nous utilisons une méthode MCMC basée sur l'algorithme de Gibbs. Enfin, deux indices originaux sont proposés pour estimer le signal impulsionnel et permettre une analyse plus pertinente de la chaîne de Markov résultante.

Abstract – We consider in this paper the problem of the blind deconvolution of positive peaks, *i.e.* we wish to restore a sparse spike train whose peaks are positive and to estimate the impulse response and the problem hyperparameters. Because of the complexity of the posterior global criterion, we use an MCMC method based on the Gibbs sampler. Finally, we propose two new indexes to give an estimate of the signal, yielding an efficient analysis of the Markov chain.

1 Introduction

Nous considérons dans cet article le problème de la déconvolution impulsionnelle positive myope non supervisée, c'est-à-dire que nous souhaitons estimer conjointement le signal impulsionnel $\mathbf{z} \in \mathbb{R}^{+N}$, la réponse impulsionnelle $\mathbf{h} \in \mathbb{R}^{M}$ et les différents hyperparamètres du problème connaissant la sortie du système $\mathbf{y} \in \mathbb{R}^{N}$. Nous modélisons le problème sous la forme classique suivante :

$$\mathbf{y} = \mathbf{H}\mathbf{z} + \mathbf{b},$$

où $\mathbf{H}$ est la matrice de Toeplitz de $\mathbf{h}$, et où le bruit $\mathbf{b}$ permet la prise en compte des incertitudes de mesure et de modèle.

Le problème de la déconvolution impulsionnelle myope peut se rencontrer en spectroscopie [4], en sismique réflexion [1, 9], en électricité (décharges partielles) ou encore en génétique (séquençage d'ADN [5]).

La déconvolution myope, qui, comme la déconvolution classique, est un problème inverse mal posé, cherche de plus à estimer un grand nombre de variables par rapport aux données disponibles. La résolution de ce type de problème impose donc d'inclure des connaissances a priori fortes pour régulariser la solution, comme par exemple imposer de restaurer un signal impulsionnel. Dans ce contexte, l'approche bayésienne associée au développement des techniques de simulations stochastiques fourni un cadre particulièrement attractif pour résoudre ce type de problèmes.

Ainsi, certains auteurs imposent un a priori Bernoulli-gaussien sur $\mathbf{z}$ et gaussien sur $\mathbf{h}$ [1, 9] ; ils utilisent ensuite l'échantillonneur de Gibbs qui permet de générer une chaîne de Markov pour chaque quantité à estimer. Cheng *et al.* [1] et Rosec *et al.* [9] proposent deux estimateurs différents pour le signal impulsionnel (une discussion à propos du choix de ces estimateurs est proposée section 4). D'autres auteurs préfèrent estimer les paramètres de $\mathbf{z}$ et $\mathbf{h}$, à savoir la position et l'amplitude des pics et la largeur de la réponse impulsionnelle [4, 5] : l'a priori Bernoulli-gaussien n'est pas utilisé puisque $\mathbf{z}$ est modélisé par un ensemble de pics dont les positions et les amplitudes sont à estimer. Différentes lois a priori sur les paramètres y sont proposées, mais tous appliquent l'algorithme RJMCMC. L'estimation des variables se fait ensuite en utilisant l'estimateur MAP (maximum a posteriori). Notons que le fait d'estimer seulement les paramètres de $\mathbf{h}$ plutôt que $\mathbf{h}$ entier permet de diminuer le nombre de variables à estimer, mais cela implique que la structure de la réponse impulsionnelle soit connue a priori, ce qui est le cas dans certaines applications, en particulier la spectroscopie.

Nous considérons ici le cas particulier (courant en spectroscopie ou en génétique) où les impulsions sont positives :

$$\forall\, n \in \{1, \ldots, N\},\ \mathbf{z}_n \geq 0.$$

De plus, pour éviter d'avoir à optimiser un critère a posteriori trop compliqué, nous suivons la majorité des travaux en se plaçant dans un cadre bayésien et en utilisant une méthode MCMC basée sur l'échantillonneur de Gibbs [1, 8]. Cet algorithme itératif génère une chaîne de Markov du vecteur de variables $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N\}$. Partant de valeurs initiales $\boldsymbol{\theta}^{(0)} = \{\boldsymbol{\theta}_1^{(0)}, \ldots, \boldsymbol{\theta}_N^{(0)}\}$, on simule à tour de rôle les variables de $\boldsymbol{\theta}$ à partir leur distributions a posteriori conditionnelles. Ainsi, à l'itération i et $\forall n \in \{1, \ldots, N\}$,

$$\boldsymbol{\theta}_n^{(i)} \sim p(\boldsymbol{\theta}_n | \mathbf{y}, \boldsymbol{\theta}_1^{(i)}, \ldots, \boldsymbol{\theta}_{n-1}^{(i)}, \boldsymbol{\theta}_{n+1}^{(i-1)}, \ldots, \boldsymbol{\theta}_N^{(i-1)}).$$

La section 2 présente les lois a priori choisies pour chaque quantité dont on déduit les lois a posteriori (en section 3), à partir desquelles l'échantillonneur de Gibbs permet de simuler chaque variable. Le choix des estimateurs est discuté dans la

section 4 : nous y proposons deux indices originaux permettant d'obtenir une estimation pertinente du signal impulsionnel. Enfin, la section 5 conclut ce travail.

2 Lois a priori

2.1 Loi a priori de $\mathbf{z}$

Le signal impulsionnel est modélisé par un processus Bernoulli-gaussien à support positif. Aussi, le signal est représenté par le couple $\mathbf{z} = (\mathbf{q}, \mathbf{x})$ où

- $\mathbf{q}$ fixe l'occurrence des pics et suit une loi de Bernoulli de paramètre λ [8] :

$$\mathbf{q}_n \sim \mathcal{B}er(\lambda) ;$$

- $\mathbf{x}$ correspond à l'amplitude des pics et suit une loi gaussienne à support positif de variance $r_1 > 0$ s'il y a un pic au point n ($\mathbf{q}_n = 1$) ou $r_0 = 0$ sinon ($\mathbf{q}_n = 0$) :

$$\mathbf{x}_n \sim \mathcal{N}^+\big(0, r_1\mathbf{q}_n + r_0(1-\mathbf{q}_n)\big).$$

  $\mathcal{N}^+(\mu, \sigma^2)$ représente la loi gaussienne de moyenne μ et de variance σ^2 tronquée en zéro.

La loi a priori de $\mathbf{z}_n$ s'écrit alors :

$$\begin{aligned} p(\mathbf{z}_n|\mathbf{z}_{-n}, \lambda, r_1, s, r_{\mathbf{b}}) &= p(\mathbf{x}_n, \mathbf{q}_n|\mathbf{z}_{-n}, \lambda, r_1, s, r_{\mathbf{b}}) \\ &= p(\mathbf{x}_n|\mathbf{q}_n, r_1)p(\mathbf{q}_n|\lambda), \end{aligned}$$

où $\mathbf{z}_{-n} = \{\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}, \mathbf{z}_{n+1}, \ldots, \mathbf{z}_N\}$.

En pratique, lorsque le nombre de pics $N_{\mathbf{q}} = \sum_n \mathbf{q}_n$ devient très faible, il apparaît que la solution qui consiste à avoir un grand nombre de pics de petite amplitude devient très probable : c'est typiquement un problème de dégénérescence. Il est donc nécessaire de fixer un a priori sur λ qui permette de limiter le nombre de pics. Nous choisissons la loi a priori conjuguée[1] :

$$\lambda \sim \mathcal{B}e(1, N+1).$$

où $\mathcal{B}e(\alpha, \beta)$ représente la famille des lois beta [8].

En outre, si aucun a priori n'est fixé sur r_1, sa loi a posteriori s'écrit :

$$r_1 \sim \mathcal{IG}\left(N_{\mathbf{q}}/2 - 1, \mathbf{x}_n^T\mathbf{x}_n/2\right),$$

où $\mathcal{IG}(\alpha, \beta)$ représente la famille des lois inverse gamma [8]. Un problème numérique se pose alors dans le cas où $N_{\mathbf{q}} \leq 2$: le premier paramètre de la loi inverse gamma n'est plus strictement positif. Pour éviter cela, nous posons un a priori sur r_1 en choisissant également une loi conjuguée :

$$r_1 \sim \mathcal{IG}(\alpha_1, \beta_1)$$

dont l'espérance et la variance sont [8] :

$$\begin{aligned} \mathbb{E}[r_1] &= \beta_1/(\alpha_1 - 1), && (\text{si } \alpha_1 > 1), \\ \text{Var}[r_1] &= \beta_1^2/\left((\alpha_1-1)^2(\alpha_1-2)\right) && (\text{si } \alpha_1 > 2). \end{aligned}$$

L'a priori a été choisi pour éviter les erreurs numériques. Cependant, afin de minimiser son influence sur l'estimation, il doit être le moins informatif possible. À ce titre, la variance doit être choisie la plus grande possible et l'espérance de manière à ne pas affecter l'estimation. On choisit donc de prendre l'espérance égale à une estimation grossière de la variance de l'amplitude des pics. Dès lors, puisque $\text{Var}[r_1] = \mathbb{E}[r_1]^2/(\alpha_1 - 2)$, il faut que α_1 soit le plus petit possible, donc $\alpha_1 = 2 + \varepsilon$, où $\varepsilon \ll 1$. D'après l'expression de l'espérance on en déduit : $\beta_1 = \mathbb{E}[r_1](\alpha_1 - 1) = \mathbb{E}[r_1](1 + \varepsilon)$.

[1] Une loi a priori conjuguée permet d'obtenir une loi a posteriori de même forme mais dont les paramètres sont fonction de la vraisemblance et de l'a priori.

2.2 Loi a priori de s

Le cas le plus général consiste à supposer la réponse impulsionnelle inconnue. Cependant, dans certaines applications (en particulier la spectroscopie), on peut supposer que la structure de $\mathbf{h}$ est connue. Seuls les paramètres de la structure sont à estimer, limitant ainsi le nombre d'inconnues du problèmes. C'est ce cas que nous traitons : dans la suite de l'article, nous prenons l'exemple d'une réponse impulsionnelle lorentzienne de paramètre s :

$$\mathbf{h}_n = \frac{s^2}{s^2 + n^2}.$$

Alors, seul le scalaire s est à estimer. Le seul a priori imposé à s est qu'il soit positif.

2.3 Loi a priori de $\mathbf{b}$

$\mathbf{b}$ est modélisé par un processus gaussien i.i.d. de moyenne nulle et de variance $r_{\mathbf{b}}\mathbf{I}$. De même que pour s, le seul a priori imposé est la positivité de $r_{\mathbf{b}}$.

2.4 Indéterminations en déconvolution myope

Dans le problème de la déconvolution myope de signaux réels, plusieurs types d'indéterminations peuvent intervenir [1] :

- l'indétermination sur le contenu spectral tient au fait qu'il n'est pas possible a priori de différencier les cas $h(n) \star z(n)$ et $h_1(n) \star (h_2(n) \star z(n))$ où $h_1(n) \star h_2(n) = h(n)$. Dans notre cas, l'a priori Bernoulli-gaussien à support positif sur z permet de lever cette indétermination ;
- l'indétermination temporelle correspond au fait que $\forall \eta$, $h(n) \star z(n) = h(n+\eta) \star z(n-\eta)$: le signal impulsionnel et la réponse impulsionnelle peuvent être décalés d'une valeur inverse l'un par rapport à l'autre. Ici, l'indétermination est levée en fixant le maximum de la réponse impulsionnelle en $n = 0$;
- enfin, l'indétermination sur l'amplitude correspond au fait que $\forall \alpha, h(n) \star z(n) = \alpha h(n) \star z(n)/\alpha$. On lève l'indétermination en fixant le maximum de la réponse impulsionnelle à 1.

3 Lois a posteriori

3.1 Fonction de vraisemblance

Il y a donc $2N+4$ variables à estimer : $\boldsymbol{\theta} = \{\mathbf{x}, \mathbf{q}, r_1, \lambda, s, r_{\mathbf{b}}\}$. L'a priori gaussien sur le bruit permet de déterminer l'expression de la vraisemblance selon :

$$p(\mathbf{y}|\mathbf{x}, s, r_{\mathbf{b}}) = \frac{1}{(2\pi r_{\mathbf{b}})^{N/2}} \exp\left(-\frac{1}{2r_{\mathbf{b}}}||\mathbf{y} - \mathbf{H}\mathbf{x}||^2\right).$$

3.2 Détermination des lois a posteriori

La règle de Bayes permet de déterminer les lois a posteriori à simuler :

$$\begin{aligned}
\mathbf{x}_n &\sim \mathcal{N}^+\big(\mu_{\mathbf{q}_n,n}, \rho_{\mathbf{q}_n,n}\big),\\
\mathbf{q}_n &\sim \mathcal{B}er\big(\lambda_{1,n}\big),\\
r_1 &\sim \mathcal{IG}\big(N_{\mathbf{q}}/2+\alpha_1, \mathbf{x}_n^T\mathbf{x}_n/2+\beta_1\big),\\
\lambda &\sim \mathcal{B}e\big(N_{\mathbf{q}}+1, 2N-N_{\mathbf{q}}+1\big),\\
s &\sim \exp\big(-||\mathbf{y}-\mathbf{H}(s)\mathbf{x}||^2/2r_{\mathbf{b}}\big),\\
r_{\mathbf{b}} &\sim \mathcal{IG}\big(N/2-1, ||\mathbf{y}-\mathbf{H}\mathbf{x}||^2/2\big),
\end{aligned}$$

où :

$$\rho_{1,n} = \frac{r_1 r_{\mathbf{b}}}{r_{\mathbf{b}} + r_1 \mathbf{H}_n^T\mathbf{H}_n}, \qquad \mu_{1,n} = \frac{r_{1,n}}{r_{\mathbf{b}}}\mathbf{e}_n^T\mathbf{H}_n,$$

$$\rho_{0,n} = 0, \qquad \mu_{0,n} = 0, \qquad \mathbf{e}_n = \mathbf{y} - (\mathbf{H}\mathbf{x} - \mathbf{H}_n\mathbf{x}_n),$$

$$\lambda_{1,n} = \left[1 + \frac{1-\lambda}{\lambda}\sqrt{\frac{r_1}{\rho_{1,n}}}\exp\left(-\frac{\mu_{1,n}^2}{2\rho_{1,n}}\right)\right]^{-1},$$

et où $\mathbf{H}_n$ est la $n^{\text{ème}}$ colonne de $\mathbf{H}$. Notons que dans la loi a posteriori de s, la dépendance de $\mathbf{H}$ par rapport à s a été rendue explicite.

3.3 Simulation des lois à posteriori

L'échantillonnage des variables $\mathbf{q}_n$, λ, r_1 et $r_{\mathbf{b}}$ est facile puisqu'il s'agit de lois classiques.

Pour $\mathbf{x}_n$, des techniques efficaces de génération de variables gaussiennes positives existent [2, 6, 7] ; en particulier, dans [6], nous proposons un algorithme d'acceptation-rejet mixte utilisant plusieurs lois candidates et dont le taux d'acceptation moyen est très élevé, donnant ainsi une méthode d'échantillonnage rapide.

Enfin, la loi de s n'étant pas standard, son échantillonnage est réalisé par un algorithme de Metropolis-Hastings à marche aléatoire avec, comme loi candidate, une gaussienne à support positif de moyenne $s^{(i-1)}$ et de variance r_s fixée. L'efficacité de l'algorithme dépend en grande partie de r_s : si la variance est trop faible, la chaîne de Markov risque de converger trop lentement puisque ses pas seront petits ; au contraire, si la variance est trop grande, l'algorithme va rejeter un grand nombre de propositions [3, 8]. En général, ce réglage est heuristique. Dans [3], Gelman *et al.* recommandent par exemple de chercher à atteindre un taux d'acceptation de 0,5 pour les modèles de dimension 1.

Enfin, afin d'accélérer la convergence de l'algorithme, les $\mathbf{z}_n$ sont échantillonnés dans un ordre aléatoire à chaque itération.

4 Choix de l'estimateur

Dans cette section, nous nous concentrons sur le choix de l'estimateur du signal impulsionnel, les autres quantités étant estimées au sens de l'EAP (espérance a posteriori) car facile à implémenter à partir d'une chaîne de Markov [1]. On suppose donc que l'échantillonneur de Gibbs a généré les chaînes de Markov de $\mathbf{x}$ et $\mathbf{q}$. Pour simplifier, nous supposons que la période de transition de la chaîne de Markov est négligeable ou a été supprimée : on considère donc que la chaîne est en régime stationnaire de $i = 1$ à $i = I$. Dans un premier temps, nous présentons les deux estimateurs proposés dans [1] et [9] et les appliquons à la chaîne de Markov générée, et nous montrons qu'ils peuvent, dans certains cas, conduire à une interprétation erronée. Nous proposons alors deux indices permettant d'obtenir une interprétation correcte des résultats.

Pour cela, considérons l'exemple d'une chaîne de Markov telle que $I = 4$ et $N = 4$ (figure 1). En $k = 1$, la chaîne de Markov ne génère jamais de pic ; en $k = 2$, elle en génère toujours un ; enfin, elle génère alternativement un pic en $k = 3$ et en $k = 4$, signifiant qu'il y a sans doute un pic situé entre les points 3 et 4, mais qu'il est difficile de le positionner avec précision. Les résultats des différents estimateurs sont également représentés, avec éventuellement l'estimation de $\mathbf{q}_n$.

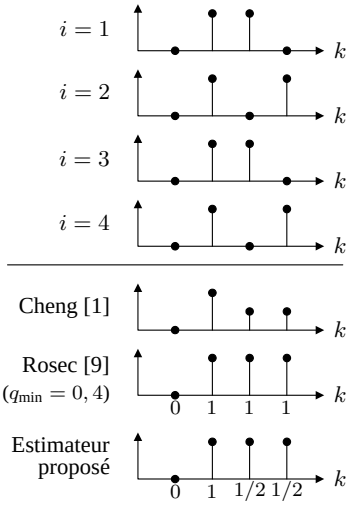


FIG. 1 – Comparaison des différents estimateurs

Dans [1], les auteurs ne considèrent pas le vecteur $\mathbf{q}$ des occurrences des pics et proposent d'estimer $\mathbf{x}_n$ au sens de l'EAP :

$$\widehat{\mathbf{x}}_n^{\text{EAP}} = \mathbb{E}[\mathbf{x}_n|\mathbf{y}] = \int \mathbf{x}_n p(\mathbf{x}_n|\mathbf{y}) d\mathbf{x}_n.$$

Cette quantité n'étant pas calculable explicitement, l'estimateur est approché par :

$$\widehat{\mathbf{x}}_n \triangleq \frac{1}{I}\sum_{i=1}^{I} \mathbf{x}_n^{(i)}.$$

Le résultat fourni par cet estimateur sur l'exemple de la figure 1 donne un résultat cohérent en $k = 1$ et $k = 2$. En revanche, l'analyse de l'estimation indique qu'il y a deux pics de faible amplitude en $k = 3$ et $k = 4$, alors qu'il n'y en a qu'un seul grand : cet estimateur est donc inadéquat.

Rosec *et al.* [9], pour leur part, prennent en compte le vecteur $\mathbf{q}$ et l'estime au sens du MAP (maximum a posteriori) : $\widehat{\mathbf{q}}_n$ ne peut donc prendre que la valeur 0 ou 1.

$$\widehat{\mathbf{q}}_n \triangleq \begin{cases} 1 & \text{si } \frac{1}{I}\sum_{i=1}^{I} \mathbf{q}_n^{(i)} > q_{\min},\\ 0 & \text{sinon,}\end{cases}$$

où $q_{\min}$ est un seuil fixé a priori. L'estimateur de l'amplitude est défini ainsi :

$$\widehat{\mathbf{x}}_n \triangleq \begin{cases} \dfrac{1}{\sum_{i=1}^{I}\mathbf{q}_n^{(i)}} \displaystyle\sum_{\substack{i=1\\ \mathbf{q}_n^{(i)}=1}}^{I} \mathbf{x}_n^{(i)} & \text{si } \widehat{\mathbf{q}}_n = 1,\\ \dfrac{1}{\sum_{i=1}^{I}(1-\mathbf{q}_n^{(i)})} \displaystyle\sum_{\substack{i=1\\ \mathbf{q}_n^{(i)}=0}}^{I} \mathbf{x}_n^{(i)} & \text{si } \widehat{\mathbf{q}}_n = 0,\end{cases}$$

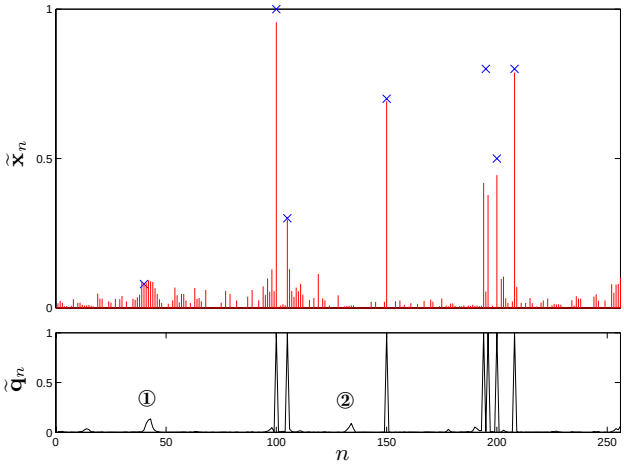


FIG. 2 – Estimation sur un signal simulé. Les croix représentent les pics réels et les lignes leur estimation.

L'estimation à partir de la chaîne de Markov de la figure 1 estime correctement les deux premiers pics, mais, avec $q_{\min} < 0,5$, estime deux pics de grande amplitude en $k = 3$ et $k = 4$ alors qu'il n'y en a qu'un seul.

La principale limitation des estimateurs précédents provient de la prise de décision sous-jacente à l'estimateur. Dans le premier cas, tous les pics sont considérés comme étant significatifs, alors que dans le deuxième, seuls les plus probables sont pris en compte. Afin d'améliorer l'analyse de la chaîne de Markov, nous proposons d'utiliser conjointement deux indices estimant respectivement la probabilité d'occurrence d'un pic et son amplitude :

$$\widetilde{\mathbf{q}}_n \triangleq \frac{1}{I}\sum_{i=1}^{I} \mathbf{q}_n^{(i)}, \qquad \widetilde{\mathbf{x}}_n \triangleq \frac{1}{\sum_{i=1}^{I} \mathbf{q}_n^{(i)}} \sum_{\substack{i=1 \\ \mathbf{q}_n^{(i)}=1}}^{I} \mathbf{x}_n^{(i)}.$$

Notons que pour $I \to \infty$, ces quantités s'écrivent également :

$$\widetilde{\mathbf{q}}_n \approx \mathbb{E}[\mathbf{q}_n|\mathbf{y}], \qquad \widetilde{\mathbf{x}}_n \approx \mathbb{E}[\mathbf{x}_n|\mathbf{y}]/\mathbb{E}[\mathbf{q}_n|\mathbf{y}].$$

L'indice $\widetilde{\mathbf{q}}_n$ n'est pas à proprement parler un estimateur de $\mathbf{q}_n$ car il peut prendre des valeurs non entières. Cependant, il permet d'obtenir une estimation de la *probabilité d'occurrence des pics*, donc de leur probabilité d'existence. Ainsi, dans le cas de la figure 1, les deux premiers pics sont toujours correctement estimés, et en $k = 3$ et $k = 4$ on estime deux pics de grande amplitude, mais de probabilité d'existence 1/2, ce qui reflète mieux le comportement de la chaîne de Markov. L'analyse conjointe des deux indices permet de déduire qu'il n'y a qu'un seul pic situé au barycentre des deux estimés (c'est-à-dire en $k = 3,5$), et dont l'amplitude est la somme des amplitudes $\widetilde{\mathbf{x}}_n$ pondérées par leur probabilité d'occurrence $\widetilde{\mathbf{q}}_n$. Ces deux indices permettent ainsi un analyse plus fine de la chaîne de Markov les estimateurs proposés par [1] et [9].

Afin de valider notre méthode et d'illustrer la pertinence de l'estimateur, nous avons simulé un signal de $N = 256$ points qui est la convolution d'un signal impulsionnel et d'une lorentzienne de largeur $s = 2$. Le RSB a été fixé à 30 dB environ. Le résultat d'une simulation de 1000 itérations est donné figure 2. La largeur de la réponse impulsionnelle a été estimée à $\widehat{s} = 1,954$ et l'estimation obtenue est relativement bonne : les pics aux alentours de 100, 150 et 200 sont bien retrouvés. En ① et ②, on trouve un ensemble de pics d'occurrence 10 % environ. Il y a donc probablement un pic aux alentours de ces endroits, dont l'amplitude est bien estimée par la somme pondérée des $\widetilde{\mathbf{x}}$ et la position par le barycentre des $\widetilde{\mathbf{q}}$, sous réserve que la zone considérée soit correctement sélectionnée. En raisonnant de la sorte, on retrouve bien le pic de faible amplitude en ①. En revanche, l'amplitude des pics estimés en ② est très faible : on peut en déduire qu'il n'y a pas de pic à cet endroit. Une analyse telle que celle présentée dans [1] et [9] n'aboutirait pas à ces conclusions. Le pic situé à 195 a été estimé par deux pics *très proches* d'occurrence 1, et d'amplitude moitié. Encore une fois, en supposant que la zone relative à ce pic soit correctement sélectionnée, sa position et son amplitude sont bien estimés par le barycentre et la somme pondérée. Il est clair que la qualité de l'analyse est très fortement liée à la sélection de la zone relative à un pic, étape qui à l'heure actuelle est laissée à l'appréciation de l'utilisateur.

5 Conclusion

Nous avons présenté dans cet article une méthode de déconvolution impulsionnelle positive myope, dans le cas où la réponse impulsionnelle est de structure connue. De plus, nous nous sommes restreint au cas courant où les pics sont positifs. Un a priori Bernoulli-gaussien à support positif a permis de modéliser le signal impulsionnel, et, en se plaçant dans un cadre bayésien, nous avons utilisé l'échantillonneur de Gibbs pour simuler les lois a posteriori de chaque variable. Enfin, nous avons proposé deux nouveaux indices qui, utilisés conjointement, permettent d'obtenir une analyse plus fine des résultats fournis par l'algorithme de déconvolution.

Références

[1] Q. Cheng, R. Chen et T.-H. Li, Simultaneous wavelet estimation and deconvolution of reflection seismic signals, *IEEE Transactions on Geoscience and Remote Sensing* **34** (1996) 377–384.

[2] L. Devroye, *Non-uniform random variate generation*, Springer-Verlag, New York (1986). Disponible sur http ://jeff.cs.mcgill.ca/˜luc/rnbookindex.html.

[3] A. Gelman, G.O. Roberts et W.R. Gilks, Efficient Metropolis jumping rules, *Bayesian Statistics* **5** Oxford University Press (1996) 599–607.

[4] S. Gulam Razul, W.J. Fitzgerald et C. Andrieu, Bayesian model selection and parameter estimation of nuclear emission spectra using RJMCMC, *Nucl. Inst. Meth. Phys. Res. A* **497** (2003) 492–510.

[5] N.M. Haan et S.J. Godsill, Bayesian model for DNA sequencing, *ICASSP'02*, Orlando, Floride, USA, (2002) **4** 4020–4023.

[6] V. Mazet, D. Brie et J. Idier, Simuler une distribution normale positive à partir de plusieurs lois candidates, *20e GRETSI*, Louvain-la-Neuve, Belgique, 6-9 septembre 2005.

[7] C.P. Robert, Simulation of truncated normal variables, *Statistics and Computing* **5** (1995) 121–125.

[8] C.P. Robert, *Méthodes de Monte Carlo par chaînes de Markov*, Economica (1996).

[9] O. Rosec, J.-M. Boucher, B. Nsiri et T. Chonavel, Blind marine seismic deconvolution using statistical MCMC methods, *IEEE Journal of Oceanic Engineering* **28** (2003) 502–512.

Identification aveugle de système acoustique et application à la déréverbération de la parole

W. BOBILLET[1], E. GRIVEL[1], R. GUIDORZI[2] et M. NAJIM[1]

[1] Equipe Signal et Image, UMR 51 31 LAPS,
351 Cours de la libération, 33405 Talence Cedex, France
[2] Dipartimento di Elettronica, Informatica e Sistemistica, Università di Bologna,
Viale del Risorgimento 2, 40136 Bologna, Italy

email: {william.bobillet, eric.grivel, mohamed.najim}@laps.u-bordeaux1.fr, rguidorzi@deis.unibo.it

Résumé *– Cet article traite de la déréverbération aveugle d'un signal de parole fondée sur deux microphones. Les réverbérations entre le locuteur et les microphones sont modélisées par des filtres à Réponse Impulsionnelle Finie (RIF). Les versions réverbérées du signal de parole sont perturbées par des bruits blancs additifs, dont les variances ne sont pas nécessairement les mêmes. Nous proposons une version modifiée de la méthode initialement développée dans le domaine du contrôle par Guidorzi et al., et qui permet d'estimer de façon aveugle les RIF, ainsi que les variances des bruits additifs. Elle tire profit des propriétés des matrices d'autocorrélation des versions réverbérées de la parole et des observations bruitées.*

Abstract *– This paper deals with the blind speech dereverberation issue when using two microphones. The reverberations between the source signal and the microphones are modelled by Finite Impulse Responses (FIR). The reverberated versions of speech are then contaminated by additive white noises whose variances could be different. The method we propose is a modified version of the identification technique developed by Guidorzi et al. in the framework of control. It makes it possible to estimate both the FIR and the variances of the additive noises and is based on the properties of the autocorrelation matrices of both the reverberated versions of the speech and the noisy observations.*

1. Introduction

Dans la plupart des applications modernes de transmission de la parole, telles les communications sans fil, les vidéoconférences ou les systèmes de reconnaissance de la parole, diverses sources de bruit dégradent le signal utile. Ainsi, une étape de rehaussement est souvent nécessaire pour améliorer l'intelligibilité de la parole.

Lorsque l'on considère des environnements clos, il faut prendre en compte l'acoustique environnant le locuteur. En effet, le signal de parole peut subir de nombreuses réverbérations en suivant de multiples trajets avant d'atteindre le microphone. Ces réverbérations entraînent une distorsion spectrale du signal qui peut être modélisée par des filtres à Réponse Impulsionnelle Finie (RIF).

Dans le contexte monovoie, l'utilisation des statistiques d'ordre deux ne permet de retrouver que des RIF à minimum de phase [2]. Pour cette raison, les statistiques d'ordre supérieur ont été exploitées [15] [16]. Néanmoins, ces méthodes nécessitent l'utilisation d'un grand nombre d'échantillons et reposent sur l'hypothèse de stationnarité du signal sur de longues fenêtres temporelles, ce qui n'est pas le cas pour le signal de parole.

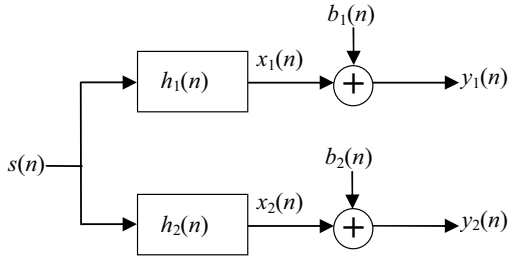


FIG 1 : parole réverbérée et bruitée, captée par deux microphones

Ainsi, l'utilisation de plusieurs microphones est souvent nécessaire pour l'égalisation aveugle de la parole $s(n)$ (Cf. FIG 1), et l'on se place ainsi dans un contexte SIMO. Parmi les méthodes d'estimation des RIF [12] [8], de nombreuses techniques en sous-espace ont été développées [14] [17]. Leur objectif est d'estimer les RIF à partir du noyau de la matrice d'autocorrélation des versions réverbérées $x_i(n)$ du signal de parole. Dans [7], l'algorithme dit EVAM (EigenVector-based Algorithm for Multichannel blind deconvolution) permet d'estimer les fonctions de transfert ainsi que leur ordre. Afin de diminuer le coût calculatoire de l'identification des RIF, Huang et al. [10] ont proposé une approche adaptative minimisant l'erreur quadratique moyenne calculée à partir des Relations Croisées[1] entres les versions réverbérées de la parole. La convergence de l'algorithme étant lente, les auteurs ont étendu la méthode au domaine fréquentiel [11]. Récemment, Gannot et al. [6] ont développé une approche fondée sur une décomposition en valeurs singulières généralisées. Néanmoins, cette méthode nécessite soit l'estimation de la matrice d'autocorrélation des bruits additifs pendant les trames de silence, soit une hypothèse de bruits blancs de même variance.

On s'intéresse dans cet article à la déréverbération d'un signal de parole dans le cas où les bruits additifs sont blancs mais ont des variances différentes. On dispose pour cela de deux microphones. L'approche proposée tire profit de la méthode d'identification de RIF développée par Guidorzi et al. [5]. Cette technique permet d'estimer non seulement les RIF, mais aussi les variances des bruits additifs. Nous proposons en particulier une nouvelle version de l'algorithme présenté dans [3]. Les critères proposés dans [4] et [5] sont ainsi modifiés et permettent d'améliorer l'estimation des RIF dans le contexte du rehaussement de la parole.

Cet article est organisé comme suit : dans la partie 2, le système SIMO est introduit ainsi que la méthode d'identification des RIF. Les critères modifiés sont également présentés. Une étude comparative de ces critères est menée dans la partie 3.

[1] $(x_1 * h_2)(n) - (x_2 * h_1)(n) = 0$

2. Identification du système SIMO

2.1 Le modèle SIMO

Les réverbérations étant modélisées par des RIF d'ordre L, notées $h_{\mathrm{i}}(n)$, le signal de parole $s(n)$ et sa i$^{\text{ème}}$ version réverbérée $x_{\mathrm{i}}(n)$, avec $\mathrm{i}=1,2$, satisfont la relation suivante :

$$x_{\mathrm{i}}(n)=(s*h_{\mathrm{i}})(n)=\sum_{k=0}^{L}h_{\mathrm{i}}(k)s(n-k) \qquad (1)$$

La séquence d'observation $y_{\mathrm{i}}(n)$ résulte de la somme de la version réverbérée $x_{\mathrm{i}}(n)$ et du bruit additif $b_{\mathrm{i}}(n)$ de variance σ_{i}^2 (Cf. FIG 1) :

$$y_{\mathrm{i}}(n)=x_{\mathrm{i}}(n)+b_{\mathrm{i}}(n). \qquad (2)$$

Dans la suite, on considèrera les vecteurs suivants, définis pour $l>0$:

$$\underline{\mathrm{s}}(n)=\left[s(n-L)\;\; s(n-L+1)\;\; \cdots\;\; s(n+l)\right]^{\mathrm{T}} \qquad (3)$$

$$\underline{\mathrm{x}}(n)=\left[x_1(n)\;\; \cdots\;\; x_1(n+l)\;\; x_2(n)\;\; \cdots\;\; x_2(n+l)\right]^{\mathrm{T}} \qquad (4)$$

$\underline{\mathrm{y}}(n)$ et $\underline{\mathrm{b}}(n)$ sont construits de façon similaire à $\underline{\mathrm{x}}(n)$. On introduit en outre la matrice de type Sylvester généralisée :

$$\mathrm{H}_l=\begin{bmatrix}\mathrm{H}_l^1\\ \mathrm{H}_l^2\end{bmatrix} \text{ où :} \qquad (5)$$

$$\mathrm{H}_l^{\mathrm{i}}=\begin{bmatrix}h_{\mathrm{i}}(L) & h_{\mathrm{i}}(L-1) & \cdots & h_{\mathrm{i}}(0) & 0\\ & \ddots & \ddots & & \ddots\\ 0 & h_{\mathrm{i}}(L) & h_{\mathrm{i}}(L-1) & \cdots & h_{\mathrm{i}}(0)\end{bmatrix}_{(l+1)\times(l+\mathrm{L}+1)} \qquad (6)$$

On peut alors réécrire les relations (1) et (2) sous la forme matricielle suivante :

$$\underline{\mathrm{x}}(n)=\mathrm{H}_l\underline{\mathrm{s}}(n), \qquad (7)$$

$$\underline{\mathrm{y}}(n)=\underline{\mathrm{x}}(n)+\underline{\mathrm{b}}(n). \qquad (8)$$

L'estimation des RIF est fondée sur l'estimation du noyau de la matrice d'autocorrélation des échantillons $\underline{\mathrm{x}}(n)$, $\mathrm{R}_x^l=\frac{1}{N-l}\sum_{k=0}^{N-l-1}\underline{\mathrm{x}}(k)\underline{\mathrm{x}}^{\mathrm{T}}(k)$. En effet, lorsque les conditions d'identifiabilité données dans [9] sont satisfaites, on a :

$$\mathrm{Null}\left(\mathrm{R}_x^l\right)=\mathrm{Span}\left(\mathrm{M}_l^{\mathrm{T}}\right) \qquad (9)$$

où

$$\mathrm{M}_l=\left[\mathrm{H}_l^2\;\; -\mathrm{H}_l^1\right]. \qquad (10)$$

Cependant, R_x^l n'est pas disponible et doit être estimée.

2.2 Estimation de R_x^l et des RIF

Les bruits additifs sont supposés stationnaires et non corrélés avec le signal de parole. Ainsi, si l'on travaille avec un grand nombre N d'échantillons, on obtient :

$$\mathrm{R}_y^l=\mathrm{R}_x^l+\mathrm{R}_b^l \qquad (11)$$

et

$$\mathrm{R}_b^l=\mathrm{diag}\left(\sigma_1^2\,\mathrm{I}_{l+1},\sigma_2^2\,\mathrm{I}_{l+1}\right)=\begin{bmatrix}\sigma_1^2\,\mathrm{I}_{l+1} & 0\\ 0 & \sigma_2^2\,\mathrm{I}_{l+1}\end{bmatrix}. \qquad (12)$$

où R_y^l et R_b^l sont calculées de manière analogue à R_x^l.

L'estimation de R_x^l est fondée sur les propriétés suivantes :

i) R_x^l *semi définie positive, i.e.* $\mathrm{R}_x^l\geq 0$.

ii) R_y^l *est définie positive, i.e.* $\mathrm{R}_y^l>0$.

ainsi que sur le théorème suivant [1] :

Le couple $P^*=\left(\sigma_1^2,\sigma_2^2\right)$ *est l'unique solution du système d'équations* :

$$\mathrm{R}_x^l(\alpha,\beta)=\mathrm{R}_y^l-\mathrm{R}_b^l(\alpha,\beta)\geq 0 \;\; \textit{pour}\;\; l\geq L \qquad (13)$$

$$\text{avec}\;\; \mathrm{R}_b^l(\alpha,\beta)=\mathrm{diag}\left(\alpha\,\mathrm{I}_{l+1},\beta\,\mathrm{I}_{l+1}\right). \qquad (14)$$

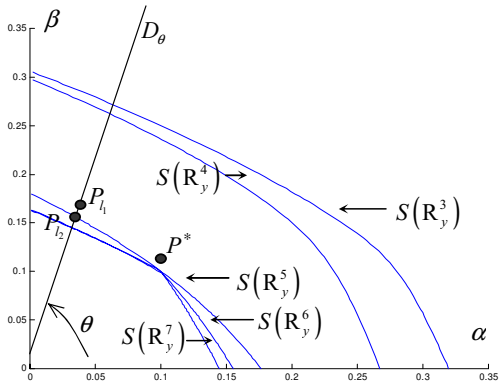


FIG 2 : courbes convexes $S(\mathrm{R}_y^l)$ pour $L=5$ et $l=3,..,7$

L'ensemble des solutions de chaque équation (13) est une courbe convexe dans le plan (α,β), notée $S(\mathrm{R}_y^l)$. Dans la suite, nous utiliserons les notations ci-après :

P_l est le point d'intersection entre $S(\mathrm{R}_y^l)$ et la droite passant par l'origine D_θ, de pente $\tan(\theta)$ (Cf. FIG 2). La matrice $\mathrm{R}_x^l(\alpha,\beta)$ résultante, définie par la relation (13), ne dépend alors que de l'angle θ et sera notée $\mathrm{R}_x^l(\theta)$.

L'estimation des variances des bruits additifs revient théoriquement à déterminer le point commun P^* à toutes les courbes $S(\mathrm{R}_y^l)$ définies pour $l\geq L$, ou encore θ^* l'angle correspondant. Cependant, en pratique, P^* n'existe pas. Ainsi, Guidorzi et al. ont introduit différents critères $J_{critère}(\theta)$ dont la minimisation le long de $S(\mathrm{R}_y^L)$ fournit l'estimation $\hat{\theta}$ de θ^*, et donc l'estimation des variances $\left(\sigma_1^2,\sigma_2^2\right)$ et des RIF définies par le vecteur $\underline{\mathrm{h}}$ suivant :

$$\underline{\mathrm{h}}=\left[h_1(0)\;\; \cdots\;\; h_1(L)\;\; h_2(0)\;\; \cdots\;\; h_2(L)\right]^{\mathrm{T}} \qquad (15)$$

Les valeurs de $J_{critère}(\theta)$ sont calculées à partir de l'estimation $\hat{\underline{\mathrm{h}}}(\theta)$ de $\underline{\mathrm{h}}$ donnée par le noyau de la matrice $\mathrm{R}_x^L(\theta)$. En effet, pour $l=L$, la relation (9) s'écrit :

$$\mathrm{Null}\left(\mathrm{R}_x^L\right)=\mathrm{Span}\left(\left[h_2(L)\cdots h_2(0)-h_1(L)\cdots-h_1(0)\right]^{\mathrm{T}}\right). \qquad (16)$$

Cependant, la taille élevée des matrices considérées rend difficile la détermination du noyau. En effet, les valeurs propres de R_x^L peuvent être très proches et il est alors difficile d'identifier celles associées au signal et celles relatives au bruit. Pour remédier à ce problème, nous proposons d'une part d'estimer les RIF en tenant compte de la

structure du noyau des matrices R_x^l pour $l > L$. D'autre part, nous proposons de généraliser les critères $J_{critère}(\theta)$ donnés dans [4] et [5] aux courbes $S(\mathrm{R}_y^l)$ d'ordre $l > L$.

2.3 Estimation alternative des RIF

D'après la relation (9), on a :

$$\mathrm{R}_x^l \mathrm{M}_l^{\mathrm{T}} = 0 \tag{17}$$

L'estimation $\hat{\underline{\mathrm{h}}}$ de $\underline{\mathrm{h}}$ vérifie alors la relation suivante :

$$\hat{\underline{\mathrm{h}}}(\theta) = \underset{\|\underline{\mathrm{u}}\|_2=1}{\arg\min}\left\{C(\underline{\mathrm{u}},\theta)\right\}. \tag{18}$$

où $C(\underline{\mathrm{u}},\theta)$ est une forme quadratique dépendant des $(2L+2)$ variables du vecteur $\underline{\mathrm{u}}$, définie par :

$$C(\underline{\mathrm{u}},\theta) = \mathrm{Trace}\left(\mathrm{M}_l(\underline{\mathrm{u}})\mathrm{R}_x^l(\theta)\mathrm{M}_l^{\mathrm{T}}(\underline{\mathrm{u}})\right) \tag{19}$$

et minimisée sous la contrainte $\|\underline{\mathrm{u}}\|_2 = 1$.

On obtient alors l'algorithme suivant.

Algorithme

i) Pour un angle $\theta \in \left[0;\frac{\pi}{2}\right]$, déterminer le point P_{l_1} de $S(\mathrm{R}_y^{l_1})$ et si nécessaire le point P_{l_2} de $\mathrm{S}(\mathrm{R}_y^{l_2})$, avec $l_2 > l_1 \geq L$. Cf. FIG 2.

ii) Calculer la matrice $\mathrm{R}_x^{l_1}(\theta)$.

iii) Estimer les RIF $\hat{\underline{\mathrm{h}}}(\theta)$ à l'aide de la relation (18).

iv) Evaluer le critère $J_{critère}(\theta)$. La valeur minimale permet alors d'obtenir les estimées $(\hat{\sigma}_1^2, \hat{\sigma}_2^2)$ et $\hat{\underline{\mathrm{h}}}$.

2.4 Critères modifiés

2.4.1 Critère de covariance de l'entrée (ICM, Input Covariance Matching criterion)

L'objectif ici est de comparer les matrices $\mathrm{R}_x^{l_1}(\theta)$ et $\mathrm{R}_x^{l_2}(\theta)$ en tenant compte de la structure du système bivoie. Pour cela, on introduit la matrice d'autocorrélation des échantillons de l'entrée $\underline{\mathrm{s}}(n)$, de dimensions $(l+L+1)\times(l+L+1)$:

$$\mathrm{R}_s^l = \frac{1}{N-l}\sum_{k=0}^{N-l-1} \underline{\mathrm{s}}(k)\underline{\mathrm{s}}^{\mathrm{T}}(k), \text{ pour } l \in \{l_1, l_2\}. \tag{20}$$

D'après la relation (7), les matrices R_x^l et R_s^l vérifient :

$$\mathrm{R}_x^l = \mathrm{H}_l \mathrm{R}_s^l \mathrm{H}_l^{\mathrm{T}}. \tag{21}$$

Les conditions d'identifiabilité [9] étant supposées vérifiées, la matrice H_l est de rang plein et donc la relation (21) est inversible :

$$\mathrm{R}_s^l = \mathrm{H}_l^{+}\mathrm{R}_x^l \mathrm{H}_l^{\mathrm{T}+} \tag{22}$$

où l'exposant $^+$ représente la pseudo inverse de Moore-Penrose. Or, on peut écrire la matrice R_s^l sous la forme :

$$\mathrm{R}_s^l = \begin{bmatrix} \mathrm{R}_s^L & \mathrm{D}_l \\ \mathrm{D}_l^{\mathrm{T}} & \mathrm{C}_l \end{bmatrix}. \tag{23}$$

La matrice R_s^L, de dimensions $(2L+1)\times(2L+1)$, ne dépend pas de l, et peut donc être calculée soit à partir de $\mathrm{R}_x^{l_1}$, soit à partir de $\mathrm{R}_x^{l_2}$. Le critère modifié ICM vise donc à minimiser la différence entre les matrices $\mathrm{R}_s^L(\theta,l_1)$ et $\mathrm{R}_s^L(\theta,l_2)$ qui sont les estimations de R_s^L calculées respectivement à partir de $\mathrm{R}_x^{l_1}(\theta)$ et $\mathrm{R}_x^{l_2}(\theta)$.

$$J_{ICM}(\theta) = \left\|\mathrm{R}_s^L(\theta,l_1) - \mathrm{R}_s^L(\theta,l_2)\right\|_{Fro} \tag{24}$$

où $\|.\|_{Fro}$ désigne la norme de Frobenius.

2.4.2 Critère de décalage du noyau (SR, Shifted Relations criterion)

Puisque la relation (9) est valable pour tout $l > L$, l'équation suivante :

$$\mathrm{R}_x^{l_2}(\theta)\mathrm{M}_{l_2}^{\mathrm{T}}(\theta) = 0 \tag{25}$$

a pour solution $\theta = \theta^*$, *i.e.* $P_{l_1} = P_{l_2} = P^*$. Le critère SR modifié est alors donné par la relation suivante :

$$J_{SR}(\theta) = \left\|\mathrm{R}_s^{l_2}(\theta)\mathrm{M}_{l_2}^{\mathrm{T}}(\theta)\right\|_{Fro}. \tag{26}$$

2.4.3 Critère de Covariance de l'Erreur (ECM, Error Covariance Matching criterion)

L'objectif est ici de minimiser la norme de la matrice d'autocovariance de l'erreur suivante, définie à l'aide des Relations Croisées :

$$\begin{aligned} e(n) &= \left(h_1(\theta) * y_2\right)(n) - \left(h_2(\theta) * y_1\right)(n) \\ &= e_x(n) + e_b(n) \end{aligned} \tag{27}$$

où $e_x(n)$ et $e_b(n)$ sont définis comme suit :

$$\begin{aligned} e_x(n) &= \left(h_1(\theta) * x_2\right)(n) - \left(h_2(\theta) * x_1\right)(n) \\ e_b(n) &= \left(h_1(\theta) * b_2\right)(n) - \left(h_2(\theta) * b_1\right)(n) \end{aligned} \tag{28}$$

Les bruits additifs $b_1(n)$ et $b_2(n)$ étant des processus blancs à moyenne nulle et non corrélés avec le signal de parole, on montre que les matrices d'autocovariance de $e(n)$, $e_x(n)$ et $e_b(n)$, respectivement notées $\mathrm{R}_e(\theta)$, $\mathrm{R}_{e_x}(\theta)$ et $\mathrm{R}_{e_b}(\theta)$, vérifient la relation suivante :

$$\mathrm{R}_e(\theta) = \mathrm{R}_{e_x}(\theta) + \mathrm{R}_{e_b}(\theta) \tag{29}$$

De plus, d'après l'equation (28), $e_b(n)$ est la somme de deux processus à Moyenne Ajustée (MA) d'entrée blanche $b_1(n)$. On a alors la relation suivante :

$$\mathrm{R}_{e_b}(\theta) = \alpha \mathrm{H}_L^1(\theta)\mathrm{H}_L^{1\mathrm{T}}(\theta) + \beta \mathrm{H}_L^2(\theta)\mathrm{H}_L^{2\mathrm{T}}(\theta). \tag{30}$$

Lorsque $\theta = \theta^*$, il vient $e_x(n) = 0$, et donc on a :

$$\mathrm{R}_{e_x}(\theta^*) = 0. \tag{31}$$

Le nouveau critère ECM est donc défini comme suit :

$$J_{ECM}(\theta) = \left\|\mathrm{R}_{e_x}(\theta)\right\|_{Fro} = \left\|\mathrm{R}_e(\theta) - \mathrm{R}_{e_b}(\theta)\right\|_{Fro}. \tag{32}$$

3. Simulations et conclusions

Nous présentons dans cette partie une étude comparative entre les différents critères, originaux et modifiés.

Puisque les RIF ne sont estimées qu'à une constante multiplicative près, leur évaluation ne doit pas dépendre de cette constante. Ainsi, comme $\underline{h}$ et $\hat{\underline{h}}$ définissent des directions dans un espace de dimension $(2L+2)$, nous introduisons l'erreur quadratique suivante [13] :

$$\xi\left(\underline{h},\hat{\underline{h}}\right)=\sin^2(\phi)=\min_{\gamma}\frac{\left\|\underline{h}-\gamma\hat{\underline{h}}\right\|^2}{\left\|\underline{h}\right\|^2}=1-\left(\frac{\underline{h}^{\mathrm{T}}\hat{\underline{h}}}{\left\|\underline{h}\right\|\left\|\hat{\underline{h}}\right\|}\right)^2$$

avec $\phi\in\left[0;\frac{\pi}{2}\right]$ l'angle formé par les droites engendrées par $\underline{h}$ et $\hat{\underline{h}}$. Le critère d'évaluation des RIF est ainsi la Racine de l'Erreur Quadratique Moyenne :

$$\mathrm{REQM}=\sqrt{\frac{1}{R}\sum_{k=1}^{R}\xi\left(\underline{h},\hat{\underline{h}}_k\right)}$$

où $R=100$ est le nombre de simulations réalisées et $\hat{\underline{h}}_k$ est la $k^{ième}$ estimation de $\underline{h}$.

Les simulations sont réalisées sur un signal de parole de 5000 échantillons, filtré par des RIF synthétiques d'ordre $L=100$. Les versions réverbérées $x_i(n)$ sont alors perturbées par des bruits additifs blancs gaussiens, avec un Rapport Signal à Bruit (RSB) allant de 20 dB à 40 dB. Les variances vérifient $\tan\left(\theta^*\right)=\frac{\sigma_1^2}{\sigma_2^2}=1.82$. Les résultats sont présentés dans les Tableau 1-3. Différentes valeurs de l_1 et l_2 sont testées. Les versions originales des critères correspondent au cas $l_1=L$ et $l_2=L+1$ (Cf. Tableau 1). Les résultats présentés sont respectivement le taux de convergence (TC) de l'algorithme, la REQM et son écart type, calculés sur les réalisations convergentes de l'algorithme.

Tableau 1 : TC, REQM et écart type. l_1=L, et l_2=L+1.

	RSB=20 dB	RSB=30 dB	RSB=40 dB
ICM original	100% 0.0964 ± 0.0490	100% 0.0321 ± 0.0209	100% 0.0078 ± 0.0041
SR original	84% 0.1187 ± 0.1105	96% 0.0368 ± 0.0228	100% 0.0079 ± 0.0041
ECM original	88% 0.1162 ± 0.0959	98% 0.0335 ± 0.0232	100% 0.0079 ± 0.0041

Tableau 2 : TC ; REQM et écart type. l_1=L+3, et l_2=L+4.

	RSB=20 dB	RSB=30 dB	RSB=40 dB
ICM modifié	100% 0.0917 ± 0.0480	100% 0.0302 ± 0.0194	100% 0.0079 ± 0.0042
SR modifié	94% 0.1086 ± 0.0600	100% 0.0314 ± 0.0209	100% 0.0079 ± 0.0041
ECM modifié	94% 0.1091 ± 0.0686	100% 0.0311 ± 0.0210	100% 0.0079 ± 0.0040

Tableau 3 : TC, REQM et écart type. l_1=L+9, et l_2=L+10.

	RSB = 20 dB	RSB = 30 dB	RSB = 40 dB
ICM modifié	100% 0.0735 ± 0.0387	100% 0.0303 ± 0.0191	100% 0.0078 ± 0.0042
SR modifié	96% 0.1114 ± 0.0539	100% 0.303 ± 0.0191	100% 0.0079 ± 0.0042
ECM modifié	96% 0.1162 ± 0.0600	100% 0.0320 ± 0.0194	100% 0.0078 ± 0.0043

Pour des RSB entre 20 dB et 30 dB, les critères modifiés permettent d'améliorer l'estimation de RIF. Le taux de convergence est inférieur à 88% pour les versions originales des critères SR et ECM. Il augmente jusqu'à 96% en utilisant les versions modifiées (Cf. Tableau 3). On peut noter aussi une diminution de la REQM et plus particulièrement de son écart type pour les trois critères.

Bien que le coût calculatoire de la méthode proposée soit élevé, les critères modifiés tiennent compte de la structure du noyau de R_x^l et améliorent la qualité des estimations des RIF, et par conséquent la qualité du processus de déréverbération.

Références

[1] S. Beghelli, R.P.Guidorzi U.Soverini, The Frisch Scheme in Dynamic System Identification, Automatica 26, 1990.

[2] J. Benesty, S. Makino, J.Chen (eds.), Speech Enhancement, Springer, chap. 11-12, 2005.

[3] W. Bobillet, E. Grivel, R. Guidorzi and M. Najim, Cancelling convolutive and additive coloured noises for speech enhancement, Proc. IEEE-ICASSP 2004, Montreal, 2004.

[4] P. Castaldi, R. Diversi, R.P. Guidorzi, U. Soverini, Blind Estimation and Deconvolution of Communication Channels with Unbalanced Noise, Proc. of the 12th IFAC Symposium on System Identification, Santa Barbara, CA, June 2000.

[5] R. Diversi, R. Guidorzi, U. Soverini, Blind Identification and Equalization of Two-Channel FIR Systems in Unbalanced Noise Environments, Signal Processing, vol. 85, no. 1, January 2005.

[6] S. Gannot, M. Moonen, Subspace Method for Multimicrophone Speech Dereverberation, EURASIP Journal on Applied Signal Processing, Special Issue on Signal Processing for Acoustic Communication Systems, vol. 2003, no. 11, pp. 1074-1090, October 2003.

[7] M.İ. Gürelli, C.L. Nikias, EVAM: An Eigenvector-Based Algorithm for Multichannel Blind Deconvolution of Input Colored Signals, IEEE Trans. on SP, vol. 43, no. 1, February 1995.

[8] Y. Hua, Fast Maximum Likelihood for Blind Identification of Multiple FIR Channels, IEEE Trans. on SP, vol. 44, no. 3, March 1996.

[9] Y. Hua and M. Wax, Strict Identifiability of Multiple FIR Channels Driven by an Unknown Arbitrary Sequence, IEEE Trans. on SP, vol. 44, no. 3, March 1996.

[10] Y.A. Huang, J. Benesty, Adaptive Multi-Channel Least Mean Square and Newton Algorithms for Blind Channel Identification, Signal Processing, vol. 82, no.8, pp. 1127-1138, August 2002.

[11] Y.A. Huang, J. Benesty, Class of Frequency-Domain Adaptive Approaches to Blind Multichannel Identification, IEEE Trans. on SP, vol. 51, no. 1, January 2003.

[12] H. Liu, G. Xu, and L. Tong, A Deterministic Approach to Blind Identification of Multi-Channel FIR Systems, Proc. IEEE ICASSP '94, pp. 581-584, April 1994.

[13] D.R. Morgan, J. Benesty, M. Sondhi, On the Evaluation of Estimated Impulse Responses, IEEE Signal Processing Letters, vol. 5, no. 7, March 1998.

[14] E. Moulines, P. Duhamel, J. Cardoso and S. Mayrargue, Subspace Methods for the Blind Identification of Multichannel FIR Filters, IEEE Trans. on SP, vol. 43, no. 2, February 1995.

[15] O. Shalvi and E. Weinstein, New Criteria for Blind Deconvolution of Nonminimum Phase Systems (Channels), Trans. on IT, vol. 36, no. 2, March 1990.

[16] J.K Tugnait, Identification of Linear Stochastic System via Second and Fourth-Order Cumulant Matching, , IEEE Trans. on IT, vol. 33, no. 3, May 1987.

[17] G. Xu, H. Liu, L. Tong, T. Kailath, , A Least Squares Approach to Blind Channel Identification, IEEE Trans. on SP, vol. 43, no. 12, December 1995.

Le lissage H_∞ : une alternative au filtrage de Kalman pour le rehaussement du signal de parole ?

D. Labarre[1], E. Grivel[1], N. Christov[2], M. Najim[1]

[1] Equipe Signal & Image, UMR LAPS 5131, ENSEIRB BP 99, F-33402 Talence, France

Email : {david.labarre ; eric.grivel ; mohamed.najim}@laps.u-bordeaux1.fr

2 Laboratoire d'Automatique, Génie Informatique et Signal, USTL, Cité Scientifique, 59655 Villeneuve d'Ascq, France

Email : nicolai.christov@univ-lille1.fr

Résumé *– La mise en oeuvre d'un filtre de Kalman repose sur des hypothèses fortes sur le processus générateur et le bruit d'observation du système. En effet, ils doivent être tous les deux blancs et gaussiens et leurs variances connues a priori. De plus, les incertitudes liées à la modélisation du système ne sont pas prises en compte dans la pratique. Pour ces deux raisons, nous proposons d'étudier l'apport des techniques d'estimation H_∞ dans le contexte du rehaussement du signal de parole fondé sur un modèle AutoRégressif (AR). Ces approches, qui consistent à minimiser les effets des perturbations sur l'erreur d'estimation du signal dans le pire cas, sont a priori attrayantes car elles ne nécessitent aucune hypothèse statistique ni sur le processus générateur ni sur le bruit d'observation. En outre, comme la qualité perceptuelle du signal rehaussé dépend de l'étape d'estimation des paramètres AR, nous proposons ici une estimation conjointe du signal de parole et des paramètres AR. Plus particulièrement, un lissage H_∞ de type retard fixe fournit une estimation du signal de parole à partir de la dernière estimation des paramètres AR, alors que les paramètres AR sont mis à jour avec un filtrage H_∞ à partir de la dernière estimation lissée du signal de parole.*

Abstract *– In real cases, the use of a Kalman filter for signal processing may be questionable. Indeed, it does not take into account modelling uncertainties and requires a priori system information difficult to be estimated. In this paper, our purpose is to evaluate how relevant the H_∞ filters are in the framework of speech processing, where the driving process for instance is usually unknown. Indeed, these H_∞ techniques, which consist in minimizing the worst possible effects of the disturbances on the estimation error, seem appealing because no statistical information on the driving process and the additive observation noise of the system are required. However, the auditory quality of the enhanced speech still depends on the way the AR parameters are estimated because they set the spectral features of the signal to be estimated. Therefore, we propose a new method based on a cross-coupled algorithm involving two estimators running in parallel. A fixed-lag H_∞ smoothing provides an estimate of the signal, while a H_∞ filter updates the AR parameters. At time k, the signal is estimated from the latest updates of the AR parameters, while the AR parameters are estimated from the latest enhanced version of the signal.*

1. Introduction

Dans de nombreuses applications du traitement du signal comme le traitement de la parole ou l'égalisation, des méthodes paramétriques fondées sur un modèle *a priori* du système sont souvent utilisées. Parmi les techniques d'estimation, le filtre de Kalman [1] présente l'avantage d'être récursif et optimal au sens des moindres carrés. Néanmoins, la mise en oeuvre de ce filtre repose sur des hypothèses fortes sur le processus générateur et le bruit d'observation du système. En effet, ils doivent être tous les deux blancs et gaussiens et leurs variances connues *a priori*. De plus, avec cette approche, les incertitudes liées à la modélisation du système ne sont pas prises en compte.

Pour pallier ces inconvénients, nous proposons d'étudier des approches alternatives fondées sur une estimation de type H_∞. Dans ce cas, l'objectif est de minimiser les effets des perturbations sur l'erreur d'estimation du signal dans le pire cas. Cette approche est à première vue séduisante car aucune hypothèse statistique sur la représentation du système n'est nécessaire.

Bien que la théorie H_∞ ait été particulièrement étudiée dans le domaine de l'automatique, les applications en traitement du signal restent peu nombreuses. Ainsi, dans [8], Sayyarrodsari *et al.* traitent le problème de l'annulation de bruit. Un filtre H_∞ est mis en oeuvre pour mettre à jour les paramètres du système, plutôt qu'un algorithme LMS habituellement utilisé. Une adaptation plus rapide et une estimation plus lisse des paramètres sont alors obtenues. Les techniques d'optimisation H_∞ ont été également employées pour l'égalisation de canaux de communications numériques [3] et la conception de bancs de filtres [10]. Dans ces deux cas, les auteurs concluent que les approches H_2 et H_∞ donnent des résultats similaires en moyenne.

Ici, nous nous focalisons sur le problème du rehaussement d'un signal de parole s_k perturbé par un bruit additif n_k, à partir d'une unique séquence d'observations y_k définie comme suit :

$$y_k = s_k + n_k \,. \tag{1}$$

Le signal de parole s_k est modélisé par le processus AutoRégressif (AR) suivant :

$$s_k = -\sum_{i=1}^{p} a_i s_{k-i} + \alpha_k \tag{2}$$

où p désigne l'ordre du modèle, $\underline{\theta} = [a_1 \ \cdots \ a_p]^T$ sont les paramètres AR et le processus générateur α_k est blanc, gaussien, de moyenne nulle et de variance σ_α^2.

Dans [6], nous avons initié notre étude sur la pertinence du filtrage H_∞ en considérant deux protocoles de simulation.

Dans le premier, les paramètres du modèle sont estimés à partir du signal de parole non bruité, ce qui permet de caractériser le comportement du filtre H_∞ face aux incertitudes de modélisation. Le second consiste à estimer les paramètres AR à partir des observations bruitées afin d'analyser le comportement du filtre H_∞ vis-à-vis des erreurs d'estimation sur les paramètres AR. Les résultats obtenus nous ont permis de conclure que l'estimation des paramètres AR reste une étape clé dans la procédure de rehaussement car elle conditionne les propriétés spectrales du signal rehaussé.

Dès lors, l'approche proposée par Shen *et al*. dans [9], et déclinée dans le contexte des communications mobiles [2], est mise en défaut. En effet, pour estimer conjointement un signal et les paramètres du modèle AR associé à partir des observations bruitées, les auteurs utilisent deux filtres H_∞ opérant en série. Les paramètres AR sont tout d'abord estimés à partir des observations bruitées par filtrage H_∞. Puis, ils sont utilisés dans le second filtrage pour retrouver les données. Bien que les auteurs justifient leur démarche par le manque d'information *a priori*, ils négligent l'étape d'estimation des paramètres AR.

Pour pallier ce problème, nous proposons ici d'utiliser deux estimateurs récursifs qui opèrent en parallèle et interagissent comme suit :

- Un lissage H_∞ de type retard fixe (*fixed-lag* en anglais) permet d'estimer le signal de parole en utilisant la dernière valeur de l'estimation des paramètres AR.
- Un filtrage H_∞ permet de mettre à jour l'estimation des paramètres AR en utilisant la dernière estimation lissée du signal de parole. Cf. Fig. 1.

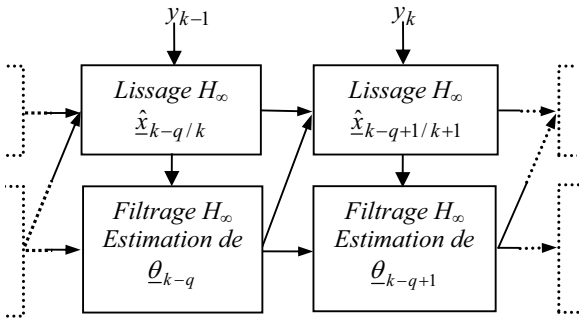


FIG. 1 : principe de la méthode proposée

L'article est organisé comme suit : la partie 2 traite du filtrage H_∞, la partie 3 de l'estimation des paramètres AR et la partie 4 de l'ajustement des matrices de poids intervenant dans les équations du filtrage et du lissage H_∞. La dernière partie présente des résultats de simulations sur l'estimation de paramètres AR et le rehaussement du signal de parole.

2. Filtrage et lissage H_∞

Outre les approches polynomiales [4], les solutions au problème du filtrage H_∞ actuellement les plus étudiées sont fondées sur une Représentation dans l'Espace d'Etat (REE) de la forme suivante :

$$\begin{cases} \underline{x}_k = \Phi_k \underline{x}_{k-1} + \Gamma \alpha_k \\ y_k = H \underline{x}_k + n_k \\ s_k = L \underline{x}_k \end{cases} \tag{3}$$

où $\underline{x}_k$ désigne le vecteur d'état, Φ_k la matrice de transition, Γ le vecteur d'entrée, H le vecteur d'observation et L un vecteur permettant de se focaliser sur une combinaison linéaire particulière des variables d'état, ici s_k.

Ces approches ont l'avantage de conduire à une implantation plus simple du filtre comparée aux approches polynomiales. Etant donné un niveau d'atténuation γ, l'estimation du signal $\hat{s}_k$ doit satisfaire le critère suivant :

$$J_\infty = \sup_{\alpha_k, n_k} \frac{\sum_{k=0}^{N-1} \| s_k - \hat{s}_k \|_2^2}{\sum_{k=1}^{N} \left[\|\alpha_k\|_2^2 + \|n_k\|_2^2 \right]} < \gamma^2 \tag{4}$$

où N est le nombre d'observations y_k disponibles.

Nous nous intéressons ici à une solution fondée sur une équation de Riccati [5]. Dans ce cas, l'estimation $\hat{s}_k$ vérifiant la condition (4) existe si et seulement si :

$$(P_{k+1/k})^{-1} + H^T H - \gamma^{-2} L^T L > 0 \tag{5}$$

où la matrice $P_{k+1/k}$ satisfait la récursion de Riccati suivante :

$$P_{k+1/k} = \Phi_k P_{k/k} \Phi_k^T + \Gamma Q \Gamma^T \tag{6a}$$

$$P_{k/k} = P_{k/k-1} - P_{k/k-1} \begin{bmatrix} H^T & L^T \end{bmatrix} T^{-1} \begin{bmatrix} H \\ L \end{bmatrix} P_{k/k-1} \tag{6b}$$

où $T = \begin{bmatrix} R & 0 \\ 0 & -\gamma^2 \end{bmatrix} + \begin{bmatrix} H \\ L \end{bmatrix} P_{k/k-1} \begin{bmatrix} H^T & L^T \end{bmatrix}$ et Q et R sont des matrices de pondération permettant de régler le comportement du filtre H_∞.

Les estimations *a posteriori* du vecteur d'état et du signal sont alors remises à jour comme suit :

$$\hat{s}_k = L \underline{\hat{x}}_{k/k} \tag{7a}$$

$$\underline{\hat{x}}_{k/k} = \underline{\hat{x}}_{k/k-1} + K_k \upsilon_k \tag{7b}$$

$$\underline{\hat{x}}_{k/k-1} = \Phi_k \underline{\hat{x}}_{k-1/k-1} \tag{7c}$$

où $\underline{\hat{x}}_{k/l}$ désigne l'estimation de $\underline{x}_k$ étant données l observations, $K_k = P_{k/k-1} H \left(R + H P_{k/k-1} H^T \right)^{-1}$ le gain du filtre et $\upsilon_k = y_k - H \underline{\hat{x}}_{k/k-1}$ le processus d'innovation.

Un lissage H_∞ à retard fixe consiste alors à estimer le signal à l'instant $k-q$ étant données k observations, où q définit le retard. Dans la suite, on supposera que $q \geq p$. En construisant le vecteur d'état comme suit :

$$\underline{x}_k = \left[s_k \cdots s_{k-q} \right]^T, \tag{8}$$

les matrices de la REE (3) sont les suivantes :

$$H = \Gamma^T = [1 \quad 0 \quad \cdots \quad 0],$$

$$L = [0 \quad \cdots \quad 0 \quad 1]$$

et $$\Phi_k = \begin{bmatrix} -a_1 & \cdots & -a_p & 0 & \cdots & 0 \\ 1 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & 1 & 0 \end{bmatrix}.$$

Le filtrage H_∞ défini par les équations (6)-(7) permet d'obtenir les versions filtrées et lissées du signal de parole de la manière suivante :

$$\hat{s}_{k-q/k-q} = H\,\underline{\hat{x}}_{k-q/k-q} \quad (9)$$

$$\hat{s}_{k-q/k} = L\,\underline{\hat{x}}_{k/k} \quad (10)$$

Remarques :
1. Les filtres de Kalman et H_∞ ont la même structure. De plus, lorsque $\gamma \to \infty$, les deux filtres coïncident.
2. Le coût calculatoire du filtre H_∞ est sensiblement plus élevé que celui du filtre de Kalman.
3. Les approches fondées sur la théorie des jeux [11] peuvent être considérées comme un problème d'estimation H_∞ car elles traitent d'un problème de type minimax.

Comme les paramètres AR ne sont pas disponibles, nous proposons maintenant de les estimer avec un second filtre récursif appliqué à la version lissée $\hat{s}_{k-q/k}$ du signal.

3. Estimation des paramètres AR

Sur une trame d'analyse, le signal est supposé stationnaire et les paramètres AR sont constants :

$$\underline{\theta}_{k-q} = \underline{\theta}_{k-q-1}. \quad (11)$$

On exprime alors l'estimation lissée du signal $\hat{s}_{k-q/k}$ en fonction des paramètres AR $\underline{\theta}_{k-q}$. Etant donnée (7), si $\{k-i\}_{i=0,\ldots,q-1}$ observations sont disponibles, on montre aisément que :

$$\hat{s}_{k-q/k} = \hat{s}_{k-q/k-q} + \sum_{i=1}^{q} K_{k-q+i}^{i+1} \upsilon_{k-q+i}. \quad (12)$$

où K_l^m désigne la $m^{\text{ième}}$ composante du vecteur gain du filtre à l'instant *l*. Or, d'après (7) et (9), la version filtrée du signal de parole à l'instant *k-q* satisfait la relation suivante :

$$\begin{aligned} \hat{s}_{k-q/k-q} &= H\left(\Phi_{k-q}\,\underline{\hat{x}}_{k-q-1/k-q-1} + K_{k-q}\upsilon_{k-q}\right) \\ &= -\left(\underline{\hat{x}}^{1:p}_{k-q-1/k-q-1}\right)^T \underline{\theta}_{k-q} + K_{k-q}^1 \upsilon_{k-q} \end{aligned}, \quad (13)$$

où $\underline{\hat{x}}^{1:p}_{k/k}$ correspond aux *p* premières composantes de $\underline{\hat{x}}_{k/k}$. Ainsi, en substituant (12) dans (13), on obtient :

$$\hat{s}_{k-q/k} = -\left(\underline{\hat{x}}^{1:p}_{k-q-1/k-q-1}\right)^T \underline{\theta}_{k-q} + \upsilon_k^\theta, \quad (14)$$

avec $\upsilon_k^\theta = \sum_{i=0}^{q} K_{k-q+i}^{i+1} \upsilon_{k-q+i}$.

Les paramètres AR sont alors estimés avec un filtre H_∞ fondé sur la REE (11)-(14). Ce filtre met en jeu deux matrices de pondération R^θ et Q^θ et un niveau d'atténuation γ^θ.

Les niveaux d'atténuation des bruits γ et γ^θ peuvent être par exemple fixés à l'aide d'une procédure d'optimisation itérative. L'ajustement des matrices de pondération Q, R, Q^θ et R^θ est plus délicat et nous proposons d'étudier ce dernier point dans le paragraphe 4.

4. Réglage des matrices de poids

Les matrices de pondération Q, R, Q^θ et R^θ permettent d'ajuster les comportements des estimateurs H_∞. Elles jouent le même rôle que les variances du processus générateur et du bruit d'observation respectivement dans le cadre H_2, comme le mentionnent Shen *et al.* dans [9]. En outre, la matrice $P_{k+1/k}$ peut être interprétée comme une borne supérieure de la matrice de covariance de l'erreur dans le cadre H_2 [11].

Nous proposons donc d'adapter ces matrices en raisonnant par analogie au cadre H_2. Nous supposons tout d'abord que les caractéristiques du bruit additif d'observation n_k varient lentement au cours du temps et l'ajustement de *R* est réalisé pendant les trames de silence.

La matrice *Q* est réglée selon les résultats présentés dans [7] de la manière suivante :

$$Q_k = \frac{k-1}{k} Q_{k-1} + \frac{1}{k} D M_k D^T. \quad (15)$$

avec $M_k = P_{k/k} - \Phi_{k-1} P_{k-1/k-1} \left(\Phi_{k-1}\right)^T + K_k \upsilon_k^2 K_k^T$ et $D = \left(\Gamma^T \Gamma\right) \Gamma^T$.

Dans le filtre de Kalman [1], l'innovation υ_k est un bruit blanc de variance $HP_{k/k-1}H^T + R$. Ainsi, nous proposons d'ajuster la matrice R^θ de la manière suivante :

$$R_k^\theta = \sum_{i=0}^{q} \left\{ \left(K_{k-q+i}^{i+1}\right)^2 \left(HP_{k-q+i/k-q+i-1}H^T + R\right) \right\}. \quad (16)$$

Enfin, d'après l'équation (11), la valeur de Q^θ est fixée à 0.

5. Résultats expérimentaux

5.1 Estimation de paramètres AR

Dans un premier temps, nous étudions les performances de l'approche proposée pour l'estimation de paramètres AR, car cette étape conditionne le rehaussement du signal de parole.

Lorsque le processus AR est perturbé par un bruit additif, Shen *et al.* [9] appliquent directement un filtre H_∞ sur les observations bruitées pour estimer les paramètres AR. Nous réalisons donc une étude comparative entre le filtre de Kalman, le filtre H_∞ et la méthode proposée. Le protocole de l'expérience est le suivant :

512 échantillons d'un processus AR d'ordre 6 sont générés selon les pôles $p_{1,2} = 0.98e^{\pm j0.1\pi}$, $p_{3,4} = 0.97e^{\pm j0.3\pi}$ et $p_{5,6} = 0.8e^{\pm j0.84\pi}$. Ils sont ensuite perturbés par un bruit

blanc gaussien à moyenne nulle avec un Rapport Signal à Bruit (RSB) de 10 dB. 100 réalisations du bruit sont utilisées.

Les résultats du tableau 1 montrent que l'utilisation directe du filtre de Kalman et du filtre H_∞ sur des observations bruitées conduisent à des estimations on-line biaisées et très similaires. La méthode proposée permet de réduire significativement les erreurs d'estimation des paramètres AR.

TAB. 1 : Estimation de paramètres AR à partir d'observations bruitées avec différentes méthodes ; RSB = 10 dB.

Paramètres	*Valeur désirée*	*Filtre de Kalman*	*Filtre H_∞* [9]	*Méthode proposée*
σ_α^2	1	10,11	10,23	3,12
a_1	-1,60	-0,99	-0,99	-1,45
a_2	0,45	0,03	0,04	0,35
a_3	0,87	0,31	0,33	0,68
a_4	-0,51	0,003	-0,01	-0,18
a_5	-0,56	-0,22	-0,23	-0,71
a_6	0,58	0,17	0,18	0,57

5.2 Rehaussement du signal de parole

Le signal /*Waziwaza*/, échantillonné à 16 kHz, est perturbé soit par un bruit blanc gaussien, soit par un bruit coloré à Moyenne Ajustée (MA) large bande. Nous avons mené une étude comparative entre la méthode proposée et celle présentée dans [9].

Même si le filtrage H_∞ permet de réduire l'influence des incertitudes du modèle sur l'estimation du signal de parole, l'estimation des paramètres AR reste une étape cruciale. Notre méthode a l'avantage d'une part de limiter les incertitudes d'estimation sur les paramètres comparées à l'approche présentée dans [9], et d'autre part de lisser le signal rehaussé, atténuant ainsi le bruit résiduel.

TAB. 2 : Gains du RSB : cas d'un bruit blanc gaussien

	RSB d'entrée	*5*	*10*	*15*
Bruit blanc gaussien	*Méthode* [9]	4.4	4.0	3.4
	Méthode proposée	8.6	7.8	5.2

TAB. 3 : Gains du RSB : cas d'un bruit coloré

	RSB d'entrée	*5*	*10*	*15*
Bruit coloré de type MA	*Méthode* [9]	6.1	5.5	3.6
	Méthode proposée	6.4	5.8	3.9

6. Conclusion

Dans cet article, notre objectif est d'évaluer la pertinence de l'estimation H_∞ pour le rehaussement du signal de parole fondé sur un modèle AR. Dans ce cadre, aucune hypothèse statistique n'est nécessaire pour caractériser le processus générateur et le bruit additif. Suite à l'étude menée dans [6], nous proposons une méthode de débruitage fondée sur deux estimateurs H_∞ en parallèle. Un lissage de type retard fixe utilise la dernière estimation des paramètres pour estimer le signal de parole, tandis qu'un filtrage H_∞ utilise la dernière version lissée du signal pour mettre à jour les paramètres AR. Cette approche a l'avantage de réduire les erreurs d'estimation sur les paramètres. Nous avons ainsi réalisé une étude comparative sur des données AR synthétiques qui montre la pertinence de l'approche proposée. Dans le contexte du rehaussement du signal de parole, l'utilisation d'un lissage permet en outre de réduire le bruit résiduel. Ainsi, les résultats expérimentaux d'une comparaison avec la seule méthode H_∞ référencée à notre connaissance confirment l'amélioration de la qualité auditive du signal de parole rehaussé obtenu avec notre approche.

Références

[1] B. D. O. Anderson, J. B. Moore, *Optimal filtering*, Prentice Hall, N.Y., 1979.

[2] J. Cai, X. Shen, J. W. Mark, *Robust Channel Estimation for OFDM Wireless Communication Systems – An H_∞ Approach*, IEEE Trans. on Wireless Communications, vol. 3, n°6, November 2004, pp.2060-2071.

[3] A.T. Erdogan, B. Hassibi, T. Kailath, *On linear H_∞ equalization of communication channels*, IEEE Trans. on SP, vol. 48, n°11, November 2000, pp. 3227-3231.

[4] M. J. Grimble, A. E. Sayed, *Solution of the H_∞ Optimal Linear Filtering Problem for Discrete-Time Systems*, IEEE Trans. on ASSP, vol. 38, n°7, July 1990, pp. 1092-1104.

[5] B. Hassibi, A. H. Sayed, T. Kailath, *Indefinite-Quadratic Estimation and Control, A unified Approach to H_2 and H_∞ Theories*, SIAM edition, 1999.

[6] D. Labarre, E. Grivel, M. Najim, *Relevance of H_∞ filtering for speech enhancement*, accepted at ICASSP'05.

[7] D. Labarre, E. Grivel, M. Najim, E. Todini, *Two-Kalman Filters Based Instrumental Variable Techniques for Speech Enhancement*, Proc. IEEE MMSP, 2004, Siena, Italy.

[8] B. Sayyarrodsari, B. Hassibi, J. How, A. Carrier, *An H∞-optimal alternative to the FxLMS algorithm*, Proc. ACC, 1998, pp.1116-1121.

[9] X. Shen, L. Deng, *A Dynamic System Approach to Speech Enhancement Using the H_∞ Filtering Algorithm*, IEEE Trans. on SAP, vol. 7, n°4, July 1999, pp. 391-399.

[10] H. Vikalo, B. Hassibi, T. Kailath, *On H_∞ optimal signal reconstruction in noisy filter banks*, Proc. ICASSP, 1999, pp. 1493-1496.

[11] I. Yaesh, U. Shaked, *Game Theory Approach to Optimal Linear State Estimation and Its Relation to the Minimum H_∞-Norm Estimation*, IEEE Trans. on AC, vol. 37, n°6, June 1992, pp. 828-831.

Recherche Linéaire Optimisée : Une Nouvelle Méthode pour Accélérer PARAFAC

Myriam RAJIH, Pierre COMON
Laboratoire I3S
2000 route des Lucioles, BP 121, F-06903 Sophia Antipolis, France
rajih@i3s.unice.fr, comon@i3s.unice.fr

Résumé – L'algorithme des moindres carrés alternés (ALS), utilisé dans la décomposition de PARAFAC, peut nécessiter un très grand nombre d'itérations avant de converger. Cette lenteur de convergence peut découler d'une taille importante des données, d'un mauvais conditionnement, ou de la présence de dégénérescences, etc. Plusieurs méthodes ont été proposées pour accélérer l'algorithme, notamment la compression préalable des données [3], et la recherche linéaire (Line Search LS) [3]. Après une brève description de PARAFAC, nous présenterons une nouvelle méthode pour accélérer ALS, qui, comparée aux méthodes existantes, a permis d'obtenir de meilleurs résultats lors des simulations. Nous proposerons ensuite une application de la nouvelle méthode à l'identification aveugle d'un mélange sous-déterminé (UDM), sachant qu'elle peut être appliquée à n'importe quelle décomposition d'ordre N.

1 Introduction

La décomposition de PARAFAC peut être considérée comme une généralisation de l'analyse de facteurs d'ordre deux aux tenseurs d'ordre N (N≥3). Elle fut introduite par Harshman en 1970 [8] qui s'est basé sur le principe des Profils Proportionnels et parallèles (PP) proposé par Cattel en 1944 [5]. Ce principe énonce que si deux (ou plusieurs) modèles d'ordre deux sont décrits par le même ensemble de vecteurs composants, tels que seuls les poids changent d'un modèle à l'autre, ces vecteurs composants conduisent alors à un nouveau modèle unique au sens de la rotation [2]. En d'autres termes, supposons que la matrice $\mathbf{X}_1$ puisse s'écrire sous la forme :

$$\mathbf{X}_1 = \mathbf{a}_1\mathbf{b}_1^T c_{11} + \mathbf{a}_2\mathbf{b}_2^T c_{12} + ... + \mathbf{a}_F\mathbf{b}_F^T c_{1F}$$

$\mathbf{a}_f$ et $\mathbf{b}_f$ ($1 \leq f \leq F$) étant les colonnes des matrices $\mathbf{A}$ et $\mathbf{B}$ respectivement, et supposons qu'une seconde matrice $\mathbf{X}_2$ puisse s'exprimer à l'aide des mêmes vecteurs composants précédents mais avec des poids différents décrits par c_{2j} (au lieu de c_{1j} précédemment) :

$$\mathbf{X}_2 = \mathbf{a}_1\mathbf{b}_1^T c_{21} + \mathbf{a}_2\mathbf{b}_2^T c_{22} + ... + \mathbf{a}_F\mathbf{b}_F^T c_{2F}$$

Alors nous pouvons ainsi construire un nouveau modèle :

$$\mathbf{X}_k = \mathbf{A}diag(\mathbf{C}(k,:))\mathbf{B}^T, k = 1, 2 \quad (1)$$

qui peut être écrit autrement : $X_{ijk} = \sum_f A_{if}B_{jf}C_{kf}$. Ce modèle trilinéaire est aussi connu sous le nom de DECOMPosition CANonique (CANDECOMP) introduite par Caroll et Chang en 1970 [4]. Le modèle PARAFAC d'ordre trois et ses généralisations aux ordres supérieurs [8] [4][12] sont très populaires en psychométrie et en chimie, mais trouvent aussi des applications en traitement du signal [11] [6] [7]. Contrairement au modèle matriciel, le modèle PARAFAC a l'avantage d'être unique sous les conditions du théorème de Kruskal [10], d'où son importance. Il existe plusieurs algorithmes pour décomposer PARAFAC, le plus répandu étant ALS sur lequel nous nous concentrerons dans toute la suite. La convergence de l'algorithme ALS est parfois extrêmement lente. Cela peut se présenter lorsque la taille des données est assez grande, ou encore lorsque deux colonnes composantes sont presque colinéaires. La compression [3] et la recherche linéaire (LS) [2] sont deux solutions possibles pour remédier à ce problème. Nous nous concentrerons dans cet article sur la solution LS et nous proposerons une nouvelle méthode qui a permis de réduire le nombre d'itérations nécessaires à la convergence de ALS.

2 Modèle et notations

Considérons le modèle PARAFAC d'ordre trois défini par l'expression (1). Ce modèle peut être écrit sous une forme compacte en utilisant le produit de Khatri-Rao $\odot$ (produit Kronecker des colonnes):

$$\mathbf{X}^{(I\times JK)} = \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T$$

où les matrices $\mathbf{A}$, $\mathbf{B}$, et $\mathbf{C}$ sont des matrices de taille $I \times F$, $J \times F$, et $K \times F$ respectivement, et $\mathbf{X}^{(I\times JK)}$ est la matrice de taille $I \times JK$ obtenue en dépliant le tenseur $\mathbf{X}$ de taille $I \times J \times K$ suivant le premier mode. ALS consiste à estimer tour à tour une des trois matrices composantes, en minimisant au sens des moindres carrés l'erreur :

$$\Upsilon = \| \mathbf{X}^{(I\times JK)} - \mathbf{A}(\mathbf{C} \odot \mathbf{B})^T \|_F^2$$

où $\| \bullet \|_F$ est la norme de Frobenius. $\mathbf{B}$ et $\mathbf{C}$ étant initialisées, l'estimée de $\mathbf{A}$ est donnée par :

$$\widehat{\mathbf{A}} = \mathbf{X}^{(I\times JK)}(\mathbf{Z}_a^+)^T \quad (2)$$

avec $\mathbf{Z}_a = \mathbf{C} \odot \mathbf{B}$ et $(^+)$ est la pseudo-inverse de Moore-Penrose. Nous estimons les matrices $\mathbf{B}$ et $\mathbf{C}$ de manière équivalente avec $\mathbf{Z}_b = \mathbf{A} \odot \mathbf{C}$ et $\mathbf{Z}_c = \mathbf{B} \odot \mathbf{A}$, et nous itérons jusqu'à satisfaction du critère de convergence. Typiquement, la convergence est atteinte lorsque l'erreur Υ affiche, entre deux itérations successives, un changement inférieur à un seuil prédéfini, lequel varie selon les données. Pour des données simples, le seuil peut être fixé à 10^{-6} par exemple, mais doit être plus petit pour des données plus complexes, 10^{-10} par exemple.

Il arrive que la convergence nécessite plusieurs milliers d'itérations. Un choix judicieux des valeurs initiales peut quelquefois aider à atteindre le minimum global très rapidement. Cependant, quand le tenseur a de grandes dimensions, ou quand deux facteurs sont presque colinéaires, les techniques d'initialisation ne permettent pas d'aller plus vite. Dans [3], Bro propose de compresser les données dans un espace réduit afin de réduire les dimensions du tenseur, et de là réduire la complexité de ALS. La solution LS a aussi été proposée pour accélérer la convergence [9] [2]. Cette solution est décrite dans la section qui suit.

3 Recherche Linéaire (LS)

Il a été noté lors des simulations que, lorsque la convergence est lente, il existe des cycles de convergence définis par une direction unique. Au cours d'un cycle donné, les facteurs composants évoluent dans la même direction jusqu'à la fin du cycle. Il en est de même pour les cycles suivants. La convergence à l'intérieur d'un cycle peut nécessiter plusieurs itérations. Afin de limiter le nombre d'itérations d'un cycle donné, Harshman et Bro proposent d'extrapoler ce comportement. Ils proposent de prédire la valeur du facteur composant après un certain nombre d'itérations en calculant une sorte de régression linéaire :

$$\mathbf{A}^{(new)} = \mathbf{A}^{(it-2)} + R_{LS}(\mathbf{A}^{(it-1)} - \mathbf{A}^{(it-2)}) \qquad (3)$$

$\mathbf{A}^{(it-1)}$ est l'estimée de la matrice $\mathbf{A}$, obtenue à l'itération $(it-1)$ de ALS. A l'itération (it), la nouvelle matrice $\mathbf{A}^{(new)}$ est utilisée à la place de $\mathbf{A}^{(it-1)}$. $(\mathbf{A}^{(it-1)} - \mathbf{A}^{(it-2)})$ définit la direction du cycle. Les matrices $\mathbf{B}$ et $\mathbf{C}$ sont obtenues de façon équivalente en utilisant le même facteur de relaxation R_{LS}. Pour être efficace, l'extrapolation ne doit pas engendrer plus de temps que les itérations qu'elle remplace. C'est justement le cas lorsque R_{LS} est une constante (entre 1.2 et 1.3) [8], ou lorsqu'il est fixé à $it^{1/3}$ [2].

A chaque itération (it), les nouveaux facteurs sont utilisés pour calculer l'erreur :

$$\Upsilon^{(new)} = \| \mathbf{X}^{(I\times JK)} - \mathbf{A}^{(new)}(\mathbf{C}^{(new)} \odot \mathbf{B}^{(new)})^T \|_F^2 \quad (4)$$

Si $\Upsilon^{(new)} \geq \Upsilon^{(it-1)}$, cela veut dire que nous sommes allés trop loin lors de l'extrapolation et que R_{LS} est trop grand. Il est alors décrémenté de $it^{1/n}$ à $it^{1/(n+1)}$, n étant fixé à 3 au début de la simulation, et les facteurs de l'itération $(it-1)$ sont utilisés au lieu des nouveaux : l'itération (it) est effectuée sans extrapolation. Par contre, si $\Upsilon^{(new)} < \Upsilon^{(it-1)}$ alors l'accélération est accomplie et nous gagnons quelques itérations.

LS est appliquée après un nombre d'itérations de l'algorithme ALS, afin de laisser le temps à ce dernier de se stabiliser. Dans [1], ce nombre est fixé à 6, mais il peut être augmenté selon les données.

La faible valeur de R_{LS} pourrait induire que l'accélération n'est pas efficace. Ceci n'est pas vrai car l'effet de R_{LS} est reconduit d'une itération à une autre, donnant lieu en final à une nette diminution du nombre d'itérations comme le montre la figure 4. Le modèle utilisé dans cette simulation est présenté dans la section 5. Le nombre d'itérations nécessaires à la convergence diminue de 10000 à 4200. Il reste toutefois assez élevé, d'où la nécessité de trouver une nouvelle méthode qui réduise significativement la complexité de ALS.

4 Recherche Linéaire Optimisée (ELS)

L'idée principale derrière ELS (Enhanced Line Search) est de chercher un facteur de relaxation R_{LS} optimal qui conduise à la solution finale du cycle en une seule itération. Définissons pour l'itération (it), la direction du cycle pour la matrice composante $\mathbf{A}$ comme $\mathbf{G}_a^{(it)} = \mathbf{A}^{(it-1)} - \mathbf{A}^{(it-2)}$. $\mathbf{G}_b^{(it)}$ et $\mathbf{G}_c^{(it)}$ sont définies de manière équivalente. Au lieu de fixer R_{LS} comme dans l'expression (3), nous cherchons le triplet optimal (R_a, R_b, R_c) qui minimise :

$$\Upsilon_{ELS} = \| \mathbf{X}^{(I\times JK)} - (\mathbf{A}^{(it-2)} + R_a\mathbf{G}_a^{(it)}) \Big((\mathbf{C}^{(it-2)} + R_c\mathbf{G}_c^{(it)}) \odot (\mathbf{B}^{(it-2)} + R_b\mathbf{G}_b^{(it)})\Big)^T \|_F^2 \quad (5)$$

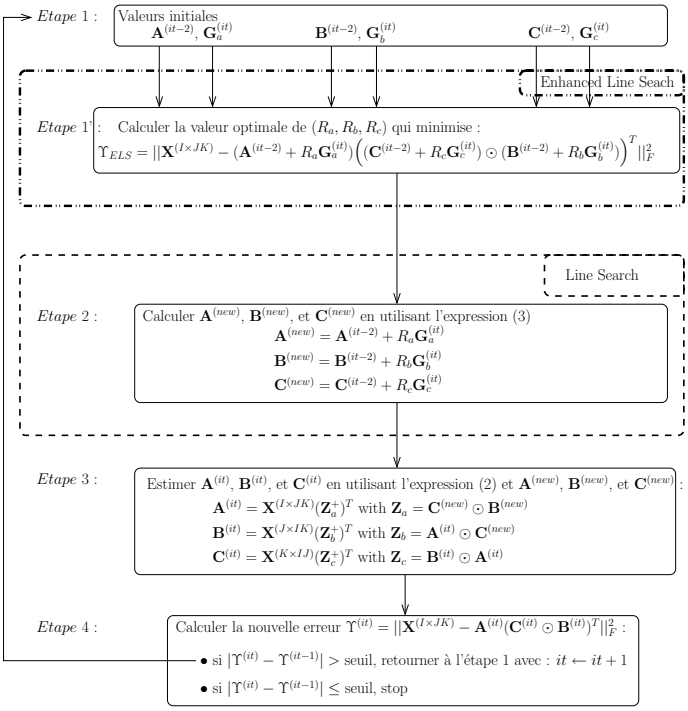


FIG. 1: Etapes de l'algorithme ALS avec la Recherche Linéaire Optimisée.

ELS est appliquée au début de l'algorithme ALS, comme le montre la figure (1), où l'étape 1' correspond à la partie ELS. Le facteur de relaxation R_{LS} n'est plus fixe comme c'était le cas pour LS, mais il est calculé à l'étape 1' de la figure (1), comme étant la valeur optimale qui donne la plus petite erreur Υ_{ELS}.

La solution optimale est obtenue en minimisant conjointement Υ_{ELS} par rapport à R_a, R_b, et R_c. Cela revient à résoudre un système de trois polynômes en trois variables de degré global 5, et engendre une complexité importante. Des solutions à complexité réduite sont obtenues en minimisant Υ_{ELS} par rapport à deux variables seulement, ou encore en prenant le même facteur $R = R_a = R_b = R_c$. Quelques optimisations possibles sont citées ci-après :

1. (R_a, R_b, R_c) donne la solution optimale
2. (R, R, R_c) où nous utilisons le même facteur R pour $\mathbf{A}$ and $\mathbf{B}$ et nous minimisons Υ_{ELS} par rapport aux deux variables R et R_c
3. (R, R, R) où nous utilisons le même facteur pour les trois matrices
4. $R(R_b, R_c)$ où nous utilisons le facteur de relaxation $R = it^{1/3}$ de LS pour la matrice $\mathbf{A}$, et nous minimisons (5) par rapport à R_b et R_c
5. $R(R, R)$ est le même que $R(R_b, R_c)$ avec $R_b = R_c$
6. $R, R(R)$ où nous optimisons par rapport à R_c uniquement

Pour s'assurer que l'extrapolation est moins coûteuse que les itérations qu'elle remplace nous avons calculé les complexités de ALS et de l'optimisation 3 précédente. A chaque itération ALS, nous exécutons les étapes suivantes :

- Calculer le facteur de relaxation optimal R en minimisant l'expression (5). Pour cela, deriver (5) par rapport à R, et résoudre le polynôme de degré 5 en une inconnue obtenu

- Calculer les nouveaux facteurs $\mathbf{A}^{(new)}$, $\mathbf{B}^{(new)}$, et $\mathbf{C}^{(new)}$ en utilisant l'expression (3), et calculer l'erreur correspondante Υ_{new} donnée par l'expression (4)

- Utiliser $\mathbf{A}^{(new)}$, $\mathbf{B}^{(new)}$, et $\mathbf{C}^{(new)}$ comme valeurs initiales pour l'itération PARAFAC au lieu de $\mathbf{A}^{(it-1)}$, $\mathbf{B}^{(it-1)}$, et $\mathbf{C}^{(it-1)}$, et estimer le premier facteur $\widehat{\mathbf{A}}$ en utilisant la formule (2)

- Exécuter la troisième l'étape de nouveau pour estimer les autres facteurs $\widehat{\mathbf{B}}$ et $\widehat{\mathbf{C}}$, en utilisant $\widehat{\mathbf{A}}$ et $\mathbf{C}^{(new)}$ pour estimer $\widehat{\mathbf{B}}$, et $\widehat{\mathbf{A}}$ et $\widehat{\mathbf{B}}$ pour estimer $\widehat{\mathbf{C}}$

Une itération (ALS+ELS) coûte environ $(F+8F^2)(JK+IK+IJ)+3FIJK+11F^3+3F^2+FIJK+5^3 = \mathbf{2240}$ multiplications, lorsque $I=2$, $J=3$, $K=3$, et $F=3$. Le détail des calcul est reporté en annexe. (ALS) seul engendre $(F+8F^2)(JK+IK+IJ)+3FIJK+11F^3+3F^2 = \mathbf{2061}$ multiplications, ce qui est équivalent en terme d'ordre de grandeur.

Lorsque ALS convergence rapidement, disons en moins de 1000 itérations, ELS n'est d'aucune utilité. Cependant, lorsque la convergence est très ralentie, ELS fait la différence, et cela à complexité équivalente. Lorsque les dimensions du tenseur d'ordre N sont très grandes et d'ordre $O(I)$ ($F << I$), les complexités de (ALS) et de (ALS+ELS) sont équivalentes et égales à $O(FI^N)$. C'est aussi le cas lorsqu'une dimension I_0 est beaucoup plus grande comparée aux autres dimensions puisque les deux algorithmes engendrent une complexité $O(I_0)$.

Il est intéressant de noter que l'erreur $\Upsilon^{(new)}$ est toujours inférieure à $\Upsilon^{(it-1)}$ quand les valeurs optimales de R_a, R_b, et R_c sont utilisées, comme dans les trois premières optimisations. Cependant, lorsque le facteur de relaxation est fixé, comme dans la méthode LS, $\Upsilon^{(new)}$ peut être supérieure à $\Upsilon^{(it-1)}$, ce qui veut dire que l'accélération peut échouer. Ceci explique la meilleure performance de ELS face à LS car l'accélération est garantie à chaque itération. Quelques optimisations possibles de ELS ont été implémentées. Les résultats de simulation montrent que ELS est très attrayante.

5 Résultats

Nous avons reporté sur les figures 2 et 3 l'impact de ELS sur ALS dans le cas de Dégénérescence de 2 Facteurs (2FD), où deux facteurs composants sont presque colinéaires de façon à ce que seule la contribution de l'un d'eux est prise en compte [2]. La présence de 2FD ralentit la convergence et peut engendrer l'apparition d'intervalles de convergence lente (swamps), où l'erreur Υ a besoin d'un très grand nombre d'itérations afin d'enregistrer une très faible diminution.

Considèrons le modèle PARAFAC d'ordre trois de l'expression (1) avec :

$$\mathbf{A} = \begin{pmatrix} 1 & cos(\theta) & 0 \\ 0 & sin(\theta) & 1 \end{pmatrix}$$

$$\mathbf{B} = \begin{pmatrix} 3 & \sqrt{2}cos(\theta) & 0 \\ 0 & sin(\theta) & 1 \\ 0 & sin(\theta) & 0 \end{pmatrix}$$

et $\mathbf{C}$ est la matrice identité de taille 3×3. La colinéarité est contrôlée à travers la variable θ qui est fixée à $\pi/60$.

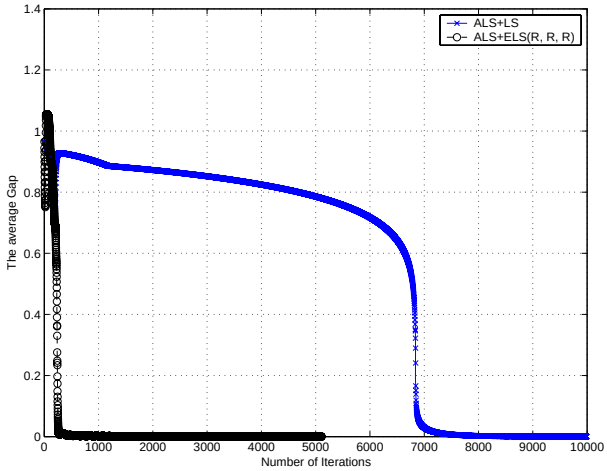


FIG. 2: Gap entre la vraie matrice composante $\mathbf{A}$ et son estimée en utilisant (ALS+LS) et (ALS+ELS) avec l'optimisation (R,R,R) pour $\theta = \pi/60$.

Nous constatons à partir des figures 2 et 3 que ELS accélère la convergence puisqu'elle fait décroître le nombre d'itérations de 7100 à 350 !

Comme signalé dans le résumé, nous proposons une application de ELS à l'identification aveugle d'un UDM. Nous utilisons ELS pour accélérer ALESCAF, l'algorithme proposé en [7] qui consiste en l'identification aveugle d'un UDM basée sur la fonction caractéristique. En utilisant les notations de [7], ALESCAF conduit à un modèle PARAFAC d'ordre quatre :

$$\mathbf{T}^{(P\times KP^2)} = \mathbf{A}(\mathbf{D} \odot \mathbf{A} \odot \mathbf{A})^T$$

Le tenseur $\mathbf{T}$ contient les dérivées d'ordre trois de la fonction caractéristique conjointe des observations, calculée en K points de la grille Ω. La matrice $\mathbf{D}$ est obtenue grâce à la propriété d'indépendance des sources, et ses entrées sont définies par :

$$D_{kn} = \psi_n^{(3)}(\sum_q A_{qn}u_q[k])$$

où $1 \leq k \leq K$ et $1 \leq n \leq N$. $\mathbf{A}$ est la matrice de canal de taille 2×3 à identifier.

Nous utilisons l'implémentation de ALS proposée par Andersson et Bro dans [1], et nous remplaçons la partie LS par les six optimisations de ELS citées dans la figure 4. Les trois sources sont BPSK et nous générons un bloc infini de données en prenant les 2^3 combinaisons possibles de $\{-1, 1\}$. Le bruit n'est pas pris en compte, et le nombre maximum d'itérations est fixé à 5000.

Nous reportons sur la figure 4 le gap entre la vraie matrice de canal et son estimée pour six optimisations de ELS, et nous les comparons avec (ALS+LS) et ALS non accéléré (sans LS). La figure montre, une fois de plus, que ELS est indispensable pour accélérer la convergence, puisque le nombre d'itérations nécessaires pour atteindre la convergence passe de 4200 avec (ALS+LS), à 2900 avec l'optimisation $R_{LS}(R,R,R)$ de ELS.

6 Conclusions

Nous avons présenté ELS, une nouvelle méthode pour accélérer l'algorithme ALS utilisé dans la décomposition du modèle PARAFAC, et nous l'avons comparée avec les méthodes

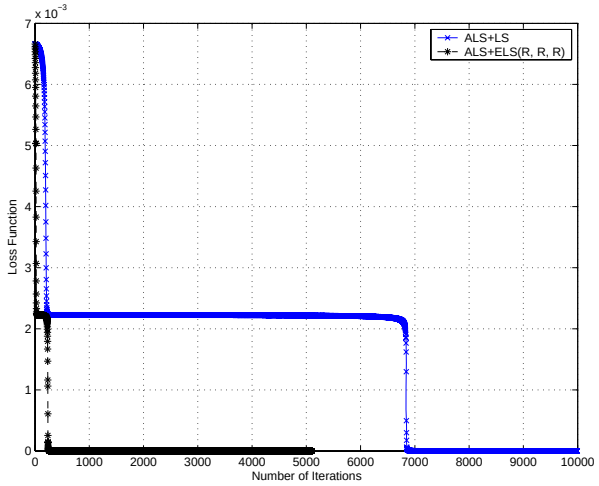


FIG. 3: L'erreur function Υ en fonction du nombre d'itérations pour ALS avec LS, et les optimisations (R, R, R) et (R, R, R_c) de ELS, avec $\theta = \pi/60$.

existantes. Les simulations ont montré que ELS est très attrayante, particulièrement dans le cas de 2FD où elle a permis de diviser le nombre d'itérations par **20**! L'application de ELS à l'identification aveugle d'un UDM a aussi donné de bons résultats puisque ELS a accéléré la convergence.

Dans les travaux futurs, le bruit sera pris en compte et une plus large classe de blocs de données sera considérée. Nous utiliserons des tenseurs à grandes dimensions afin de confirmer que ELS est aussi attractive dans ce cas.

7 Annexe

Pendant une itération de ALS, les opérations suivantes sont exécutées (la matrice $\widehat{\mathbf{A}}$ est estimée en premier) :

1. Calculer le produit Khatri-Rao pour obtenir la matrice $\mathbf{Z}_a$. Cette opération coûte FJK multiplications

2. Calculer $\mathbf{Z}_a^+$ par SVD réduite de $\mathbf{Z}_a$, ce qui nécessite $7JKF^2 + \frac{11}{3}F^3$ multiplications

3. Estimer le facteur $\widehat{\mathbf{A}}$ grâce à l'expression (2), ce qui nécessite $IJKF + JKF^2 + F^2$ multiplications

Les étapes précédentes sont exécutées pour l'estimation de chacun des trois facteurs, ce qui implique $(F + 8F^2)(JK + IK + IJ) + 3FIJK + 11F^3 + 3F^2$ multiplications pour une itération complète de ALS. La complexité générée par ELS est $FIJK + O((2N-1)^3)$ lorsque nous choisissons l'optimisation 3. précédente pour un model PARAFAC d'ordre 3 (N=3).

En général, pour un tenseur d'ordre N de dimension $I_1 \times I_2 \times ... \times I_N$, la complexité générée par ALS lorsque toutes les dimensions sont du même ordre $O(I)$ est : $\frac{11}{3}NF^3 + NF^2 + NFI^N + 8NF^2I^{N-1} + NF\frac{I^2(1-I^{N-2})}{1-I}$. Parallèlement, ELS nécessite $FI^N + O((2N-1)^3)$ multiplications. On en déduit que ALS et (ALS+ELS) sont à complexité équivalente puisqu'ils nécessitent tous les deux $O(FI^N)$ multiplications, lorsque les dimensions sont grandes et de même ordre de grandeur $O(I)$. $O(I_0)$ multiplications sont nécessaires lorsqu'une dimension I_0 est très grande par rapport aux autres dimensions.

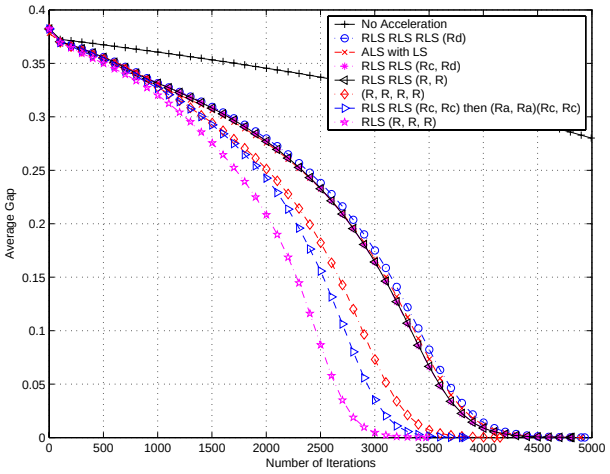


FIG. 4: Gap entre la vraie matrice de canal et son estimée pour $(P, N) = (2, 3)$, en utilisant l'algorithme ALESCAF.

References

[1] C. A. ANDERSSON, R. BRO, "The n-way toolbox for matlab", *Chemometrics and Intelligent Laboratory Systems*, vol. 52, no. 9, pp. 1–4, Sept. 2000.

[2] R. BRO, *Multi-way Analysis in the Food Industry : Models, Algorithms, and Applications*, PhD thesis, University of Amsterdam, Amsterdam, 1998.

[3] R. BRO, C. A. ANDERSSON, "Improving the speed of multiway algorithms. part ii : Compression", *Chemometrics and Intelligent Laboratory Systems*, , no. 42, pp. 105–113, 1998.

[4] J. D. CARROLL, J. J. CHANG, "Analysis of individual differences in multidimensional scaling via n-way generalization of Eckart-Young decomposition", *Psychometrika*, vol. 35, no. 3, pp. 283–319, Sept. 1970.

[5] R. B. CATTELL, "Parallel proportional profiles and other principles for determining the choice of factors by rotation", *Psychometrika*, vol. 9, pp. 267–283, 1944.

[6] P. COMON, "Blind channel identification and extraction of more sources than sensors", in *SPIE Conference*, San Diego, July 19-24 1998, pp. 2–13, republished in IEEE Trans. Sig. Proc., Jan. 2004.

[7] P. COMON, M. RAJIH, "Blind identification of under-determined mixtures based on the characteristic function", *ICASSP Conference*, Mar 18–23 2005.

[8] R. A. HARSHMAN, "Foundations of the Parafac procedure: Models and conditions for an explanatory multimodal factor analysis", *UCLA Working Papers in Phonetics*, vol. 16, pp. 1–84, 1970.

[9] R. A. HARSHMAN, "Determination and proof of minimum uniqueness conditions for parafac1", *UCLA Working Papers in Phonetics*, vol. 22, pp. 111–117, 1972.

[10] J. B. KRUSKAL, "Three-way arrays: Rank and uniqueness of trilinear decompositions", *Linear Algebra and Applications*, vol. 18, pp. 95–138, 1977.

[11] L. DE LATHAUWER, *Signal Processing Based on Multilinear Algebra*, PhD thesis, K. U. Leuven, E. E. Dept. -ESAT, Belgium, 1997.

[12] N. D. SIDIROPOULOS, R. BRO, "On the uniqueness of multilinear decomposition of n-way arrays", *Journal of Chemometrics*, vol. 14, pp. 229–239, 2000.

Une méthode de sous-espace pour l'extraction aveugle d'une source cyclostationnaire

Roger BOUSTANY, Jérôme ANTONI

Université de Technologie de Compiègne
Centre de Recherche de Royallieu, BP 20529, 60205 Compiègne Cedex, France
roger.boustany@utc.fr, jerome.antoni@utc.fr

Résumé – Le besoin de séparer des mélanges de signaux apparaît dans beaucoup d'applications du traitement du signal. La solution à ce problème réside dans les techniques émergentes de la séparation aveugle de sources (SAS) qui se fondent généralement sur la connaissance du nombre de sources statistiquement indépendantes dans le mélange. Cette communication traite le cas où le nombre de sources est inconnu et l'indépendance statistique n'est pas forcément satisfaite, mais où il n'y a qu'un seul signal d'intérêt (SI) cyclostationnaire à extraire. La méthode proposée exploite la cyclostationnarité de ce SI par une décomposition en sous-espaces des signaux observés via ses statistiques cycliques d'ordre 2. Cette approche est d'abord développée pour les mélanges instantanés et est ensuite étendue au cas convolutif dans le domaine fréquentiel où le problème des permutations ne se pose pas comme c'est le cas pour la SAS classique. Des expériences sur des signaux ECG et industriels sont finalement effectuées et illustrent la bonne performance de la méthode proposée.

Abstract – The need for blindly separating mixtures of source signals arises in many signal processing applications. The solution to this problem was found using emerging blind source separation (BSS) techniques which usually rely on the knowledge of the number of statistically independent sources present in the mixture. This paper deals with the case where the number of sources is unknown and the statistical independence may not apply, but where there is only one cyclostationary signal of interest (SOI) to be separated. A method for extracting this SOI by exploiting its cyclostationarity through a subspace decomposition of the observations is proposed. It is first developed for instantaneous mixtures and is then extended to the convolutive case in the frequency domain where it does not suffer from the permutation problem as does classical BSS. Experiments on ECG and industrial data are finally performed and illustrate the high performance of the proposed method.

1 Introduction

Dans de nombreux domaines du traitement du signal, l'expérimentateur doit faire face au problème de la séparation de sources présentes dans un mélange. Au cours de ces dernières années, plusieurs méthodes de séparation aveugle se basant sur l'indépendance statistique des sources ont été mises en oeuvre. Pour approcher l'indépendance statistique des sources recherchées, elles profitent habituellement d'hypothèses supplémentaires sur la nature des signaux, telles que leur non-gaussiannité [1], leur non-blancheur [2] ou leur non-stationnarité — et plus particulièrement leur cyclostationnarité [3], [4], [5].
La séparation aveugle de sources (SAS) consiste à retrouver m *sources* inconnues $\mathbf{s}(t) = (s_1(t), \ldots, s_m(t))^T$ à partir des mesures de n signaux de mélange ou *observations* $\mathbf{x}(t) = (x_1(t), \ldots, x_n(t))^T$, $t \in \mathbb{R}$. Suivant l'application, le modèle du mélange est soit instantané

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) + \mathbf{n}(t), \tag{1}$$

soit convolutif

$$\mathbf{x}(t) = \mathbf{H}(t) * \mathbf{s}(t) + \mathbf{n}(t), \tag{2}$$

où $*$ est le produit de convolution, $\mathbf{A}$ et $\mathbf{H}(t)$ représentent respectivement des matrices de rang plein de «scalaires» a_{ji} et de «filtres linéaires» $h_{ji}(t)$ modélisant le transfert de le source i au capteur j et où $\mathbf{n}(t)$ est un bruit additif décorrélé avec les sources. Cependant, dans de nombreuses situations, le nombre de sources est inconnu et supérieur au nombre de capteurs. Deux conditions nécessaires à la plupart des algorithmes classiques pour réaliser la séparation se trouvent donc mises à défaut. Par ailleurs, l'indépendance statistique mutuelle des sources est une condition restrictive qui n'est pas souvent respectée dans les applications réelles, telles que celles impliquant des signaux industriels. L'objectif de cette communication est de proposer une approche se substituant à la SAS classique dans le but d'extraire la contribution sur les capteurs[1] d'un signal d'intérêt (SI) cyclostationnaire — désigné par $s_1(t)$ — noyé dans un nombre inconnu de signaux non-désirés. Ceux-ci sont de deux types :

- les signaux non-désirés spatialement cohérents donnés par les contributions $\{\sum_{i=2}^{m} h_{ji}(t)s_i(t)\}_{j=1}^{n}$ et désignés par *interférences*.
- les signaux non-désirés spatialement non-cohérents donnés par les contributions $\{n_j(t)\}_{j=1}^{n}$ et simplement désignés par *bruits*.

L'approche proposée exploite la cyclostationnarité du SI par une décomposition en sous-espace des observations. Par opposition aux algorithmes SCORE [6], la méthode «SUBLEX» présentée dans le paragraphe 3 permet facilement l'utilisation de plusieurs fréquences cycliques pour améliorer l'estimation du SI. De plus, elle ne nécessite la détermination que de la moitié du nombre d'inconnus de SCORE. Une extension de la méthode aux mélanges convolutifs est également présentée pour couvrir une plus large étendue d'applications. Les expériences

[1] La contribution d'une source sur un capteur est le signal enregistré quand uniquement cette source est active

du paragraphe 4 sur des signaux ECG et sur des signaux industriels illustrent la bonne performance de SUBLEX. Le paragraphe 2 passe en revue brièvement les fondements de la cyclostationnarité à l'ordre 2.

2 Les signaux cyclostationnaires et leurs descripteurs d'ordre 2

Dans ce qui suit, tous les signaux sont supposés à puissance finie et stochastiques. Un signal (presque-)cyclostationnaire au second ordre $s(t)$ avec des fréquences cycliques $\alpha_i \in \mathcal{A}$ est tel que sa fonction d'auto-corrélation est (presque-)périodique et admet par conséquent un développement en série de Fourier

$$\begin{aligned} R_s(t,\tau) &= \mathbb{E}\{s(t)s^*(t-\tau)\} \\ &= \sum_{\alpha_i \in \mathcal{A}} R_s^{\alpha_i}(\tau) e^{j2\pi\alpha_i t}, \end{aligned} \quad (3)$$

où la fonction d'auto-corrélation cyclique $R_s^{\alpha_i}(\tau)$ est identiquement non-nulle sur l'ensemble $\mathcal{A}$.
La corrélation spectrale définie par la double transformée de Fourier de la fonction d'auto-corrélation par rapport à t et τ

$$S_s(\alpha, f) = \int\int R_s(t,\tau) e^{-j2\pi\alpha t} e^{-j2\pi f\tau} dt d\tau \quad (4)$$

devient en y insérant l'éq.(3)

$$S_s(\alpha, f) = \sum_{\alpha_i \in \mathcal{A}} S_s^{\alpha_i}(f)\delta(\alpha - \alpha_i), \quad (5)$$

ce qui veut dire que la «puissance» d'un signal cyclostationnaire est distribuée suivant des lignes spectrales parallèles à l'axe de f et positionnées aux fréquences cycliques $\alpha = \alpha_i$. La quantité spectrale $S_s^{\alpha_i}(f)$ est connue sous le nom de spectre de puissance cyclique et sera utilisée dans la méthode SUBLEX relative aux mélanges convolutifs. L'approche SUBLEX des mélanges instantanés décrite ci-après n'utilise pour l'instant que la fonction d'auto-corrélation cyclique.

3 La méthode SUBLEX

3.1 Principes

Le modèle du mélange bruité décrit par (1) est d'abord considéré. Comme indiqué auparavant, le SI résulte d'une seule source cyclostationnaire avec une fréquence cyclique α non partagée avec les interférences et le bruit — ces derniers étant supposés donc soit stationnaires soit cyclostationnaires avec des fréquences cycliques différentes. La matrice de corrélation cyclique des observations est :

$$\begin{aligned} \mathbf{R}_x^{\alpha}(\tau) &= \left\langle \mathbb{E}\{\mathbf{x}(t)\mathbf{x}^H(t-\tau)\} e^{-j2\pi\alpha t} \right\rangle \\ &= \mathbf{A}\mathbf{R}_s^{\alpha}(\tau)\mathbf{A}^H, \end{aligned} \quad (6)$$

avec $\langle(\ldots)\rangle = \lim_{T\to\infty} T^{-1} \int_{-T/2}^{T/2} (\ldots) dt$. C'est une matrice de rang 1 car $\mathbf{R}_s^{\alpha}(\tau)$ ne possède qu'un seul élément non-nul situé sur sa diagonale et ceci est vrai indépendamment du bruit $\mathbf{n}(t)$. Par conséquent,

$$\mathbf{R}_x^{\alpha}(\tau) = R_{s_1}^{\alpha}(\tau)\mathbf{a}_1\mathbf{a}_1^H, \quad (7)$$

où $R_{s_1}^{\alpha}(\tau)$ est la fonction d'auto-corrélation cyclique du SI $s_1(t)$ et $\mathbf{a}_1$ est la colonne correspondante de la matrice $\mathbf{A}$. Par suite, la décomposition en valeurs propres de la matrice $\mathbf{R}_x^{\alpha}(\tau)\mathbf{R}_x^{\alpha}(\tau)^H$ permet d'obtenir la colonne $\mathbf{a}_1$ de $\mathbf{A}$ à un coefficient près, mais sans que ceci soit suffisant pour séparer les sources. Pour extraire le SI, un sous-espace complémentaire $\mathcal{B}$ engendré par la base des vecteurs $\mathbf{B}$ orthogonale à $\mathbf{a}_1$ est construit tel que

$$\mathbf{r}(t) = \mathbf{B}^H\mathbf{x}(t) \quad (8)$$

est un vecteur de dimension $(n-1)$ dont les éléments orthogonaux entre eux et avec $s_1(t)$ sont des références sur les interférences. La matrice $\mathbf{B}$ de dimension $n \times (n-1)$ est par exemple déterminée à l'aide de la décomposition QR.
Les contributions des interférences sur les capteurs dénotées par $\mathbf{x}_{.s_1}(t)$ sont ensuite estimées par moindres carrés :

$$\widehat{\mathbf{x}}_{.s_1}(t) = \mathbf{C}\mathbf{r}(t), \quad (9)$$

où

$$\mathbf{C} = \mathbf{R}_{xr}^0(0)\mathbf{R}_r^0(0)^+, \quad (10)$$

avec

$$\begin{aligned} \mathbf{R}_{xr}^0(0) &= \left\langle \mathbb{E}\{\mathbf{x}(t)\mathbf{r}^H(t)\} \right\rangle, \\ \mathbf{R}_r^0(0) &= \left\langle \mathbb{E}\{\mathbf{r}(t)\mathbf{r}^H(t)\} \right\rangle, \end{aligned}$$

et $^+$ dénote la pseudo-inverse. Les contributions du SI $\widehat{\mathbf{x}}_{/s_1}(t)$ sont finalement extraites des signaux capteurs par :

$$\widehat{\mathbf{x}}_{/s_1}(t) = \mathbf{x}(t) - \widehat{\mathbf{x}}_{.s_1}(t).$$

3.2 Évaluation dans le cas non-bruité

SUBLEX est conçue pour remplacer la SAS dans le cas des mélanges bruités, quand il n'y a qu'un seul SI cyclostationnaire avec une fréquence cyclique α et un nombre inconnu de sources interférentes. Reste à démontrer que SUBLEX sépare le SI aussi bien que la SAS dans le cas non-bruité et en considérant alors — sans perte de généralité — que $\mathbb{E}\{\mathbf{s}(t)\mathbf{s}(t)^H\} = \mathbf{I}$. En effet, soit

$$\mathbf{A} = \mathbf{Q}\mathbf{\Omega} = \left(\frac{\mathbf{a}_1}{\|\mathbf{a}_1\|} \quad \mathbf{B} \right) \mathbf{\Omega} \quad (11)$$

la décomposition QR qui a permis d'avoir $\mathbf{B}$ de (8), Ω étant une matrice triangulaire supérieure de dimension $n \times m$. En insérant (11) dans (8), il résulte $\mathbf{r}(t) = \begin{pmatrix} \mathbf{0} & \mathbf{\Omega}_{22} \end{pmatrix} \mathbf{s}(t)$, avec $\mathbf{\Omega}_{22}$ la matric bloc de dimension $(n-1) \times (m-1)$ extraite de $\mathbf{\Omega}$, telle que $\mathbf{\Omega} = \begin{pmatrix} \Omega_{11} & \mathbf{\Omega}_{12} \\ \mathbf{0} & \mathbf{\Omega}_{22} \end{pmatrix}$. Ceci implique après quelques manipulations : $\mathbf{C} = \mathbf{A} \begin{pmatrix} \mathbf{0}^T \\ \mathbf{\Omega}_{22}^+, \end{pmatrix}$ où $\mathbf{\Omega}_{22}^+ = \mathbf{\Omega}_{22}^H(\mathbf{\Omega}_{22}\mathbf{\Omega}_{22}^H)^+$ est la pseudo-inverse de $\mathbf{\Omega}_{22}$. Finalement, remplaçant $\mathbf{C}$ dans (11) par son expression, il résulte

$$\begin{aligned} \widehat{\mathbf{x}}_{/s_1}(t) &= \mathbf{A}\mathbf{s}(t) - \mathbf{A} \begin{pmatrix} \mathbf{0}^T \\ \mathbf{\Omega}_{22}^+ \end{pmatrix} \mathbf{B}^H\mathbf{Q}\mathbf{\Omega}\mathbf{s}(t) \\ &= \mathbf{A}\mathbf{s}(t) - \mathbf{A} \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \mathbf{s}(t) \\ &= \mathbf{A} \begin{pmatrix} s_1(t) \\ \mathbf{0} \end{pmatrix}, \end{aligned} \quad (12)$$

à condition que $\mathbf{\Omega}_{22}^+\mathbf{\Omega}_{22} = \mathbf{I}$, c'est-à-dire que $\mathbf{\Omega}_{22}^+$ est l'inverse à gauche de $\mathbf{\Omega}_{22}$ ou bien $n \geq m$. Dans ce cas et en absence de

bruit, l'erreur sur $\widehat{\mathbf{x}}_{/s_1}(t)$ est nulle. Dans tous les autres cas — incluant les cas où les sources du mélange ne sont pas mutuellement indépendantes — la méthode fournit le meilleur résultat possible au sens des moindres carrés, au contraire de la SAS classique.

3.3 Considérations pratiques

En pratique, les signaux mesurés ont une taille finie et en conséquence les statistiques sont empiriques. La matrice empirique $\widehat{\mathbf{R}}_x^{\alpha}(\tau)$ peut être de rang supérieur à 1 ce qui implique une mauvaise estimation de $\mathbf{a}_1$. Pour remédier à ce problème, la méthode est (i) appliquée sur les observations blanchies $\widetilde{\mathbf{x}}$ et (ii) une diagonalisation conjointe d'un ensemble de matrices hermitiennes $\widehat{\mathbf{R}}_{\widetilde{x}}^{\alpha}(\tau)\widehat{\mathbf{R}}_{\widetilde{x}}^{\alpha}(\tau)^H$ indexées par différentes valeurs de α prises dans un ensemble fini $\mathcal{B} \subseteq \mathcal{A}$ et de τ, est effectuée pour extraire de manière plus robuste le vecteur $\mathbf{a}_1$.

3.4 Extension au cas des mélanges convolutifs

Le modèle convolutif décrit par (2) s'écrit dans le domaine fréquentiel, en termes d'incréments spectraux :

$$\mathbf{dX}(f) = \mathbf{H}(f)\mathbf{dS}(f) + \mathbf{dN}(f). \qquad (13)$$

La méthode SUBLEX est ensuite appliquée en utilisant la matrice spectrale cyclique $\mathbf{S}_x^{\alpha}(f)$ au lieu de la matrice de corrélation cyclique $\mathbf{R}_x^{\alpha}(\tau)$. En pratique, un pré-blanchiment des observations et une diagonalisation conjointe de $\widehat{\mathbf{S}}_{\widetilde{x}}^{\alpha}(f)\widehat{\mathbf{S}}_{\widetilde{x}}^{\alpha}(f)^H$ sur $\mathcal{B} \subseteq \mathcal{A}$ sont exécutés pour assurer une estimation plus robuste du vecteur $\mathbf{a}_1(f)$ de $\mathbf{H}(f)$. Il faut remarquer que contrairement à la SAS classique, cette méthode ne présente pas de problème de permutation puisqu'un seul SI est reconstruit.

4 Applications

La première application illustre l'utilisation de la version «instantanée» de SUBLEX pour l'extraction de l'électrocardiogramme d'un foetus — un signal contenant des informations importantes sur la santé du foetus — à partir des mesures prises sur l'abdomen de la mère. Les signaux sont constitués de 8 enregistrements de 2500 points échantillonnés à $f_s = 500$Hz [7]. Afin de déterminer la fréquence cyclique α du SI généré par le foetus, la corrélation spectrale décrite par (5) a été calculée sur l'un des signaux de mesures. La figure (1) montre une cyclostationnarité importante à la fréquence cyclique $\alpha = 8.8\times10^{-3}f_s$ et ses harmoniques. En examinant la corrélation spectrale d'un enregistrement provenant d'une électrode placée sur le thorax de la mère, la fréquence cyclique $\alpha = 8.8 \times 10^{-3}f_s$ n'existe pas, ce qui veut dire que cette fréquence cyclique peut être attribuée à l'activité électrique du foetus. La méthode SUBLEX a ensuite été appliquée sur 3 mesures abdominales en considérant $N_\alpha = 9$ fréquences cycliques pour l'estimation de $\mathbf{a}_1$. Les résultats de la figure (2) illustrent la bonne performance de la méthode.

La deuxième application concerne des mesures prises sur un système mécanique comprenant un étage d'engrenages avec un rapport de réduction de $32 : 49$, une vitesse d'entrée $f_r = 3$Hz

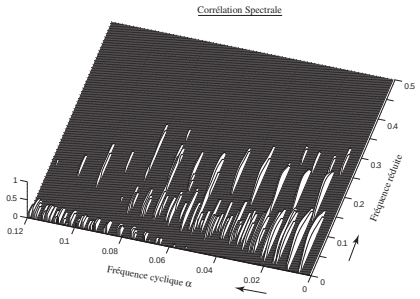


FIG. 1 – Corrélation spectrale d'un signal ECG (estimée avec une fenêtre demi-sinus de 32 points et 75% de recouvrement) montrant la fréquence cyclique foetale à $\alpha = 8.8 \times 10^{-3}f_s$. Le signal est pré-blanchi et uniquement les valeurs au-dessus d'un niveau de signification de 5% sont affichées.

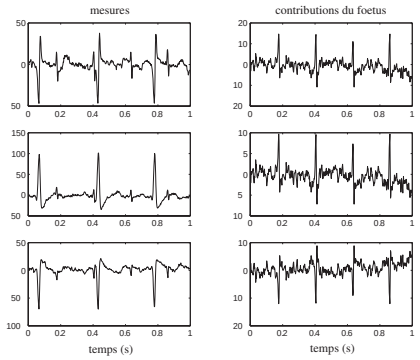


FIG. 2 – Extraction de l'électrocardiogramme foetal à partir de 3 enregistrements.

et un couple de 60N.m. L'axe moteur est supporté par un roulement à billes à deux rangées dont les caractéristiques sont les suivantes :

- Diamètre de la bague externe = 44.85mm
- Diamètre de la bague interne = 32.17mm
- Diamètre d'une bille B_d = 7.12mm
- Nombre de billes N_b = 12/rangée.

Ce roulement présente un défaut localisé sur la bague interne. L'objectif est d'extraire la contribution des vibrations générées par le passage des billes sur ce défaut [8]. La fréquence théorique de passage d'une bille sur un défaut de bague interne est donnée par la formule [9] :

$$BPFI = \frac{f_r N_b}{2}\left(1 + \frac{B_d}{P_d}\cos\theta\right) \approx 22.39Hz, \qquad (14)$$

où P_d et θ représentent respectivement le diamètre primitif et l'angle de contact du roulement. Les signaux étudiés ont une taille de 100000 points échantillonnés à $f_s = 16384$Hz. la corrélation spectrale de l'une des mesures (fig. 3) permet de déterminer de manière plus précise la fréquence cyclique du défaut. Une forte cyclostationnarité à $\alpha = 1.36 \times 10^{-3}f_s = 22.28$Hz et ses harmoniques confirme la valeur théorique de (14). L'extension de la méthode SUBLEX aux mélanges convolutifs a ensuite été appliquée avec $\alpha = 1.36\times10^{-3}f_s$ et $N_\alpha = 5$. La première colonne de la figure (4) présente les signaux de mesure, et la deuxième colonne les contributions du défaut extraites par SUBLEX. Un zoom sur quelques cycles présenté sur la figure (5) illustre la très bonne extraction des chocs dus au défaut en présence d'interférences comme celles dues à l'engrène-

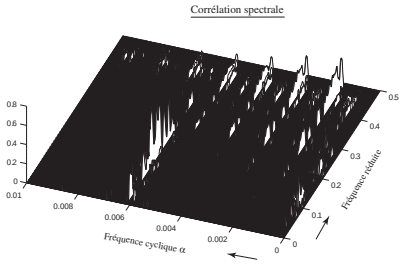


FIG. 3 – Corrélation spectrale d'un signal industriel (estimée avec une fenêtre demi-sinus de 256 points et 75% de recouvrement) montrant la fréquence cyclique d'un défaut de bague interne à $\alpha = 1.36 \times 10^{-3} f_s$. Le signal est pré-blanchi et uniquement les valeurs au-dessus d'un niveau de signification de 5% sont affichées.

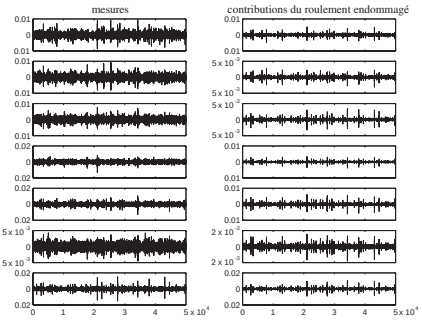


FIG. 4 – Extraction de la signature du défaut de roulement.

ment des pignons, et dans un bruit de fond considérable. Les spectres d'amplitude des enveloppes des signaux de mélange avant et après SUBLEX sont représentés à la figure (6). Avant SUBLEX, ils révèlent la dominance de la fréquence d'engrènement des pignons f_g et de ses harmoniques correspondants au fonctionnement normal du système, lesquels fortement masquent la signature du défaut de roulement. Après SUBLEX, cette fréquence f_g et ses harmoniques sont éliminés du signal extrait dont l'enveloppe ne contient plus que la fréquence *BPFI* et ses harmoniques modulés par la fréquence de rotation f_r comme prévu par la théorie pour un défaut de bague interne [9].

5 Conclusion

Une méthode SUBLEX de sous-espace a été présentée pour l'extraction d'une source cyclostationnaire mélangées avec un nombre *inconnu* de sources interférantes — une situation que la SAS classique aborde avec difficulté. L'approche proposée réalise une *parfaite* séparation du SI en absence de bruit si le nombre de capteurs est supérieur ou égal au nombre de sources. Dans les autres cas, elle garantit un débruitage significatif. Une extension de la méthode au cas des mélanges convolutifs a été également présentée dans le but de couvrir une plus large étendue d'applications. La version «instantanée» de SUBLEX a ensuite été illustrée sur des signaux ECG pour en extraire l'électrocardiogramme foetal, et sa version fréquentielle sur des signaux industriels pour en extraire la signature d'un défaut de roulement noyé dans un mélange convolutif complexe.

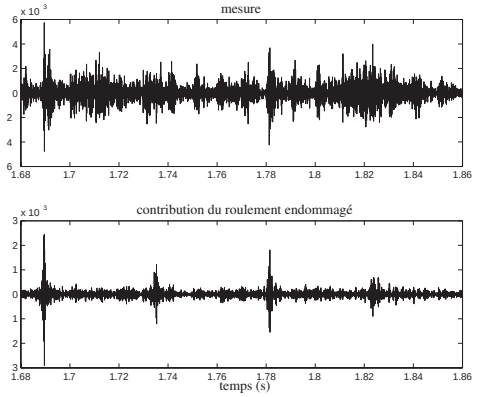


FIG. 5 – Zoom sur quelques cycles révélant la présence de chocs caractérisant un défaut localisé de roulement à billes.

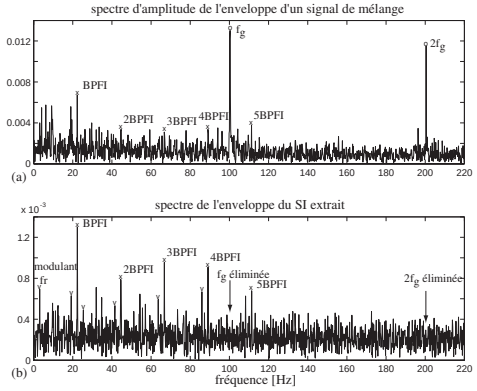


FIG. 6 – Spectres d'amplitude des enveloppes (a) d'un signal de mesure et (b) de la contribution du roulement endommagé extraite après application de SUBLEX.

Références

[1] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non gaussian signals," in *IEE Proceedings-F*, vol. 140, pp. 362–370, Dec. 1993.

[2] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso and E.Moulines, "A blind source separation technique using second order statistics," *IEEE Trans. on Signal Processing*, vol. 45, no. 2, pp. 434–444, Feb. 1997.

[3] K. Abed-Meraim, Y. Xiang and Y. Hua, "Blind source separation using second-order cyclostationary statistics," in *IEEE Trans. on Signal Processing*, vol. 49, no. 4, pp. 694–701, Apr. 2001.

[4] A. Ferréol and P. Chevalier, "On the behavior of current second and higher order blind source separation methods for cyclostationary sources," in *IEEE Trans. on Signal Processing*, vol. 48, no. 6, pp. 1712–1725, Jun. 2000.

[5] J. Antoni, F. Guillet, M. ElBadaoui and F. Bonnardot, "Blind separation of convolved cyclostationary processes," in *Signal Processing*, 85, pp.51–66, 2005.

[6] B.G. Agee, S.V. Schell and W.A. Gardner, "Spectral self-coherence restoral : A new approach to blind adaptive signal extraction using antenna arrays," in *Proceedings of the IEEE*, vol. 78, no. 4, pp. 753–767, Apr. 1990.

[7] J.-F. Cardoso, "Multidimensional independent component analysis," in *Proc. ICASSP'98*, Seattle, USA, May. 12-15. 1998, pp. 1941–1944.

[8] R. B. Randall, J. Antoni, S. Chobsaard, "The Relationship between Spectral Correlation and Envelope Analysis for Cyclostationary Machine Signals. Application to Ball Bearing Diagnostics," in *Mechanical Systems and System Processing*, vol. 15, no. 5, pp. 945–962, 2001.

[9] S. Braun, "Mechanical signature analysis : theory and applications", Academic Press, 1986.

Déconvolution aveugle dans le domaine fréquentiel fondée sur le taux d'information mutuelle

Anthony Larue, Jérôme Mars, Christian Jutten

Laboratoire des Images et des Signaux,
961 rue de la Houille Blanche 38402 Saint Matin d'Hères BP 46, France
larue@lis.inpg.fr,mars@lis.inpg.fr,jutten@lis.inpg.fr

Résumé – Nous introduisons le taux d'information mutuelle comme mesure de blancheur regroupant l'ensemble des statistiques d'ordre supérieur. Ensuite, nous l'utilisons pour construire un algorithme de déconvolution aveugle fréquentiel (DAF) d'un système SISO incorporant aussi une régularisation limitant l'amplification du bruit. Enfin, le traitement de données réelles d'explosions sous-marines et de sismologie est proposé.

Abstract – We introduce the mutual information rate as a whiteness measure using all higher order statistics. Then, we purpose a new frequency blind deconvolution algorithm of a SISO system including a regularization to avoid the noise amplification. Finally, we process real data of underwater explosions recordings and seismovolcanic signals.

1 Introduction

Le problème de déconvolution aveugle se rencontre dans de nombreux domaines d'applications (télécommunications, contrôle non-destructif, détection de choc, géophysique...). L'observation résulte de la convolution entre un signal d'entrée $r(t)$ et un filtre direct w plus un bruit additif $b(t)$ soit: $d(t) = (w \star r)(t) + b(t) = \sum_{i=-\infty}^{+\infty} w(i)r(t-i) + b(t)$. $\star$ est le produit de convolution. Dans le contexte de la déconvolution aveugle, seule l'observation $d(t)$ est connue, alors que le filtre direct et le signal d'entrée sont inconnus. Mais sous l'hypothèse d'un processus d'entrée iid et non Gaussien, le signal d'entrée est unique à un rapport d'amplitude et un retard près. Le modèle de convolution et l'algorithme de déconvolution sont résumés par la figure 1.

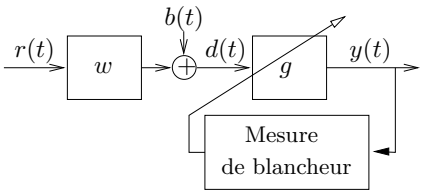


Figure 1: Modèle de convolution et déconvolution

Généralement, les algorithmes de déconvolution sont réalisés en ajustant un filtre inverse g suivant un critère de blancheur portant sur le signal de sortie $y(t)$. Pour mesurer le caractère iid de la sortie, on peut utiliser les statistiques d'ordre 2 (densité spectrale et fonction de corrélation). Ces méthodes (égalisation spectrale, Yule-Walker, Durbin) sont simples mais ne permettent pas l'estimation de la phase du filtre. Soit, on fixe la phase du filtre par des informations *a priori*, soit on utilise les statistiques d'ordre supérieur (à deux). Classiquement, on préfère les statistiques d'ordre 4 (tricorrelation[1, 2], kurtosis [3] et trispectre) à celles d'ordre 3 (bicorrélation, bispectre) car ces dernières sont nulles pour des distributions symétriques. Les performances de ces méthodes sont très influencées par la qualité des estimateurs des statistiques d'ordre supérieur (SOS). De plus, la sélection arbitraire des SOS n'assure pas le choix optimal des statistiques utilisées pour les signaux traités. Ainsi, nous proposons de mesurer la blancheur par une mesure plus générale : le taux d'information mutuelle (TIM) de la sortie.

2 Algorithme de déconvolution

L'information mutuelle d'un vecteur aléatoire $\mathbf{z} = (z_1, \ldots, z_n)$ de dimension n est définie par:

$$I(\mathbf{z}) = \sum_{i=1}^{n} H(z_i) - H(z_1, z_2, \ldots, z_n) \qquad (1)$$

où $H(z_i)$ est l'entropie de Shannon marginale de z_i, i.e. $H(z_i) = -\int_{\mathbb{R}} p_{z_i}(u) \log p_{z_i}(u) du$ et $H(z_1, \ldots, z_n)$ l'entropie de Shannon conjointe $H(\mathbf{z}) = -\int_{\mathbb{R}^n} p_{\mathbf{z}}(\mathbf{u}) \log p_{\mathbf{z}}(\mathbf{u}) d\mathbf{u}$. L'information mutuelle $I(\mathbf{z})$ est un critère d'indépendance pour la séparation de sources car elle est toujours positive et $I(\mathbf{z})$ s'annule si et seulement si les composantes de $\mathbf{z}$ sont mutuellement indépendantes. En déconvolution aveugle, on doit mesurer l'indépendance entre les échantillons d'un processus, soit une séquence de longueur infinie. Ainsi, on définit le taux d'entropie [4] d'un processus aléatoire $Z = \{Z_t\}$ par :

$$\mathcal{H}(Z) = \lim_{T \to \infty} \frac{1}{T} H(Z_1, \ldots, Z_T) \qquad (2)$$

Cette limite existe pour un processus stationnaire [4]. Puis, le taux d'information mutuelle (TIM) du processus aléatoire Z est défini par:

$$I(Z) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^{t=T} H(Z_t) - \mathcal{H}(Z) \qquad (3)$$

où $H(Z_t)$ représente l'entropie marginale du processus Z_t et $\mathcal{H}(Z)$ le taux d'entropie de Z défini en (2). Pour

plus de clarté, Z représente le processus contenant tous les échantillons, et Z_t est le processus avec un seul échantillon $z(t)$. Le taux d'information mutuelle $I(Z)$ est toujours positif et s'annule si et seulement si Z est un processus iid [4]. Ainsi, le TIM est une mesure de blancheur acceptable pour construire des algorithmes de déconvolution aveugle (Cf figure 1). Sous hypothèses de stationnarité, l'entropie marginale du processus Z_t est indépendante du temps. Tous les termes de la somme de (3) sont égaux, donc nous avons:

$$\lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{t=T} H(Z_t) = H(Z_\tau) \quad \forall\ \tau \tag{4}$$

Par conséquent, en ajoutant l'hypothèse de stationnarité, le taux d'information mutuelle (3) se récrit simplement:

$$I(Z) = H(Z_\tau) - \mathcal{H}(Z) \tag{5}$$

Dans (5), τ est arbitrairement choisi. Mais, il n'y a pas de raison de privilégier une valeur de τ. De plus, plusieurs réalisations du processus Z_τ sont nécessaires pour estimer convenablement l'entropie marginale $H(Z_\tau)$. Ainsi, en pratique, pour estimer l'entropie de Z_τ, nous utilisons tous les échantillons $z(1),\ldots,z(T)$ comme T réalisations du processus Z_τ. Dans la suite, nous considérerons comme équivalent le processus Z_t et une réalisation de ce dernier $z(t)$. Dans le contexte de la déconvolution aveugle, on peut obtenir un critère plus simple [5, 6], en remarquant que le taux d'entropie de la sortie $y(t) = (g\star d)(t)$ s'exprime de la façon suivante:

$$\mathcal{H}(Y) = \mathcal{H}(D) + \frac{1}{2\pi}\int_0^{2\pi} \log\left|\sum_{t=-\infty}^{+\infty} g(t)\exp\left(-jt\theta\right)\right| d\theta \tag{6}$$

Avec (6), le TIM de la sortie Y est:

$$I(Y) = H(y(\tau)) - \mathcal{H}(D) - \frac{1}{2\pi}\int_0^{2\pi} \log\left|\sum_{t=-\infty}^{+\infty} g(t)\exp\left(-jt\theta\right)\right| d\theta \tag{7}$$

Puis, comme le taux d'entropie des données $\mathcal{H}(D)$ est indépendant du filtre inverse g, il est équivalent de mesurer la blancheur de la sortie avec (7) ou $\tilde{I}(Y)$:

$$\tilde{I}(Y) = H(y(\tau)) - \frac{1}{2\pi}\int_0^{2\pi} \log\left|\sum_{t=-\infty}^{+\infty} g(t)\exp\left(-jt\theta\right)\right| d\theta \tag{8}$$

qui est minimum quand le processus Y est iid. Ce critère utilise toutes les statistiques d'ordre supérieur par l'intermédiaire de l'entropie. Avec cette mesure de blancheur, on peut envisager différents algorithmes de déconvolution. Ces algorithmes sont equivalent à des méthodes de maximum de vraisemblance, dans lesquelles on remplace la distribution du signal de réflectivité supposée connue par le maximum de vraisemblance par la distribution de la sortie estimée à chaque itération. Taleb *et al.* [5] propose dans le cas d'un système de Wiener, de minimiser (8) par rapport à la réponse impulsionnelle du filtre $g(t)$. En prenant, un nombre fini de coefficients pour $g(t)$, la méthode est adaptée à l'utilisation d'un modèle MA pour le filtre inverse donc un filtre direct w autorégressif. Cet algorithme (MAMV) n'est pas très robuste au bruit. En présence de bruit additif Gaussien, des expériences montrent que ses performances sont identiques à la méthode d'ordre 2 de Yule-Walker. **Nous proposons de construire un critère de déconvolution dans le domaine fréquentiel à l'aide de la mesure de blancheur (8) égale à une constante près au TIM.** Nous travaillerons en fréquence, car l'amplification du bruit (supposé large bande) sur la sortie est contrôlée simplement en limitant les grandes valeurs de $G(f)$. Ensuite, on ajoute une contrainte de continuité sur la réponse en fréquence du filtre inverse. Ainsi, le critère de déconvolution en fonction de la réponse en fréquence $\mathbf{G} = [G(0),\ldots,G(T-1)]$ est:

$$\begin{aligned} J(\mathbf{G}) &= H(y(\tau)) - \frac{1}{T}\sum_{f=0}^{T-1}\log|G(f)| + \\ &\quad \lambda_1\sum_{f=0}^{T-1}|G(f)-G(f+1)|^2 + \lambda_2\sum_{f=0}^{T-1}|G(f)|^p \end{aligned} \tag{9}$$

Les deux premiers termes sont l'estimation de la mesure de blancheur (8). Le terme pondéré par l'hyperparamètre λ_1 est la contrainte de continuité sur le spectre du filtre inverse. Le dernier terme est l'équivalent du facteur de bruit pour le filtrage de Wiener, pénalisant avec la norme $\mathbb{L}^p$ les grandes valeurs de $G(f)$ qui amplifient des bandes de fréquences représentant uniquement du bruit. Ces deux termes de régularisation peuvent être analysés comme deux *a priori* sur la distribution de la réponse en fréquence du filtre (i) une loi conditionnelle gaussienne entre $G(f)$ et $G(f+1)$, (ii) une loi Gaussienne généralisée d'ordre p pour la loi marginale de $G(f)$. Le minimum de (9) est obtenu par une descente de gradient [7] par rapport aux coefficients complexes $G(f),\ f = 0,\ldots,T-1$. Le gradient de (9) par rapport à $G(f)$ est :

$$\begin{aligned} \widehat{\nabla J(\mathbf{G})} &= \frac{1}{2T^2}\Psi_Y(f)D^*(f) - \frac{1}{2T}\frac{1}{G^*(f)} + \\ &+ \lambda_1(2G(f) - G(f+1) - G(f-1)) \\ &+ \lambda_2\frac{p}{2}\frac{|G(f)|^p}{G^*(f)} \end{aligned} \tag{10}$$

où $\Psi_Y(f)$ est la transformée de la fonction score définie par $\psi_Y(u) = -\frac{d}{du}\log p_Y(u)$. L'information des SOS est contenue dans cette fonction. **L'algorithme de déconvolution aveugle en fréquence (DAF)** est le suivant:

1. *initialisation du filtre inverse $G(f)$ et de $y(t)$*
2. *estimation de la fonction score ψ_Y;*
3. *calcul du gradient (10);*
4. *mise à jour de $G(f) \leftarrow G(f) - \mu\widehat{\nabla J(\mathbf{G})}$;*
5. *calcul de la sortie $y(t)$;*
6. *étape de normalisation.*

μ est le pas du gradient. Nous itérons entre les étapes 2 et 6 jusqu'à la convergence de l'algorithme. L'étape de normalisation permet de lever l'indétermination d'amplitude sur la sortie. La fonction score est estimée par une méthode à noyaux développée par Pham [8] dans le cadre de la séparation de sources. Cet estimateur a de bonnes performances quelque soit la distribution du signal car il est basé sur une méthode très générale à noyaux. L'utilisation d'un noyau à support borné (splines cubiques) donne un coût de calcul assez faible.

3 Exemples

Le but des premiers tests est de montrer la capacité de notre algorithme à estimer des filtres quelque soit leurs phases. L'expérience des figures 2 et 3 utilise comme signal d'entrée (en (a)) un signal de réflectivité (applications : contrôle non-destructif, géophysique) modélisé par un signal Bernoulli-Gaussien avec 50 réflecteurs pour 400 échantillons. Les observations en (b) sont le résultat du filtrage de ce signal d'entrée par un filtre ARMA(10,10) à phase minimale (respectivement nulle) pour la figure 2 (resp. 3). La qualité de la déconvolution peut être analysée dans le domaine temporel en comparant le signal déconvolué en (c) au signal d'entrée en (a), et dans le domaine fréquentiel, on compare en (d) (resp. (e)) le module (resp. la phase) du filtre inverse estimé (en trait continu) à la valeur théorique (trait pointillé).

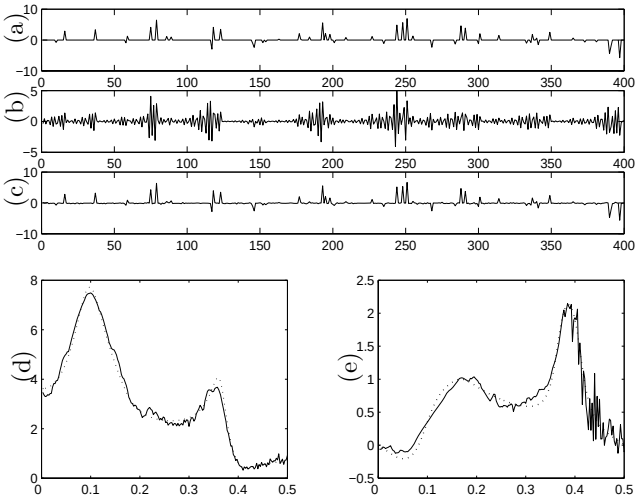


Figure 2: (a) Signal Bernoulli-Gaussien d'entrée. (b) Observation après filtre ARMA(10,10) à phase minimale. (c) Signal déconvolué par algorithme DAF. (d) (resp. (e)) Module (resp. phase) du filtre inverse estimé (trait continu) comparé avec la valeur théorique (trait pointillé)

Cette expérience avec des signaux simulés montre que l'estimation de filtre à phase non-minimale est réalisable avec l'algorithme DAF grâce à l'utilisation des SOS par le biais du TIM. L'erreur d'estimation sur la phase est due aux grands nombres de paramètres à optimiser. De plus, l'erreur est la plus grande dans la plage [0.4 0.5] qui est la partie la moins énergétique du spectre des données.

La figure 4 représente en (a) un enregistrement réel en piscine de la pression lors d'une explosion sous-marine. L'observation est constituée de trois ondes : (i) l'onde directe, (ii) une réflexion sur le fond, (iii) une réflexion à la surface. En appliquant le modèle de convolution à ces données, le signal d'entrée contient les informations relatives entre l'arrivée directe de l'onde et les réflexions. Le filtre direct représente l'onde émise. Cette onde est caractéristique de l'effet bulle avec une croissance exponentielle de la pression suivie d'une décroissance exponentielle. Pour les méthodes paramétriques, on retient un

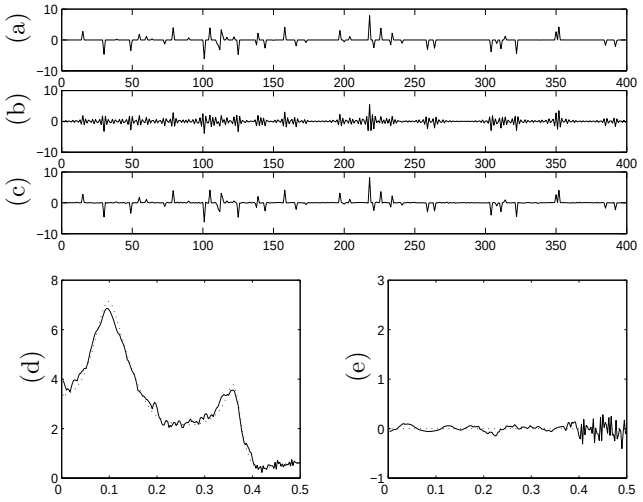


Figure 3: (a) Signal Bernoulli-Gaussien d'entrée. (b) Observation après filtre ARMA(10,10) à phase nulle. (c) Signal déconvolué par algorithme DAF. (d) (resp. (e)) Module (resp. phase) du filtre inverse estimé (trait continu) comparé avec la valeur théorique (trait pointillé)

modèle MA non causal pour cette onde. Dans la figure 4, nous comparons trois algorithmes: (b) l'égalisation spectrale, (c) la méthode de Durbin, (d) l'algorithme DAF.

L'algorithme DAF donne le meilleur compromis entre la qualité de la déconvolution (pics étroits, sans pics secondaires) et le niveau de bruit (100 premiers échantillons).

L'algorithme a une approche très générale (hypothèse de blancheur, non paramétrique), une autre application est la seismologie [9]. Le signal enregistré figure 5(a) par un géophone à la surface du volcan Purace (Colombie) présente une très forte résonance. Dans le modèle de la figure 1, le signal d'entrée est l'excitation due à une explosion à l'intérieur du volcan. La cheminée du volcan joue le rôle de filtre résonant. Le but de la déconvolution est de séparer les effets résonants (caractéristiques de la géométrie du volcan) et l'excitation liée à l'activité du volcan. L'excitation liée à une explosion sera supposée blanche. La figure 5 compare quatre algorithmes. Nous comparons la sortie estimée (colonne de gauche) correspondant au signal d'excitation et la réponse en fréquence du filtre direct estimée (colonne de droite) pour les fréquences réduites comprises entre 0 et 0.15 pour quatre algorithmes : (b) et (c) égalisation spectrale (ordre 2, fréquentiel); (d) et (e) Yule-Walker (ordre 2, temporel); (f) et(g) MAMV [5] (TIM, temporel); (h) and (i) DAF (TIM, fréquentiel). L'algorithme MAMV minimise le TIM par rapport au coefficient de la réponse impulsionnelle du filtre inverse. Ainsi, nous avons les quatre combinaisons possible de méthodes avec le choix de deux domaines de travail et l'utilisation uniquement des statistiques d'ordre 2 ou de l'ensemble des statistiques d'ordre supérieur par le biais du TIM.

Pour juger la qualité de la déconvolution, sur les allures temporelles, on doit obtenir une excitation très localisée temporellement, avec aucun effet de résonance. Le com-

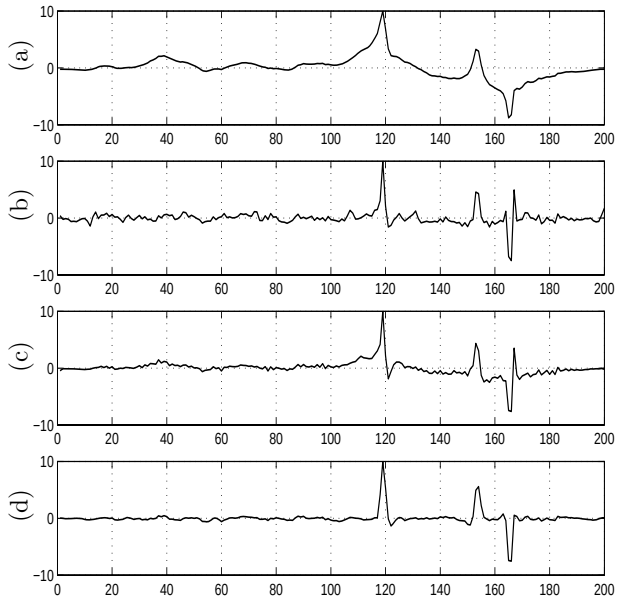


Figure 4: (a) Observation de l'explosion sous-marine (b) Egalisation spectrale, (c) Méthode de Durbin, (d) Algorithme DAF avec $\mu = 0.01$, $\lambda_1 = 0.1$, $\lambda_2 = 0.1$ et $p = 3$

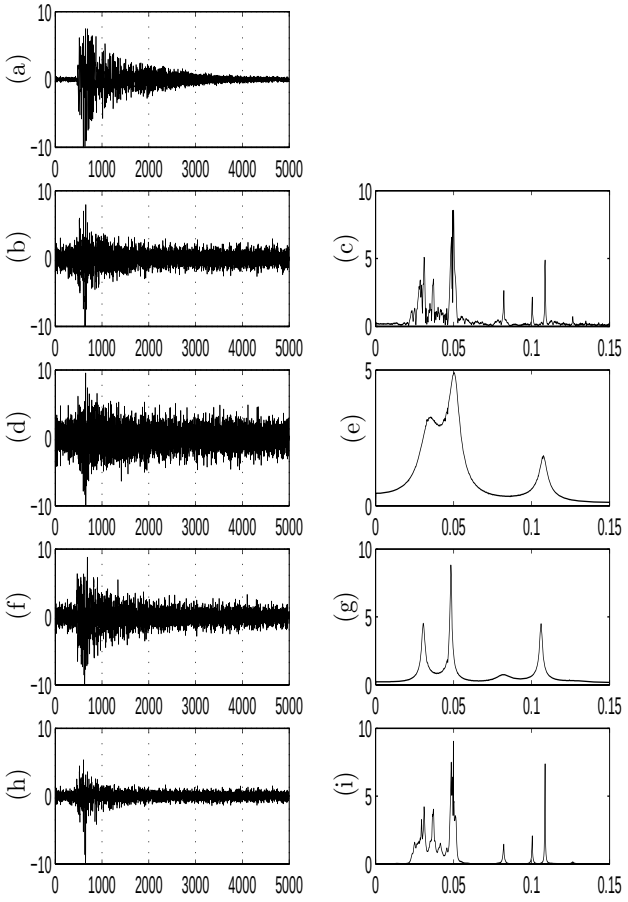


Figure 5: (a) Observations seismovolcaniques, sortie et filtre direct estimés :(b)(c) égalisation spectrale; (d) et (e) Yule-Walker; (f) et (g) MAMV; (h) et (i) DAF

promis entre la qualité de la déconvolution et l'amplification du bruit est tout simplement liée au niveau de bruit sur la sortie. Le spectre du filtre direct estimé permet de donner des informations sur les fréquences de résonance présentent dans le signal. Ces fréquences caractérisent de la géométrie du volcan. La méthode doit réaliser un compromis entre extraire le maximum de fréquences de résonance tout en ayant un spectre assez lisse. En suivant tous ces critères, on observe que l'algorithme DAF donne les meilleurs résultats en terme de localisation de l'excitation, de niveau de bruit et de nombre de fréquences de résonance extraites.

4 Conclusion

Nous avons proposé un nouvel algorithme de déconvolution dans le domaine fréquentiel. Il est particulièrement intéressant pour des applications réelles, grâce à la régularisation vis à vis du bruit. Il s'adapte aux données en sélectionnant les meilleures statistiques d'ordre supérieur pour le traitement par l'intermédiaire de l'estimation de la fonction score de la sortie.

References

[1] P. Comon, "MA identification using fourth order cumulants," *Signal Processing*, vol. 26, 1992.

[2] J. A. Cadzow, "Blind deconvolution via cumulant extrema," *IEEE Signal Processing Magazine*, pp. 24–42, May 1996.

[3] M. Boumahdi, "Blind identification using the kurtosis with applications to field data," *Signal Processing*, vol. 48, no. 3, pp. 205–216, 1996.

[4] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, 1991.

[5] A. Taleb, J. Solé i Casals, and C. Jutten, "Quasi-nonparametric blind inversion of Wiener systems," *IEEE Trans. on Signal Processing*, vol. 49, no. 5, pp. 917–924, May 2001.

[6] D.-T. Pham, "Contrast functions for blind separation and deconvolution sources," in *Proc. of the Int. Conf. ICA'01*, San-Diego, December 2001, pp. 37–42.

[7] A. Van Den Bos, "Complex gradient operator and hessian," *IEE Proceedings on Vision, Image and Signal Processing*, vol. 141, pp. 380–382, 1994.

[8] D.-T. Pham, "Fast algorithm for mutual information based independent component analysis," *IEEE Trans. on Signal Processing*, vol. 52, no. 10, pp. 2690–2700, October 2004.

[9] P. Lesage, F. Glangeaud, and J. Mars, "Applications of autoregressive models and time-frequency analysis to the study of a volcanic tremor and long-period events," *Journal of Volcanology and Geothermal Research*, vol. 114, no. 3, pp. 391–417, 2002.

Définition d'une force d'attraction aux LPE pour la segmentation par contours actifs

Annabelle GOUZE, Cédric DE ROOVER, Benoît MACQ

Laboratoire Télécommunications, Université catholique de Louvain
Place du Levant 2, 1348 Louvain la Neuve, Belgique
gouze@tele.ucl.ac.be, deroover@tele.ucl.ac.be, macq@tele.ucl.ac.be

Résumé – Dans cet article, nous traitons du problème de la segmentation vidéo. Nous proposons une nouvelle méthode qui combine les contours actifs et une pré-segmentation par les LPE (lignes de partage des eaux). Les méthodes LPE traitent efficacement les images couleurs par une sursegmentation en régions de couleurs homogènes. D'autre part, les contours actifs permettent d'obtenir des résultats réguliers. L'algorithme de segmentation proposé se compose de deux étapes. Dans un premier temps, les contours actifs évoluent sous un critère permettant de détecter un object en mouvement dans une séquence vidéo à caméra fixe. Une incertitude sur le mouvement peut induire une imprécision sur la position du contour final ainsi que des artéfacts. Aussi, nous affinons les résultats en introduisant une force d'attraction aux LPE. La méthode proposée améliore la segmentation en augmentant la précision et la régularité des résultats retournés par les contours actifs basés mouvement et les LPE.

Abstract – In this paper we address the problem of video segmentation. We propose a new method which combines an active contour and a watershed presegmentation. On the one hand, watershed methods are very efficient to provide an oversegmentation of homogeneous color regions. On the other hand, active contours methods are efficient to obtain a smooth segmentation. The proposed segmentation algorithm is performed in two steps. We first apply an active contour method to segment moving objects in video for static camera. The incertitude on the motion induces artifacts on the resulting contours. Then, the propagation is driven by the distance to the watershed boundaries to refine the results. The proposed method provides finer and smoother results than moving object segmentation by active contours and than watersheds.

1 Introduction

La segmentation d'images et de séquences vidéo vise à partitionner les données en différentes zones d'intérêt afin de permettre l'interprétation et l'analyse d'image. La segmentation peut servir à délimiter de manière précise les objets de l'arrière-plan pour être utilisés par des applications telle que la post-production. Cette dernière nécessite une segmentation fine afin de délimiter avec précision le plan objet de l'arrière-plan et ne requière pas une évolution temps-réel. Plusieurs techniques ont été développées aux cours de ces dernières décennies, notamment les méthodes par contours actifs et par LPE (lignes de partages des eaux). La segmentation par contours actifs permet de délimiter les contours d'un ou plusieurs objets sémantiques. Les contours actifs se révèlent efficace pour le suivi d'objets en mouvement, cependant la méthode souffre d'un manque de précision sur la position du contour dû à un trop faible mouvement et à la difficulté d'estimer l'arrière plan. D'autre part, les méthodes de segmentation LPE retournent une image sursegmentée. Un algorithme de fusion peut être employé en vue de restreindre le nombre de régions. La segmentation obtenue souffre, néanmoins, d'un manque de régularité. De récents travaux ont combiné les snakes et les LPE. Nguyen [11] et Park [13] ont décrit comment représenter une segmentation LPE comme un problème de minimisation d'énergie. Nguyen opère une segmentation LPE en plusieurs régions et régularise le résultat en introduisant une équation d'énergie basée sur les distances topologiques au centre de gravité des régions.
L'article aborde le problème autrement et décrit comment exploiter de manière efficace la pré-segmentation LPE dans un algorithme de contours actifs. En effet, nous présentons un nouveau critère d'énergie basé sur la distance aux lignes LPE. La dérivation de cette énergie détermine une force d'attraction vers les bords des lignes LPE les plus proches et permet d'obtenir une courbe régulière. Notre algorithme se décompose en deux étapes. Dans un premier temps, nous traquons les objets en mouvement avec un simple critère mouvement basé région. Et dans un second temps, nous appliquons la nouvelle force d'attraction aux LPE afin d'affiner les résultats de la segmentation. Cette évolution séquentielle permet de conduire la courbe sans être ralenti ou stoppé par les lignes LPE indésirables.
Cet article introduit en section 2 les méthodes de segmentation par LPE et par contours actifs. La section 3 décrit la méthode originale combinant les segmentations par contours actifs et par LPE. Enfin, la section 4 montre des résultats très concluants.

2 Introduction à la segmentation

2.1 LPE : lignes de partage des eaux

Toute image en niveaux de gris, telle qu'une image de gradient couleur, peut être considérée comme une surface topographique, contenant des monts, des plateaux et des vallées. La transformation morphologique par LPE a pour but de diviser cette surface topographique en différents bassins séparés par des lignes de partage des eaux [10]. Ces LPE partitionnent l'image en différentes régions homogènes. Le désavantage majeur des algorithmes LPE est leur haute sensibilité au bruit, il en résulte une sur-segmentation due à un grand nombre de

LPE. Cette sur-segmentation peut être résolue soit en fusionnant les régions, soit en limitant directement le nombre de régions créées par l'utilisation de marqueurs. Afin de faciliter le processus de fusion, Salembier a réorganisé les régions sur-segmentées dans un arbre de partition binaire [15]. Malgré le processus de fusion, des points faibles subsistent. Les contours ne sont pas réguliers et le suivi temporel des régions demeure non trivial.

2.2 Contours actifs

La segmentation par contours actifs consiste à déformer un contour initial et à le faire évoluer vers les bords d'un ou plusieurs objets à segmenter. L'évolution est conduite par une force (une équation aux dérivées partielles) émanant de la minimisation d'une fonctionnelle d'énergie [18, 2, 8, 12]. L'EDP fait évoluer le contour vers un minimum (local) d'énergie assimilé aux bords de l'objet. A l'origine, les snakes [9] ou les contours actifs géodésiques [1] sont calculés grâce à la minimisation d'une intégrale dont les caractéristiques dépendaient des contours. Par la suite, de nouvelles fonctionnelles d'énergie combinant des intégrales de contours et des intégrales de régions sont apparues. Elles furent introduites par [3] et [14], et développés par plusieurs auteurs [18, 2, 8, 12, 17].

3 Contours actifs contraints par les LPE

Dans cette section, nous proposons un critère basé sur la segmentation LPE, ensuite nous introduisons le schéma de segmentation par contour actif. L'idée est de segmenter séquentiellement les objets en mouvement en deux étapes complémentaires. La première étape permet de cibler les objets en mouvement et la seconde consiste à affiner le premier résultat de segmentation en attirant les contours vers les bord des régions de couleurs homogènes délimitées par les LPE. Cela suppose que la détection des objets retourne un contour proche des LPE correspondant à la meilleure solution.

3.1 Définition de la force d'attraction aux LPE

Considérons une carte de sur-segmentation définie par les LPE, notre but est d'attirer le contour actif vers la ligne LPE la plus proche. Pour ce faire, nous introduisons un critère qui contraint la distance du contour aux bords des régions LPE. Le critère, J_W, est exprimé comme suit :

$$J_W(\tau) = \int_{\Gamma(\tau)} d^2(\Gamma(\tau), W) ds \tag{1}$$

où W est un ensemble de courbes correspondantes aux LPE, d est une distance géométrique des points d'évolution à W. d peut être calculé par les algorithmes de carte de distance [4]. La définition de d pour $X \in \Gamma(\tau)$ est :

$$d(X, W) = min_{Y \in W} |X - Y| = |X - Y_{min}| . \tag{2}$$

Dans [6], Gastaud a aussi défini une énergie basée sur la distance signée à un simple contour de référence :

$$J_C(\Gamma) = \int_{\Gamma} \varphi(u(\Gamma, \Gamma^{ref})) ds \tag{3}$$

où φ est une fonction dérivable de la distance géométrique u :

$$u(X, \Gamma^{ref}) = \begin{cases} -min_{Y \in \Gamma^{ref}} |X - Y| & \text{si } X \text{ est à extérieur de } \Gamma^{ref} \\ min_{Y \in \Gamma^{ref}} |X - Y| & \text{si } X \text{ est à l'intérieur de } \Gamma^{ref} \end{cases} \tag{4}$$

X est un point de Γ et Y les coordonnées d'un point du contour de référence Γ^{ref}. En utilisant la méthode de dérivation de Gastaud qui suppose pour dériver l'intégrale que $\Gamma(\tau)$ est paramétrée par p dans $[0, 1]$, et ainsi que u et d sont des fonctions continues en p, le critère est alors exprimé comme suit :

$$J_W(\tau) = \int_{\Gamma(\tau)} d^2(X, W) \left| \frac{\partial X}{\partial p} \right| dp \tag{5}$$

L'hypothèse de continuité signifie que nous ne coupons pas le squelette des régions LPE. Cette hypothèse nous permet d'obtenir l'EDP, ensuite nous généraliserons en étendant la force d'évolution aux cas singuliers. Notons toutefois que cette étape de segmentation est réalisée en vue d'affiner un premier pas, les chances de couper le squelette sont fortement réduites. Le gradient en chaque point de la carte de distance est donné par $\mathbf{N_W} = \nabla d = \frac{X-Y}{|X-Y|}$. Notons que $d = |u|$ et que

$$\nabla u = \begin{cases} -\frac{X-Y}{|X-Y|} & \text{si } X \text{ est à l'extérieur de } \Gamma^{ref} \\ \frac{X-Y}{|X-Y|} & \text{si } X \text{ est à l'intérieur de } \Gamma^{ref} \end{cases} \tag{6}$$

En définissant par N la normale unitaire interne à Γ, nous obtenons $< \nabla d, \mathbf{N} > d =< \nabla u, \mathbf{N} > u$. La dérivée de (5) s'obtient alors en exploitant les résultats de [6] :

$$J'_W(\tau) = \int_0^1 < \frac{\partial X}{\partial \tau}, \left[2 < \mathbf{N_W}, \mathbf{N} > d - \kappa d^2 \right] \mathbf{N} > \left| \frac{\partial X}{\partial p} \right| dp \tag{7}$$

où κ est la courbure du contour. A partir dérivée, nous obtenons l'équation d'évolution suivante :

$$\frac{\partial \Gamma_W}{\partial \tau} = \left[-2 < \mathbf{N_W}, \mathbf{N} > d + \kappa d^2 \right] \mathbf{N} = F_W \mathbf{N} \tag{8}$$

Nous généralisons cette équation d'évolution aux points singuliers. En $d = 0$, le point d'évolution X appartient aux LPE et $\mathbf{N_W}$ n'est pas défini. Néanmoins, les LPE sont atteintes et la force d'attraction doit être nulle. Aussi nous définissons la force d'évolution F_W en $d = 0$ par $F_W = 0$. Si il existe plus d'une valeur Y_{min} pour un X donné, le point X est alors un point singulier. Dans ce cas, X se situe sur le squelette de la région et $\mathbf{N_W}$ n'est pas défini de manière unique. Le choix de $< \mathbf{N_W}, \mathbf{N} >$ dépend de l'implémentation et de la discrétisation. Numériquement, nous calculons $\mathbf{N_W}$ par un gradient centré pour éviter de privilégier une mauvaise direction. Comme partout, la courbe évolue dans la direction de N, mais la force et le sens de propagation sont surtout influencés par κ (puisqu'il est facteur de d^2 et que d est grand sur le squelette).

3.2 Segmentation en deux étapes

3.2.1 Premier pas : segmentation par contours actifs par détection d'objets en mouvement

La première étape met en oeuvre la détection des objets en mouvement pour un simple critère défini dans [7] par :

$$J_M(\tau) = \iint_{\Omega_{obj}(\tau)} \alpha_{obj} dxdy + \iint_{\Omega_{back}(\tau)} \alpha_{back} |I_n - B_n| dxdy + \int_{\Gamma(\tau)} \lambda ds \tag{9}$$

où α_{obj}, α_{back} et λ sont des constantes positives. B_n représente l'estimation du fond et I_n la trame courante. L'équation d'évolution est définie par :

$$\frac{\partial \Gamma_M}{\partial \tau} = (\alpha_{obj} - \alpha_{back} |I_n - B_n| + \lambda \kappa) \mathbf{N} \tag{10}$$

Cette étape conduit le contour actif près des bords de l'objet à segmenter. Cependant, la précision peut être améliorer en affinant le résultat par une seconde évolution.

3.2.2 Second pas : Segmentation par contours actifs conduits par une force d'attraction aux LPE

Pré-segmentation par LPE

La technique d'immersion introduite dans [16] retourne une première partition de l'image en régions de couleurs homogènes. Le nombre de régions est ensuite réduit par le processus de fusion présenté en [5]. Ce processus est basé sur un critère spatio-temporel. Aussi les LPE ne correspondent pas aux lignes de plus fort gradient couleur de l'image. Les LPE introduites dans notre algorithme correspondent aux bornes non-lisses des régions de la partition.

Application des contours actifs

Le second pas consiste à attirer le contour vers les LPE, et à contraindre la régularité du contour. Aussi, nous définissons un critère combinant l'énergie des distances aux LPE (introduit en 3.1) et un terme constant. Ce dernier a pour rôle de régulariser le contour final en contraignant la longueur du contour actif. L'énergie est exprimée comme suit :

$$J_W(\tau) = \int_{\Gamma(\tau)} \left[\lambda_W d^2(\Gamma(\tau), W) + \lambda_\Gamma\right] ds \qquad (11)$$

où λ_W et λ_Γ sont deux constantes positives. L'équation (8) et la dérivée de (11) retournent l'équation d'évolution :

$$\frac{\partial \Gamma_W}{\partial \tau} = \left[\lambda_W \left(-2 < \mathbf{N_W}, \mathbf{N} > d + \kappa d^2\right) + \lambda_\Gamma \kappa\right] \mathbf{N}. \quad (12)$$

Le premier terme attire le contour vers les LPE et le second le lisse. Dans le cas où le contour actif coupe le squelette des régions LPE, le problème est traité comme évoqué dans 3.1. Le terme de régularité et la rigidité de la courbe permettent de ramener les points de part et d'autre du squelette. Notons toutefois, que l'algorithme est appliqué aux résultats finals de la première étape de segmentation, aussi le contour est proche des LPE. Le risque de couper le squelette est ainsi réduit.

4 Résultats :

Les performances de la méthode proposée sont ici évaluées sur la séquence *Akiyo* pour $\alpha_{obj} = 0.9$, $\alpha_{back} = 1.2$, $\lambda = 6$, $\lambda_W = 0.2$ et $\lambda_\Gamma = 1.5$ et la séquence *Mother* (Fig. 2) for $\alpha_{obj} = 1.5$, $\alpha_{back} = 1.2$, $\lambda = 9$, $\lambda_W = 0.2$ et $\lambda_\Gamma = 2.25$. Les figures 1 (a), (b) et (c) présentent des résultats intermédiaire et final retournés par le premier pas de l'algorithme. Nous observons quelques disparités entre le contours et le bord de l'objet. La figure 1 (d) présente l'image sur-segmentée par les LPE, à partir de laquelle la carte de distance est calculée. Les figures 1 (e) et (f) montrent la propagation et les résultats retournés par le second pas. Nous constatons une nette amélioration des résultats de la première segmentation tant du point de vue de la finesse que de la régularité.

5 Conclusion

Dans cet article, nous avons présenté une nouvelle méthode de segmentation vidéo combinant les méthodes par contours actifs et par LPE. L'algorithme proposé est réalisé séquentiellement en deux étapes. La première exploite un critère de mouvement. La seconde attire le contours vers la ligne de partage des eaux la plus proche. Ces lignes séparent les régions de couleurs homogènes. La partie originale des présents travaux est la définition d'un critère d'énergie basé sur les LPE. Les résultats obtenus sont performants. Il est toutefois à noter que la détection de mouvement doit être suffisamment bonne pour que les LPE globalement les plus proches correspondent à la meilleure solution. Cet approche est applicable pour la détection d'objet en mouvement pour une caméra statique. Une alternative pour des travaux ultérieurs est d'étendre la méthode aux séquences à caméra mobile en changeant le premier critère.

6 Remerciement

Les auteurs remercient A. Herbulot, E. Debreuve et le Prof Barlaud pour leur aide dans la réflexion théorique et la région wallonne pour la réalisation de ce projet.

Références

[1] V. Caselles, R. Kimmel et G. Sapiro, *Geodesic active contours*, IJCV, vol 22, n 1, p 61-79, 1997.

[2] Chan, T. et Vese, L., *Active contours without edges*, IEEE Trans. on Im. Process., vol 10, n 2, p 266-277, 2001.

[3] L. Cohen, E. Bardinet et N. Ayache, *Surface reconstruction using active contour models*, SPIE Conf. on Geom. Meth. in Comput. Vis., 1993.

[4] O. Cuisenaire et B. Macq, *Fast Euclidean distance transformations by propagation using multiple neighbourhoods, Computer vision and Image understanding*, Comput. vis. and Im. understanding, vol 76, n 2, p 163-172, 1999.

[5] C. De Roover, M. Gabbouj et B. Macq, *An accurate semiautomatic segmentation scheme based on watershed and change detection mask*, Proc. of SPIE, Im. and Video Comm. and Process., vol. 5685, jan, 2005.

[6] M. Gastaud, M. Barlaud et G. Aubert, *Combining Shape Prior and Statistical Features For Active Contour Segmentation*, IEEE TCSVT spec. sess. on Audio and Video Anal. for Interact. Multim. Services, mai, 2004.

[7] S. Jehan-Besson, M. Barlaud, G. Aubert, *A 3-Step Algorithm using Region-Based Active Contours for Video Objects Detection*, EURASIP JASP, 2002.

[8] S. Jehan-Besson, M. Barlaud et G. Aubert, *DREAM2S : Deformable Regions driven by an Eulerian Accurate Minimization Method for image and video segmentation*, Int. Journ. of Comput. Vis., vol 53, n 1, 2003.

[9] Kass, M., Witkin, A. et Terzopoulos, D., *Snakes : Active contour models*,IJCV ; vol 1, p 321-332, 1988.

[10] F. Meyer et S. Beucher, *Morphological segmentation*, Journal of Visual Comm. and Image Rep., vol 1, n 1, p.21-46, 1990.

[11] H.T. Nguyen, M. Worring et R. van den Boomgaard, *Watersnakes : energy-driven watershed segmentation*, IEEE Trans. on Pat. Anal. and Mach. Intell., vol 25, n3, p. 300-342,2003.

[12] N. Paragios et R. Deriche, *Geodesic Active Regions and Level Set Methods for Supervised Texture Segmentation*,Int. Journ. of Comput. Vis., vol 46, n 3, 2002.

[13] J. Park et J.M. Keller, *Snakes on the Watershed*, IEEE Trans. On Pat. Anal. and Mach. Intell.,vol. 23, n 10, p. 1201-1205, 2001.

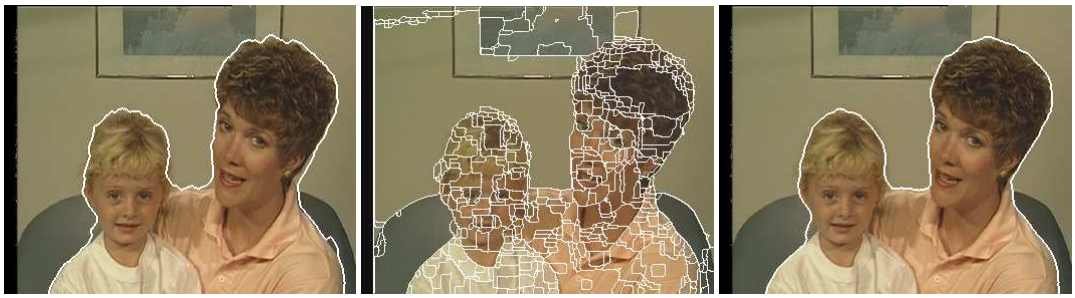
[14] R. Ronfard, *Region-based strategies for active contour models*, Int. Journ. of Comput. Vis., vol 13, n 2, p 229-251, 1994.

[15] P. Salembier et L. Garrido, *Binary Partition Tree as an Efficient Representation for Image Processing,Segmentation, and Information Retrieval*, IEEE Trans. on Im. Process., vol 9, n 4, 2000.

(a) Contour initial (b) Pas 1 : résultat intermédiaire (c) Pas 1 : Segmentation finale, contour initial du 2ème pas

(d) Pas 2 : pré-segmentation LPE (données) (e) Pas 2 : résultat intermédiaire (f) Pas 2 : segmentation finale

FIG. 1 – Segmentation en deux phases (Akiyo) : (b) propagation et (c) résultat de la segmentation d'objets en mouvement (pas 1) ; (e) propagation et (f) résultat de la segmentation par attraction aux LPE (pas 2)



(a) Pas 1 : résultat (b) Pas 2 : Pré-segmentation LPE (c) Pas 2 : segmentation finale

FIG. 2 – Segmentation de *Mother* : (a) segmentation par détection d'objet en mouvement, (b) LPE servant à calculer la carte de distance, (c) Résultat après les 2 étapes de segmentation.

[16] L. Vincent et P. Soille, *Watersheds in Digital Spaces : An Efficient Algorithm Based on Immersion Simulations*, IEEE Trans. on Pat. Anal. and Mach. Intell., vol 13, n 6, 1991.

[17] A. Yezzi, A. Tsai et A. Willsky, *A statistical approach to snakes for bimodal and trimodal imagery*, IEEE ICIP, Kobe, 1999.

[18] S Zhu et A. Yuille, *Region competition : unifying snakes, region growing, and Bayes/MDL for multiband image segmentation*, PAMI, vol 18, p 884-900, sept. 1996.

Détection de signalisation routière par ajustement de formes prototypes

Guillaume DUTILLEUX, Pierre CHARBONNIER
Laboratoire des Ponts et Chaussées de Strasbourg
11 rue Jean Mentelin, BP 9, 67035 STRASBOURG Cedex, FRANCE
guillaume.dutilleux@equipement.gouv.fr
pierre.charbonnier@equipement.gouv.fr

Résumé – Nous présentons une approche originale en matière de détection automatique robuste de signalisations verticales de danger en lumière naturelle. Le détecteur s'appuie sur la combinaison de formes prototypes et d'un algorithme d'optimisation globale de type stratégie d'évolution hybridé avec une méthode locale de type gradient conjugué. L'étape de pré-détection fait appel à un calcul de dominante chromatique robuste face aux éclairements incontrôlés. L'emploi d'une forme prototype adaptée au problème permet de s'affranchir du bruit. La fonction minimisée est la somme de termes d'*a priori* et de deux termes d'attache aux données. Le réglage du seuil de détection et l'évaluation globale du détecteur s'appuient sur le tracé de courbes COR.

Abstract – We present an original approach regarding the robust automatic detection of vertical danger road signs under natural light. The detector is based on the combination of template shapes and an optimization algorithm called evolution strategy hybridized with a local search method of the conjugate gradient kind. The pre-detection step uses a computation of chromatic dominance robust to uncontrolled illumination conditions. Using a template shape helps reducing the sensitivity to noise. The cost function is the sum of *a priori* terms and of two likelihood terms. The adjustment of the detection threshold and the global performance evaluation of the detector are founded on ROC curves.

1 Introduction

La détection et la reconnaissance automatiques de panneaux de signalisation font l'objet d'un intérêt soutenu depuis plusieurs années (voir par exemple [4] pour une bibliographie récente), en relation avec l'alimentation de bases de données d'images routières que les gestionnaires d'infrastructures utilisent en particulier pour l'inventaire de signalisation et les études de sécurité. Les systèmes d'aide à la conduite sont un autre domaine d'application.

Cet article traite de la conception d'un système de détection hors-ligne qui fonctionne sur des séquences d'images enregistrées en lumière naturelle visible. On traite ici le cas des panneaux de danger. Le système de détection s'accommode bien d'une illumination incontrôlée ou d'images où la balance des blancs est incorrecte. Le système est aussi conçu pour être robuste vis-à-vis des occultations et des signaux endommagés. L'approche choisie est basée sur la combinaison de formes prototypes [1][2] et d'une stratégie d'évolution (ES), qui représente une alternative intéressante aux algorithmes génétiques [5][6][7]. Cet algorithme d'optimisation globale est *hybridé* avec une méthode de recherche locale de type gradient, dans le but d'obtenir une convergence fine. Dans ce qui suit, les caractéristiques principales du détecteur sont d'abord présentées. Les aspects de traitement d'image sont d'abord décrits. Ils concernent la pré-détection robuste de la composante rouge, la définition d'une forme prototype adaptée au problème et du critère optimisé. Ensuite, les caractéristiques de l'algorithme d'optimisation choisi sont explicitées. La dernière partie fournit des résultats pour des images réelles et illustre la fiabilité du détecteur global.

2 Le système de détection

2.1 Traitement d'images

2.1.1 Pré-détection

Pour identifier les pixels présentant une dominantes rouge, on travaille dans l'espace RGB. Le pré-détecteur considère qu'un pixel est rouge si ses composantes vérifient

$$\begin{aligned} R &> \alpha(G+B) \\ R - \max(G,B) &> 2\alpha[\max(G,B) - \min(G,B)] \end{aligned} \quad (1)$$

Le premier critère revient à poser une contrainte sur la composante rouge normalisée, $R/(R+G+B)$, qui doit être dominante. La normalisation donne sa robustesse au pré-détecteur face aux variations d'illumination. Le deuxième critère vérifie que le pixel ne tende ni vers le jaune, ni vers le magenta. Le paramètre α peut être ajusté globalement pour une séquence d'images. Ses valeurs habituelles varient de 0.5 à 0.75. On privilégie des valeurs de α relativement faibles pour ne pas risquer de voir un panneau disparaître lors de la pré-détection. En contrepartie, le bruit dans l'image de pré-détection obtenue est plus important, mais l'utilisation d'une forme prototype adaptée (Cf. 2.1.2) permet de s'affranchir de ce problème. Enfin, une balance des blancs incorrecte est malheureusement assez fréquente dans les séquences à traiter. Comme les images analysées contiennent toujours une portion de route, une égalisation de type « monde gris » [3] peut être effectuée automatiquement dans la plupart des situations.

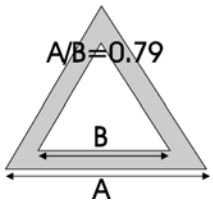


FIG. 1 – Forme prototype

2.1.2 Forme prototype

Pour la détection de panneaux de danger, une forme prototype particulière est utilisée (Cf. figure 1), dans le but de minimiser les fausses détections. Les deux triangles emboîtés définissent une couronne où l'on peut rechercher des pixels rouges.
La forme prototype est ajustée sur l'image à l'aide de trois transformations :
- mise à l'échelle,
- rotation,
- translations horizontale et verticale.

Ceci définit un espace de recherche à 4 dimensions. Un jeu de paramètres de transformation est représenté par la variable Θ dans la suite.

2.1.3 Fonction d'énergie

La fonction d'énergie est définie classiquement comme une somme de termes d'*a priori* et de termes de vraisemblance. Le calcul de l'*a priori* sur la forme obtenue par transformation concerne les points suivants :
- la position usuelle des panneaux dans l'image (dans la moitié droite),
- la gamme de taille possible pour un panneau,
- l'orientation usuelle d'un panneau.

Chaque critère sur un paramètre x supposé se trouver dans un intervalle de pertinence $[x_{min}, x_{max}]$ est représenté par une fonction de pénalité en forme de vallée :

$$\begin{aligned} x < x_{min},\ g(x) &= a(x_{min} - x) \\ x_{min} \leq x \leq x_{max},\ g(x) &= 0 \\ x > x_{max},\ g(x) &= b(x - x_{max}) \end{aligned} \tag{2}$$

Les bornes des différents intervalles de pertinence peuvent être considérées comme des données établies par la « vérité terrain ». Les différentes pentes a et b ont été réglées empiriquement.

La terme global d'*a priori* est une somme pondérée des différents termes de pénalité.

La vraisemblance se compose de deux termes : un terme orienté contour et un terme orienté région. Le terme orienté contour est largement inspiré de [1]. Il dépend de l'angle entre le gradient local de l'image et le vecteur directeur local de la forme prototype. Sa valeur approche 0 quand le modèle est aligné sur les contours d'un objet d'intérêt et 1 quand le modèle est sur un domaine homogène. Ce terme est calculé le long du triangle extérieur. Pour une forme Θ, le terme orienté région est calculé sur la carte binaire de pré-détection de dominante rouge. Il se compose de deux termes reliés aux deux zones du modèle choisi : la couronne $\partial\Omega$ et l'intérieur du petit triangle Ω (voir figure 2) :

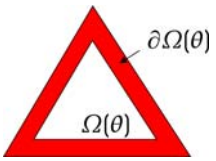


FIG. 2 – Domaines d'intégration

$$U_c(\Theta) = \frac{\int\int_{\partial\Omega(\Theta)} I(x)dx}{\int\int_{\partial\Omega(\Theta)} dx} \quad \text{et} \quad U_i(\Theta) = \frac{\int\int_{\Omega(\Theta)} I(x)dx}{\int\int_{\Omega(\Theta)} dx} \tag{3}$$

Le terme global orienté région s'écrit :

$$U_t(\Theta) = \min(1, 1 + U_c(\Theta) - U_i(\Theta)) \tag{4}$$

Quand Θ est parfaitement ajusté, U_t vaut zéro. Au contraire, $U_t(\Theta)$ est égal à 1 lorsque Θ est sur une zone uniforme « rouge » ou « non rouge » de l'image.

2.2 Optimisation

2.2.1 Optimisation globale par stratégie d'évolution

La fonction d'énergie globale obtenue en sommant les termes d'*a priori* et de vraisemblance est hautement multimodale. De ce fait, un algorithme d'optimisation globale est nécessaire. Ici, c'est une stratégie d'évolution (ES) qui a été choisie [5]. Cet algorithme appartient à la classe des algorithmes évolutionnaires, au même titre que les algorithmes génétiques (GA) [7]. Comme les GA, les ES manipulent une population de solutions candidates, et utilisent des opérateurs de croisement, mutation et sélection. A la différence des GA, les ES manipulent des vecteurs aléatoires gaussiens plutôt que des chaînes binaires. De ce fait, elles ne nécessitent pas de codage/décodage des paramètres à optimiser. Ce fondement probabiliste permet à la théorie de la convergence des ES d'être mieux établie que celle des GA [6]. Un autre point intéressant par rapport aux GA est que les ES incorporent naturellement un mécanisme d'auto-adaptation des paramètres internes, que l'on peut rapprocher du schéma de refroidissement du recuit simulé [8]. Ainsi la probabilité de mutations de grande amplitude décroît au fil des itérations.

Pour le problème d'optimisation discuté ici, une configuration à 4 parents et 7 fois plus de descendants à été retenue. Comme dans [4], la population initiale est basée sur une analyse des composantes connexes de l'image de dominante rouge. Lorsque le nombre de composantes connexes est insuffisant, les individus sont obtenus par tirage de variables aléatoires uniformes centrées et mises à l'échelle relativement à la définition des termes d'*a priori*. En remplacement de la sélection déterministe, un opérateur mixte roulette/déterministe est préféré ici, de façon à réduire la probabilité de convergence prématurée.

2.2.2 Recherche locale par gradient conjugué

En pratique, la convergence est habituellement très rapide au début de l'optimisation. Plus tard, les solutions candidates ont souvent tendance à tourner autour de la solution sans véritablement la trouver. De façon à résoudre ce problème, l'ES a été hybridée ici avec une méthode locale, de type gradient conjugué. Ceci permet d'exploiter la topologie de la fonction optimisée au voisinage de la solution, et donc à l'optimisation de converger plus vite et plus finement.

Du fait de la complexité des expressions des termes d'attache aux données, le gradient est calculé numériquement, par différences finies. En outre, comme les termes d'*a priori* ne sont pas différentiables sur la totalité de leur domaine de définition, le gradient est calculé sur le terme de vraisemblance uniquement. Ceci n'est pas préjudiciable, étant donné que la recherche locale présente surtout un intérêt au voisinage de la solution, donc dans une zone où l'*a priori* est nul.

3 Résultats

La figure 3 montre une image où le panneau est à peine visible du fait d'une situation de fort contre-jour, ainsi que de la présence de végétation et d'une levée de terrain sur le bas-côté. La figure 4 montre l'image de pré-détection de rouge, où le panneau est bien visible. Avec une telle image, la détection automatique ne pose pas de difficulté.

En guise d'illustration du fonctionnement global de l'algorithme, la figure 5 montre une image originale, la figure 6 l'image de pré-détection de rouge associée, la figure 7 le résultat de l'optimisation. Ici, l'image de pré-détection contient de nombreux leurres potentiels pour le détecteur. L'utilisation d'une forme prototype spécifique permet de les éviter.

Enfin, nous montrons figure 8 une comparaison des courbes COR (Caractéristique Opérationnelle du Récepteur) obtenues pour différents scénarios d'utilisation de l'algorithme, à partir des résultats de 2 analyses d'une séquence de 52 images, dont 32 contenant des panneaux. Pour le calcul de ces courbes, une vérité-terrain concernant la position des objets d'intérêt a été obtenue par analyse manuelle de la séquence d'images. A la fin de l'optimisation, soit 500 itérations de l'ES, on considère la valeur résiduelle de la fonction d'énergie correspondant au prototype optimal : on considère qu'il y a eu détection d'un panneau lorsque celle-ci est inférieure à un certain seuil. Si la position de l'objet est dans la zone de tolérance définie autour de la vérité-terrain (qui peut être vide s'il n'y a effectivement pas de panneau dans l'image), il s'agit d'une bonne détection, sinon, il s'agit d'une fausse alarme. Pour cette évaluation, les coordonnées des centres des formes sont arrondies au pixel près et la zone de tolérance a un rayon de 2 pixels. La probabilité de bonne détection est calculée comme le rapport du nombre de détections valides sur le nombre de panneaux présents dans la séquence. La probabilité de fausse alarme est le rapport du nombre de détections erronées sur le nombre total de fausses détections possibles. Ces probabilités sont représentées sous formes de courbes, paramétrées par la valeur du seuil de détection. On utilise ici une échelle semi-logarithmique pour améliorer la lisibilité des courbes COR. La courbe idéale correspondrait au cas de 100% de bonnes détections pour 0% de fausses alarmes. Les courbes montrent la supériorité d'une initialisation tenant compte des composantes connexes issues de l'étape de pré-détection du rouge, par rapport à une initialisation purement aléatoire. D'autre part, elles illustrent l'intérêt de l'approche hybride ES-gradient conjugué par rapport à l'algorithme ES classique.

4 Perspectives

Un détecteur de panneaux de signalisation à partir de prises de vues en lumière naturelle visible a été présenté. Il se montre peu sensible aux conditions d'illumination grâce à un pré-détecteur de dominante rouge robuste. L'emploi d'une stratégie d'évolution hybridée avec une méthode locale permet d'affiner la convergence. L'application du détecteur sur des séries de plusieurs dizaines d'images montre des résultats prometteurs en termes de robustesse.

Le développement du détecteur se porte actuellement sur le traitement efficace du cas général de la détection de signalisation : identification précoce de l'absence de panneau, traitement des images contenant plusieurs panneaux, en vue de l'utilisation du détecteur en phase opérationnelle.

Références

[1] M.P. Dubuisson Jolly, S. Lakshmanan et A.K. Jain, "Vehicle segmentation and classification using deformable templates," *IEEE Trans. On Pattern Analysis and Machine Intelligence,* Vol. 18, N°3, pp. 293-308, mars 1996.

[2] M. Mignotte, C. Collet, P. Perez et P. Bouthemy, "Hybrid genetic optimization and statistical model-based approach for the classification of shadow shapes in sonar imagery," *IEEE Trans. On Pattern Analysis and Machine Intelligence,* Vol. 22, N°2, pp. 129-141, février 2000.

[3] G. Buchsbaum, "A Spatial Processor Model for Object Colour Perception", *Journal of the Franklin Institute*, 310 :1-26, 1980.

[4] A. de la Escalera, J.M. Armingol, J.M. Pastor, et F.J. Rodriguez, "Visual sign information extraction and identification by deformable models for intelligent vehicles," *IEEE Trans. on Intelligent Transportation Systems,* Vol. 5, N°2, pp. 57-68, 2 juin 2004.

[5] I. Rechenberg, *Evolutionsstrategie'94* Fromann-Holzboog, 1994 (en allemand).

[6] T. Bäck, *Evolutionary algorithms in theory and practice,* Oxford university Press, 1996.

[7] Z. Michalewicz, *Genetic Algorithmes + Data Structures = Evolution Programs*, Springer, 1999.

[8] J. Dréo, A. Pétrowski, P. Siarry, E. Taillard, *Métaheuristiques pour l'optimisation difficile*, Eyrolles, 2003.

FIG. 3 – Image originale

FIG. 4 – Pré-détection de rouge

FIG. 5 – Image originale

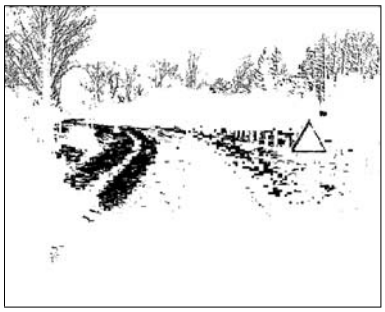


FIG. 6 – Pré-détection de rouge

FIG. 7 – Résultat de l'optimisation

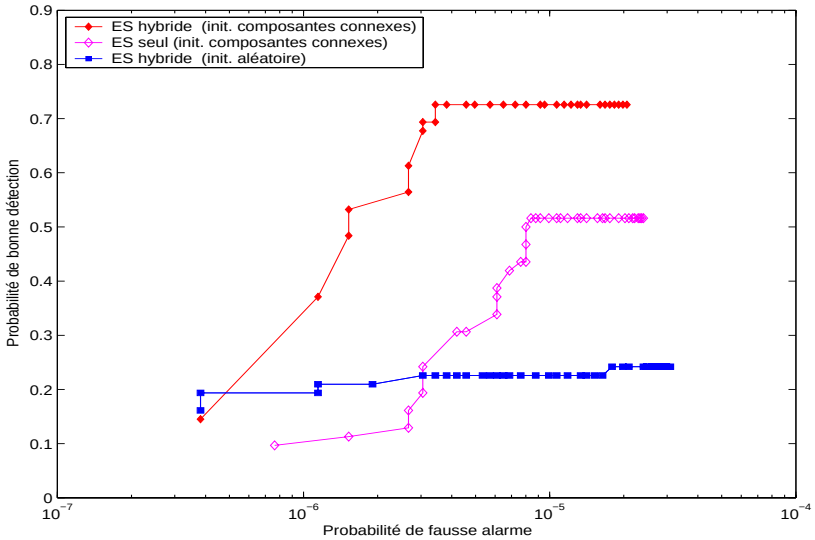


FIG. 8 – Évaluation de l'algorithme : courbes COR

Utilisation des Gradients de Forme et des Contours Actifs Basés Régions pour la Segmentation des Vecteurs Mouvement

Stéphanie Jehan-Besson[1], Stefan Duffner[1*], Ariane Herbulot[2], Michel Barlaud[2], Gilles Aubert[3]

[1]Laboratoire GREYC, 6, Bd du Marechal Juin, 14050 Caen Cedex, France

[1*]maintenant à France Telecom R&D, Rennes

[2]Laboratoire I3S, 2000 route des Lucioles, 06903 Sophia Antipolis, France

[3]Laboratoire J.A. Dieudonné, Parc Valrose, 06108 Nice Cedex 2, France

jehan@greyc.ensicaen.fr, stefan.duffner@rd.francetelecom.com
herbulot@i3s.unice.fr, barlaud@i3s.unice.fr
gaubert@math.unice.fr

Résumé – L'objectif est ici la segmentation de régions homogènes en utilisant des caractéristiques vectorielles. L'application visée est la segmentation des régions en mouvement en utilisant les vecteurs mouvement. Afin de caractériser l'homogénéité de la région en utilisant toutes les composantes des vecteurs mouvement, nous estimons puis minimisons l'entropie jointe de ces composantes. Cette quantité est évaluée en utilisant la densité de probabilité jointe des composantes (celle-ci est non paramétrique puisque nous ne faisons aucune hypothèse sur la distribution sous-jacente). Sur les résultats synthétiques, on peut montrer que l'on arrive à différencier des régions en prenant en compte les deux composantes des vecteurs mouvement. Ainsi on peut différencier des vecteurs de même longueur mais de directions différentes. Sur des séquences vidéo réelles, des premiers résultats prometteurs ont été obtenus pour la segmentation d'objets en mouvement.

Abstract – In this paper, we propose to focus on the segmentation of vector fields (e.g. motion vectors) using region-based active contours. We search for a domain that minimizes a criterion based on homogeneity measures of the vector fields. We choose to evaluate, within each region to be segmented, the average quantity of information carried out by the vector fields, namely the joint entropy of vector components. We do not make any assumption on the underlying distribution of joint probability density functions (pdfs) of vector components, and so we evaluate the entropy using non parametric pdfs. A local shape minimizer is then obtained through the evolution of a deformable domain in the direction of the shape gradient. Shape derivation tools allow us to easily derive complex criteria involving such non parametric pdfs. We first apply this segmentation method to synthetic vector fields and show interesting results where motion vector fields may be separated using both their length and their direction. Then, optical flow is estimated in real video sequences and segmented using this technique. This leads to promising results for the segmentation of moving video objects.

1 Introduction

Dans ce papier, nous proposons de segmenter des champs de vecteurs mouvement en utilisant les contours actifs basés régions. L'objectif est de segmenter les différents mouvements dans une séquence d'images. Les applications pratiques sont nombreuses : post-production cinématographique, vidéo-surveillance, indexation multimédia ou compression orientée objet. Les approches variationnelles et les contours actifs basés régions ont montré leur efficacité pour la segmentation du mouvement. Certains auteurs choisissent de minimiser l'image des différences [6], alors que d'autres considèrent des modèles paramétriques pour chaque région [2]. Une autre approche consiste en l'utilisation de la longueur des vecteurs mouvement [9] ou du mouvement dominant [10].

Dans l'approche proposée ici, nous considérons à la fois la longueur et la direction des vecteurs mouvement en utilisant l'entropie jointe des composantes des vecteurs mouvement. Nous ne faisons aucune hypothèse sur la distribution de probabilité suivie par ces composantes et nous choisissons donc d'évaluer l'entropie en utilisant des densités de probabilité non paramétriques. Un domaine optimal est obtenu par l'évolution d'un modèle déformable. Cette évolution est régie par une équation aux dérivées partielles calculée en dérivant le critère proposé. Comme la minimisation se fait ici par rapport à un domaine, nous utilisons pour la dérivation les outils de dérivation de domaine comme cela a été proposé dans [7, 1].

Des résultats sur des images synthétiques montrent que les champs de vecteurs peuvent être séparés en utilisant à la fois leur longueur et leur direction. Pour les séquences vidéo réelles, le flot optique est tout d'abord estimé puis mis à profit pour la segmentation des objets en mouvement. Des premiers résultats encourageants ont été obtenus.

2 Introduction d'un critère basé sur la théorie de l'information

Nous proposons d'utiliser un critère basé sur l'entropie présenté dans [4] et appliqué ici à la segmentation de données vectorielles et notamment des vecteurs mouvement. Notons de manière générale $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), .., f_n(\mathbf{x})]^T$ la caractéristique d'intérêt (qui peut être par exemple le vecteur des composantes couleur de l'intensité de l'image) et $q(\mathbf{f}(\mathbf{x}), \Omega)$ la densité de probabilité d'avoir $\mathbf{f}(\mathbf{x})$ sachant que $\mathbf{x} \in \Omega$. L'entropie jointe des composantes f_i a alors pour expression :

$$E(f_1, f_2, .., f_n, \Omega) = \int_\Omega -q(\mathbf{f}(\mathbf{x}), \Omega)\ \ln q(\mathbf{f}(\mathbf{x}), \Omega)\, \mathrm{d}\mathbf{x}$$
$$= \int_\Omega -q(f_1(\mathbf{x}), .., f_n(\mathbf{x}), \Omega)\ \ln q(f_1(\mathbf{x}), .., f_n(\mathbf{x}), \Omega)\, \mathrm{d}\mathbf{x}$$

Afin de segmenter les régions homogènes, nous utilisons une compétition entre l'homogénéité de deux régions, le fond noté Ω_{out} et les objets d'intérêt notés Ω_{in}. Dans notre étude, nous cherchons la partition de l'image qui minimise le critère suivant :

$$\begin{aligned} E_C(\Omega_{in}, \Omega_{out}) &= E(f_1, .., f_n, \Omega_{in}) \cdot |\Omega_{in}| \\ &+ E(f_1, .., f_n, \Omega_{out}) \cdot |\Omega_{out}| \end{aligned} \quad (1)$$

Ce critère est basé sur la maximisation de l'information mutuelle entre $\mathbf{f}(\mathbf{x})$ et le label de la région $L(\mathbf{x})$, voir [8, 4] pour plus de détails. Il correspond à l'entropie conditionnelle entre la caractéristique $\mathbf{f}(\mathbf{x})$ et le label $L(\mathbf{x})$.

2.1 Estimation de la densité de probabilité non paramétrique

Afin d'évaluer ce critère puis de le dériver, nous devons estimer la densité de probabilité $q(\mathbf{f}(\mathbf{x}), \Omega)$. Pour cela nous ne faisons pas d'hypothèse sur la distribution de probabilité sous jacente et nous utilisons une densité de probabilité non paramétrique. Celle-ci est estimée en utilisant la méthode de Parzen, ce qui nous donne l'expression suivante :

$$q(\mathbf{f}(\mathbf{x}), \Omega) = \frac{1}{|\Omega|} \int_\Omega K(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\hat{\mathbf{x}}))\, \mathrm{d}\hat{\mathbf{x}} \quad (2)$$

avec K un noyau Gaussien de moyenne 0 et de variance σ.

2.2 Calcul de l'équation d'évolution

Nous cherchons ici un domaine optimal, or l'ensemble des domaines de $\mathbb{R}^n$ n'a pas une structure d'espace vectoriel. Afin de dériver le critère, nous utilisons donc les outils de dérivation de domaines développés par [3] et mis à profit pour la segmentation par contours actifs dans [5, 7]. Nous calculons ainsi la dérivée Eulérienne du critère en suivant la méthodologie exposée dans [7, 4]. Nous pouvons alors calculer la dérivée Eulérienne dans la direction $\mathbf{V}$ de $J(\Omega) = E(\mathbf{f}, \Omega)\,|\Omega|$, ce qui nous donne :

$$dJ_r(\Omega, \mathbf{V}) = \int_{\partial\Omega} \Big[K\big(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{s})\big)\big(\ln q(\mathbf{f}(\mathbf{x}), \Omega) + 1\big) - 1 + |\Omega|\, q(\mathbf{f}(\mathbf{s}), \Omega) \ln q(\mathbf{f}(\mathbf{s}), \Omega)\Big] (\mathbf{V} \cdot \mathbf{N})\, \mathrm{d}\mathbf{s}$$

De cette dérivée, nous pouvons déduire qu'afin d'obtenir un minimum local du critère (1), nous allons faire évoluer une courbe $\Gamma(s, \tau)$, avec s la longueur d'arc et τ le paramètre d'évolution selon l'équation d'évolution suivante :

$$\begin{aligned} \frac{\partial \Gamma(s,\tau)}{\partial \tau} = \Big[& -|\Omega_{in}|\ q(\mathbf{f}(\hat{\mathbf{x}}), \Omega_{in}) \ln q(\mathbf{f}(\hat{\mathbf{x}}), \Omega_{in})\big) \\ - & \int_{\Omega_{in}} K\big(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\hat{\mathbf{x}})\big)\big(\ln q(\mathbf{f}(\mathbf{x}), \Omega_{in}) + 1\big)\, \mathrm{d}\mathbf{x} \\ + & |\Omega_{out}|\ q(\mathbf{f}(\hat{\mathbf{x}}), \Omega_{out}) \ln q(\mathbf{f}(\hat{\mathbf{x}}), \Omega_{out})\big) \\ - & \int_{\Omega_{out}} K\big(\mathbf{f}(\mathbf{x}) - \mathbf{f}(\hat{\mathbf{x}})\big)\big(\ln q(\mathbf{f}(\mathbf{x}), \Omega_{out}) + 1\big)\, \mathrm{d}\mathbf{x}\Big] \mathbf{N} \end{aligned}$$

avec $\hat{x} = \Gamma(s, \tau)$.

3 Application à la segmentation des vecteurs mouvement

Jusqu'à présent la caractéristique $\mathbf{f}(\mathbf{x})$ pouvant être utilisée n'a pas été spécifiée. Classiquement $\mathbf{f}(\mathbf{x})$ représente une ou plusieurs composantes de l'intensité de l'image [4]. Dans ce travail, nous choisissons les deux coordonnées du vecteur mouvement soit $\mathbf{f}(\mathbf{x}) = [u, v]^T$. En utilisant une telle approche, nous prenons en compte à la fois la longueur et la direction du mouvement comme cela est démontré dans les résultats expérimentaux suivants. Dans chacun des exemples, le contour initial et le contour final sont donnés.

3.1 Images synthétiques

La premier exemple (Figure 1) représente un rectangle se déplaçant en translation vers la droite sur un fond se déplaçant en translation vers la gauche. Le rectangle est ici correctement segmenté bien que les vecteurs mouvement soient de même longueur. Dans le deuxième exemple, trois rectangles de tailles et de mouvements différents sont placés sur un fond en translation. La Figure 2 nous montre l'évolution du contour. L'évolution des densités de probabilités associées au fond et à l'objet (soient respectivement $q(u, v, \Omega_{out})$ et $q(u, v, \Omega_{in})$) est donnée Figure 3. La première colonne nous donne l'évolution de $q(u, v, \Omega_{out})$ tandis que la deuxième colonne nous montre celle de $q(u, v, \Omega_{in})$. Dans la deuxième colonne, on voit clairement le pic $(0, 1)$ correspondant au mouvement du fond disparaître au fur et à mesure des itérations. Ainsi notre approche permet la segmentation de trois mouvement différents si le mouvement du fond est homogène. Le troisième exemple montre un zoom sur un disque placé sur un fond en translation vers la droite (Figure 4). Notre méthode permet de segmenter le disque.

3.2 Séquences vidéo réelles

Afin d'appliquer la méthode sur des séquences vidéo réelles, une estimation du mouvement doit être opérée. La segmentation sera d'ailleurs dépendante de la précision des vecteurs mouvement obtenus. Nous utilisons ici la méthode proposée par [12].

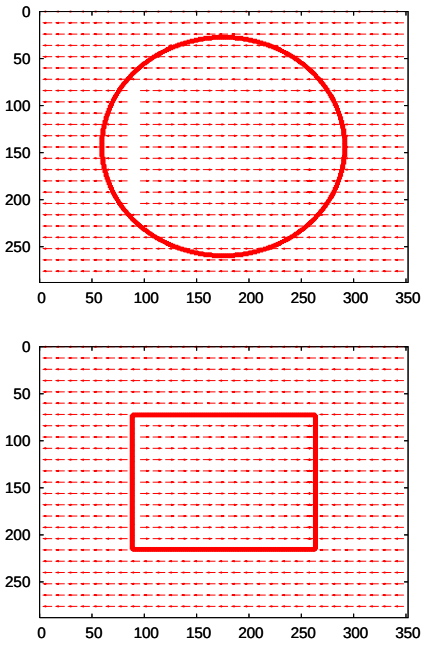


FIG. 1 – Segmentation d'un rectangle en translation

Nous noterons une image de la séquence par $I(x, y, t)$ où (x, y) représente la localisation dans l'espace image et t le temps. Nous cherchons alors à calculer le vecteur mouvement $(u, v)^T$ pour chaque pixel. L'hypothèse de base est la conservation de l'intensité dans le temps, ce qui après un développement limité se traduit par l'équation suivante :

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0 \tag{3}$$

Résoudre cette équation est un problème mal posé et une seconde contrainte de régularisation doit être ajoutée. Nous recherchons le flot optique (u, v) qui minimise la fonctionnelle suivante :

$$\int_\Omega \left(\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t}\right)^2 + \alpha\ \Psi\left(|\nabla u|^2 + |\nabla v|^2\right)\ \mathrm{d}x\,\mathrm{d}y \tag{4}$$

où Ψ est une fonction différentiable et croissante, ∇ l'opérateur gradient et α un paramètre de régularisation. Le minimum satisfait nécessairement les équations d'Euler-Lagrange et un minimum peut être calculé en plongeant les équations dans le schéma dynamique suivant :

$$\begin{aligned} u_k &= \nabla\cdot\left(\Psi'(|\nabla u|^2 + |\nabla v|^2)\nabla u\right) \\ &- \frac{1}{\alpha}\frac{\partial I}{\partial x}\left(\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t}\right) \end{aligned} \tag{5}$$

$$\begin{aligned} v_k &= \nabla\cdot\left(\Psi'(|\nabla u|^2 + |\nabla v|^2)\nabla v\right) \\ &- \frac{1}{\alpha}\frac{\partial I}{\partial y}\left(\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t}\right) \end{aligned} \tag{6}$$

avec k le paramètre d'évolution et Ψ' la dérivée de Ψ. Nous choisissons la fonction considérée par [11] et Weickert [12] :

$$\Psi(s^2) = \lambda^2\sqrt{1 + s^2/\lambda^2} \tag{7}$$

La paramètre λ est une constante positive.

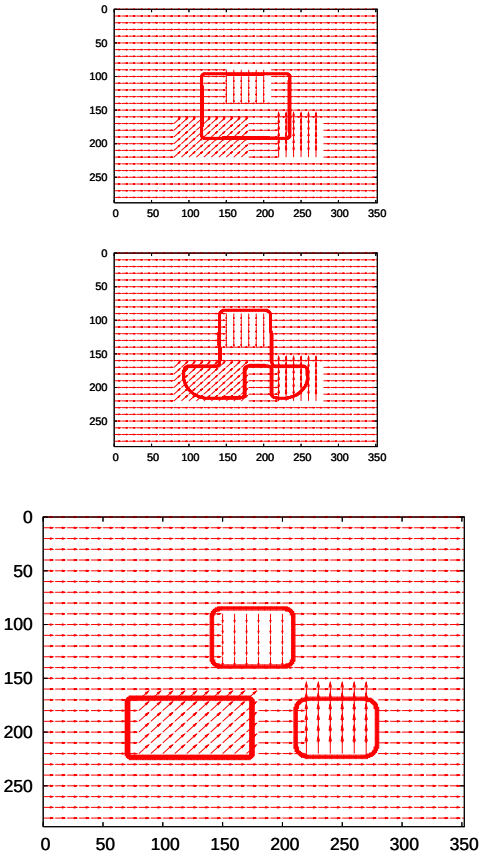


FIG. 2 – Segmentation de trois rectangles en mouvement

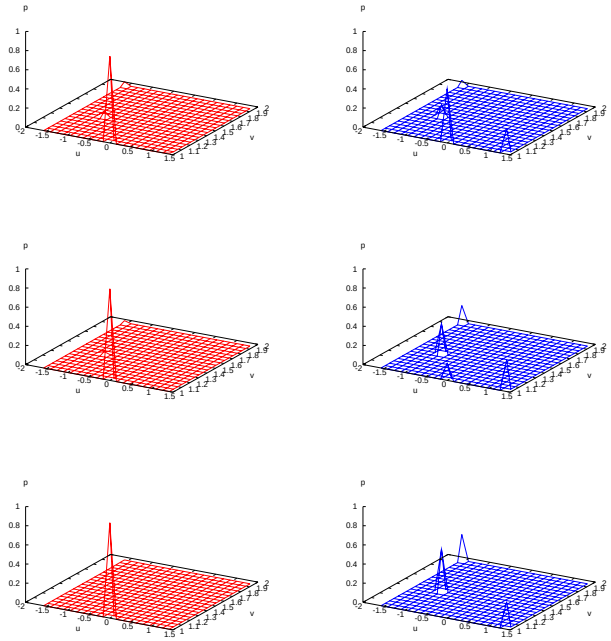
Notons que dans l'état actuel des recherches, le mouvement global de la caméra est compensé [6]. Les résultats obtenus sont montrés Fig.5 pour la séquence "edberg". L'objet en mouvement est ici correctement segmenté après compensation du mouvement de la caméra.

4 Conclusion

Dans cet article, nous proposons la segmentation de données vectorielles en minimisant un critère basé sur l'entropie jointe non paramétrique des composantes de ces vecteurs. Nous avons testé ce critère pour la segmentation de champs de vecteurs mouvement homogènes. Les résultats sur des images synthétiques montrent que l'on peut séparer des champs de vecteurs en utilisant à la fois leur longueur et leur direction. Ceci a ensuite été appliqué à la segmentation des objets en mouvement dans les séquences vidéo. Des premiers résultats prometteurs ont été obtenus en compensant le mouvement global de la caméra. Les recherches futures viseront à essayer de se passer de cette étape préliminaire pour la segmentation.

Références

[1] G. Aubert, M. Barlaud, O. Faugeras, and S. Jehan-Besson. Image segmentation using active contours : Cal-



a. $q(u, v, \Omega_{out})$ b. $q(u, v, \Omega_{in})$

FIG. 3 – Rectangles en mouvement : évolution des histogrammes au cours des itérations. La région Ω_{out} représente la région extérieure au contour.

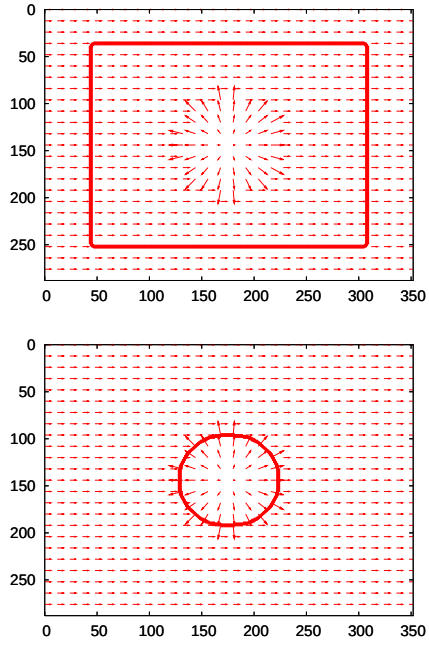


FIG. 4 – Segmentation d'un zoom sur un disque

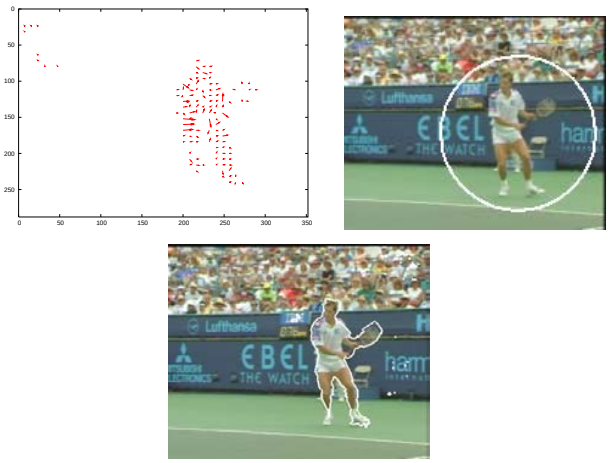


FIG. 5 – Résultats sur la séquence "edberg"

culus of variations or shape gradients ? *SIAM Applied Mathematics*, 63(6) :2128–2154, 2003.

[2] D. Cremers and S. Soatto. Variational space-time motion segmentation. In *IEEE Int. Conf. on Computer Vision*, volume 2, pages 886–892, October 2003.

[3] M.C. Delfour and J.P. Zolésio. *Shape and geometries*. Advances in Design and Control, SIAM, 2001.

[4] A. Herbulot, S. Jehan-Besson, M. Barlaud, and G. Aubert. Shape gradient for image segmentation using information theory. In *International Conference on Acoustics, Speech and Signal Processing*, Montreal, 2004.

[5] S. Jehan-Besson, M. Barlaud, and G. Aubert. Video object segmentation using eulerian region-based active contours. In *International Conference on Computer Vision*, Vancouver, Canada, 2001.

[6] S. Jehan-Besson, M. Barlaud, and G. Aubert. A 3-step algorithm using region-based active contours for video objects detection. *EURASIP Journal on Applied Signal Processing, Special issue on Image Analysis for Multimedia Interactive Services*, 2002(6) :572–581, June 2002.

[7] S. Jehan-Besson, M. Barlaud, and G. Aubert. DREAM^{2}S : Deformable regions driven by an eulerian accurate minimization method for image and video segmentation. *International Journal of Computer Vision*, (53) :45–70, 2003.

[8] J. Kim, J. Fisher III, A. Yezzi Jr., M. Cetin, and A. Willsky. Nonparametric methods for image segmentation using information theory and curve evolution. In *International Conference on Image Processing*, September 2002.

[9] F. Ranchin and F. Dibos. Moving objects segmentation using optical flow estimation. In *Workshop on Mathematics and Image Analysis*, Paris, september 2004.

[10] T. Roy, M. Barlaud, E. Debreuve, and G. Aubert. Vector field segmentation using active contours : regions of vectors with the same direction. In *IEEE Int. Workshop on Variational, Geometric and Level Sets Methods in Computer Vision*, Nice, October 2003.

[11] C. Schnörr. Segmentation of visual motion by minimizing convex non-quadratic functionals. In *Proc. 12th Int. Conf. Pattern Recognition*, volume A, pages 661–663, Jerusalem, October 1994. IEEE Computer Society Press, Los Alamitos.

[12] J. Weickert. On discontinuity-preserving optic flow. In *Proc. Computer Vision and Mobile Robotics Workshop, Santorini*, pages 115–122, September 1998.

Extraction de la forme et de la perspective dans des textures artificielles et naturelles par modèles corticaux

Corentin MASSOT, jeanny HÉRAULT

Laboratoire des Images et des Signaux
46, Avenue Felix Viallet, 38031 Grenoble Cedex, France
corentin.massot@lis.inpg.fr, jeanny.herault@lis.inpg.fr

Résumé – Comment le système visuel extrait-il une information de perspective à partir de surfaces texturées? Afin de répondre à cette question, nous proposons un algorithme biologiquement plausible basé sur un modèle simplifié du système visuel. Tout d'abord de nouveaux filtres log-normaux sont présentés en remplacement des filtres de Gabor classiquement utilisés. Notamment ces filtres sont à variables séparables en fréquence et en orientation ce qui permet d'obtenir une méthode robuste d'estimation de la fréquence locale moyenne dans l'image. Basée sur cette nouvelle approche, une décomposition de l'image en imagettes, après un pré-traitement de type rétinien, permet d'estimer la variation de fréquence sur l'ensemble de l'image. La relation entre la variation de fréquence locale et les paramètres géométriques de la surface vue en projection perspective est calculée, ce qui permet d'obtenir l'orientation et la forme 3D de l'image originale. La robustesse de la méthode est testée et discutée sur différents types de textures, à la fois régulières et irrégulières, de même que sur des scènes naturelles.

Abstract – How does the visual cortex extract perspective information from textured surfaces? To answer this question, we propose a biologically plausible algorithm based on a simplified model of the visual processing. First, new log-normal filters are presented in replacement of the classical Gabor filters. Particularly, these filters are separable in frequency and orientation and this characteristic is used to derive a robust method to estimate the local mean frequency in the image. Based on this new approach, a local decomposition of the image into patches, after a retinal pre-treatment, leads to the estimation of the local frequency variation all over the surface. The analytical relation between the local frequency and the geometrical parameters of the surface, under perspective projection, is derived and finally allows to solve the so-called problem of recovering the shape from the texture. The accuracy of the method is evaluated and discussed on different kind of textures, both regular and irregular, and also on natural scenes.

1 Extraction de la perspective dans les scènes naturelles

Comment le système visuel obtient-il une information de perspective dans les images de scènes naturelles? Basé sur les connaissances actuelles du fonctionnement du système visuel primaire, nous proposons un algorithme permettant d'obtenir une information sur l'orientation et la forme d'une surface sur laquelle est plaquée une texture, c'est-à-dire un motif plus ou moins régulier possédant différentes composantes de fréquence et d'orientation (figure 1).

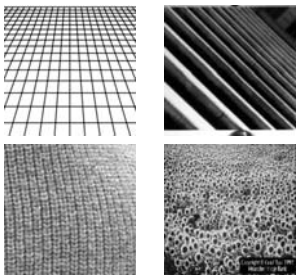


FIG. 1: Exemples de textures étudiées: quadrillage régulier, texture tirée de Super *et al* ([1]), mailles d'un pull tirée de Clerc *et al* ([2]), champ de tournesols.

Depuis le début des années 90, l'utilisation de l'information spectrale a conduit à plusieurs algorithmes efficaces d'extraction de la forme par la texture. Parmis ceux-ci on pourra notamment citer [3], [4] qui permettent de traiter des textures régulières ou possédant au moins deux composantes d'orientation et [1], [2], [5] qui ne font aucune hypothèse sur la forme du spectre et qui obtiennent également des résultats sur des textures irrégulières. Ce type de textures est plus difficile à traiter compte tenu des fortes non-stationnarités présentes dans l'image (par exemple dans un champ de tournesols, le changement de taille des fleurs ou la présence d'une personne créent des non-stationnarités locales).

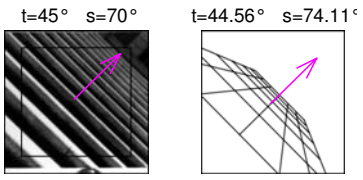


FIG. 2: Exemple de résultat obtenu sur la textures de Brodatz de la figure 1; à gauche sont indiqués les angles réels, sur l'image est reportée la normale estimée au plan; à droite sont indiqués les angles estimés et l'image représente un quadrillage incliné équivalent.

Notre travail s'inscrit dans le cadre d'une étude du fonctionnement du système visuel biologique. Les principaux mécanismes du système visuel primaire sont maintenant bien identifiés [6]. La rétine effectue un ensemble de pré-traitements

permettant notamment de séparer l'information d'illumination de celle de texture [7]. Chaque région de l'espace est analysée à travers un ensemble de champs récepteurs en recouvrement les uns par rapport aux autres. Le signal est ensuite transmis à l'aire visuelle primaire V1, où à chaque champ récepteur est associé un ensemble de cellules simples et complexes qui effectuent un échantillonnage du signal en fréquences et en orientations. Ces cellules sont organisées en macro-colonnes de même orientation et de même fréquence et la réponse des cellules complexes peut se modéliser par un filtrage passe-bande de type filtre de Gabor du spectre d'énergie correspondant à la région de l'espace visuel analysée.

La projection d'une image du monde réel sur une surface induit des déformations affines (des gradients) de la texture qui recouvre la surface. Or en vision monoculaire statique, le système visuel est capable de retrouver les caractéristiques tridimensionnelles d'une image. Dans ce papier, une nouvelle technique permettant de retrouver la forme par l'analyse des gradients de texture, basée sur le modèle simplifié, précédemment décrit du système visuel, est présentée. Elle repose sur la décomposition locale d'une texture homogène en imagettes locales (de manière similaire à l'organisation des macro-colonnes de V1). La section 2 presente un nouveau type de filtres log-normaux, mieux adaptés que les filtres de Gabor pour simuler les filtres corticaux. Ils sont utilisés pour réaliser l'échantillonnage du spectre. Ces filtres permettent de développer une technique pour estimer la fréquence locale qui tire avantage de l'ensemble des échelles disponibles après filtrage. La section 3 présente le modèle de projection perspective utlisé et la relation entre la variation de fréquence locale et les paramètres géométriques de la surface. La section 4 décrit l'algorithme final ainsi obtenu et montre différents résultats sur des textures à la fois régulières et irrégulières, de même que sur des scènes naturelles.

2 Filtres log-normaux et estimation de la fréquence locale

Les filtres de Gabor sont habituellement utilisés pour modéliser les cellules corticales. Ils présentent l'avantage d'être localisés à la fois dans l'espace et en fréquence mais ils ne sont pas à variables séparables en fréquence et en orientation. De plus ils ne sont pas symétriques exprmiés sur une échelles log-polaire et ont une transmission non nulle en $f = 0$ ce qui donne des réponses bruités en basse fréquence ([8]). Nous présentons ici de nouveaux filtres *log-normaux* qui sont mieux adaptés à notre problème et qui représentent une meilleur approximation de la réponse des cellules corticales.

2.1 Filtres log-normaux

L'expression de ces nouveaux filtres repose sur la loi log-normale:

$$|G_{i,j}(f,\theta)|^2 = A\frac{1}{f^2}exp\left(-\frac{1}{2}\left(\frac{ln(f/f_i)}{\sigma_r}\right)^2\right).cos^{2n}(\frac{\theta-\theta_j}{2}) = |G_r.G_\theta|^2$$

où $G_{i,j}$ est la fonction de transfert du filtre, f_i, la fréquence centrale, θ_j, l'orientation centrale , σ_r, la bande passante suivant les fréquences, n contrôle la bande passante suivant les orientations. A est un facteur de normalisation tel que $||G_{i,j}(f,\theta)||^2 = 1$ soit en prenant $A = \frac{2^{2n}}{2\pi C_{2n}^n}\frac{1}{\sigma_r\sqrt{2\pi}}$. Enfin G_r et G_θ représentent respectivement la composante radiale et la composante angulaire du filtre; ces filtres sont donc à variable séparables en fréquence et en orientation.

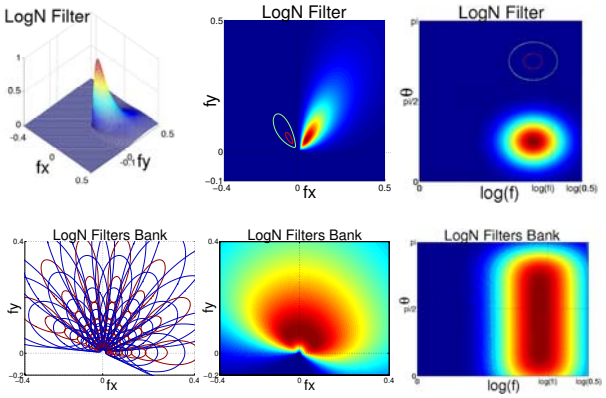


FIG. 3: *1ère ligne: filtres log-normaux, à gauche: représentation 3D; au milieu et à droite: filtre et contours représentant 50% et 90% de l'énergie maximale en coordonnées cartésiennes dans le plan des fréquences et en coordonnées log-polaires. 2ème ligne: banc de filtres log-normaux, à gauche: contours des filtres composant le banc; au milieu et à droite: banc de filtres en coordonnées cartésiennes dans le plan des fréquence et en coordonnées log-polaires.*

La première ligne de la figure 3 présente les filtres log-normaux en coordonnées cartésiennes et log-polaires. Contrairement aux filtres de Gabor, les filtres log-normaux sont insensibles à la composante continue quelle que soit la largeur de bande choisie. Ces filtres sont assymétriques en coordonnées cartésiennes mais deviennent symétriques en coordonnées log-polaires, de manière similaire à la forme des réponses des cellules complexes du cortex visuel primaire ([6]). La deuxième ligne de la figure 3 présente un banc de filtres log-normaux. Il est possible d'observer la bonne couverture du spectre et la régularité de l'échantillonnage, notamment en log-polaire.

2.2 Estimation de la fréquence locale moyenne

Dans cette partie nous présentonds une méthode permettant d'estimer la fréquence moyenne à une position donnée de l'image. Prenons la composante fréquentielle d'un filtre log-normal à la fréquence centrale f_i:

$$G_i^2(f) = \frac{1}{f^2}exp\left(-\frac{1}{2}\left(\frac{ln(f/f_i)}{\sigma_r}\right)^2\right)$$

De manière similaire à Knutsson *et al* [9], le rapport de filtres adjacents en f_i et f_{i+1} s'exprime par:

$$\frac{G_{i+1}^2(f)}{G_i^2(f)} = exp\left(-\frac{1}{2\sigma_r^2}[(ln(f/f_{i+1})^2 - (ln(f/f_i))^2]\right) = \left(f/\sqrt{f_i f_{i+1}}\right)^{\frac{ln(f_{i+1}/f_i)}{\sigma_r^2}}$$

En posant $\sigma_r^2 = ln(f_{i+1}/f_i)$, nous obtenons finalement:

$$G_{i+1}^2(f) = \frac{f}{\sqrt{f_i f_{i+1}}} G_i^2(f)$$

Afin d'extraire une information d'échelle indépendante des orientations locales, nous considérons des filtres par bande de fréquence (notés FBF) qui résultent de la somme sur l'ensemble des orientations j des réponses des filtres centrés à la même fréquence f_i sur le spectre d'énergie de l'image $S(f,\theta)$:

$$C_i = \int_f G_i^2(f) \int_\theta S(f,\theta) \sum_j G_j^2(\theta) f df d\theta \quad (2) \qquad \text{et}$$

$$\frac{C_{i+1}}{C_i} = \frac{1}{\sqrt{f_i f_{i+1}}} < f > \quad (3)$$

Le rapport de deux FBF adjacents donne l'équation 3 et représente une estimation de la fréquence locale moyenne pour la bande située à la $i$ème fréquence centrale du banc de filtres utilisé. Finalement en sommant sur l'ensemble des estimations, nous obtenons:

$$< f >= \sum_i \frac{C_i}{\sum_i C_i} < f >_i \quad (4)$$

Il est à noter que cette méthode tire partie des différentes estimations effectuées sur l'ensemble des bandes de fréquence disponibles, ce qui la rend robuste aux variations d'échelle entre les différentes textures étudiées. De plus cette méthode repose sur la séparation entre l'information d'échelle et l'information angulaire, ce qui permet d'utiliser uniquement l'information de fréquence sans hypothèse ou perturbation aucune provenant de l'information d'orientation contenue dans la texture analysée. Cette méthode s'avère simple et efficace car elle ne repose que sur la combinaison au premier ordre de différents filtres et s'adapte bien à un modèle *feedforward* du traitement visuel de bas niveau ce qui n'est pas le cas des méthodes similaires, telles que [1] ou [5].

3 Géométrie et estimation de l'orientation de surfaces planes

3.1 Projection perspective

Nous considérons un modèle de projection perspective pour le passage du monde tridimensionnel au monde bidimensionnel de l'image. La figure 4 presente le système de coordonnées de la projection d'une surface plane. (x_w, y_w, z_w) represente les coordonnées du monde, (x_s, y_s), les coordonnées de la surface et (x_i, y_i), les coordonnées de l'image. d (resp. $zw0$) représente la coordonnée de l'image (resp. de la surface) sur l'axe z_w. τ (*tilt*) represente l'angle entre x_i et la projection de la normale z_s à la surface sur le plan de l'image. σ (*slant*) est l'angle entre z_w et la normale à la surface en $zw0$. Sa valeur est comprise entre 0 et $\pi/2$.

La relation entre les coordonnées (x_s,y_s) de la surface et les coordonnées (x_i,y_i) de l'image s'exprime par (voir également [1]):

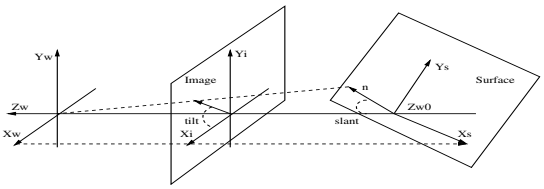


FIG. 4: *Modèle de la projection perspective*

$$\begin{bmatrix} x_s \\ y_s \end{bmatrix} = \frac{\begin{bmatrix} (cos(\sigma) & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} (cos(\tau) & sin(\tau) \\ -sin(\tau) & cos(\tau) \end{bmatrix}}{a_i} \begin{bmatrix} x_i \\ y_i \end{bmatrix} = \frac{A}{a_i} \begin{bmatrix} x_i \\ y_i \end{bmatrix} \quad (5)$$

avec $a_i = \frac{-sin(\sigma)sin(\tau)x_i + cos(\tau)sin(\sigma)y_i + dcos(\sigma)}{d + zw0}$ correspondant à un facteur de zoom en fonction de la position spatiale (x_i, y_i), noté x_i (de même on notera f_i le vecteur (f_{ix}, f_{iy})). Autour de l'axe optique, le spectre d'énergie de l'image $I_i(f_i)$ est relié à celui de la surface $I_s(f_s)$ par:

$$I_i(f_i) = \int_{x_i} i_s(x_s) e^{(-j2\pi x_i^t f_i)} dx_i$$

$$= \int_{f_s} I_s(f_s) \int_{x_i} e^{j2\pi x_i^t (\frac{A^t}{a_i} f_s - f_i)} dx_i df_s$$

En supposant la luminosité constante avec la projection, on obtient $i_i(x_i) = i_s(x_s)$ avec $x_s = \frac{A}{a_i} x_i$. En prenant a_i constant dans une région limitée centrée autour de x_i, le changement de variable $f_s \rightarrow a_i A^{-t} f_i$ donne:

$$I_i(f_i, x_i) = \int_{f_s} I_s(f_s) \delta(\frac{A^t}{a_i} f_s - f_i) df_s \quad \text{d'où} \quad \boxed{f_i = \frac{A^t}{a_i} f_s} \quad (6)$$

L'équation 6 donne la relation entre la fréquence de l'image f_i et la fréquence de la surface f_s.

3.2 Estimation de l'orientation de surfaces planes

Afin de relier la variation de fréquence avec l'orientation de la surface couverte par une texture homogène, la méthode ne repose que sur une une hypothèse de stationnarité locale. La variation de fréquence sur l'image ets alors considérée comme ne provenant que de l'inclinaison de la surface en profondeur. A partir de l'équation 6, la variation locale de fréquence de l'image f_i s'exprime par:

$$df_i = -\frac{\nabla^t a_i dx}{a_i^2} A^t f_s = -\frac{\nabla^t a_i dx}{a_i} f_i \quad (7)$$

En coordonnées polaires, la fréquence de l'image peut s'exprimer par $f_i = v_i[cos(\varphi_i)\ sin(\varphi_i)]^t$. L'équation 7 donne:

$$df_i = dv_i[cos(\varphi_i)\ sin(\varphi_i)]^t + v_i[-sin(\varphi_i)\ cos(\varphi_i)]^t d\varphi_i$$

$$= -\frac{\nabla^t a_i dx}{a_i} v_i[cos(\varphi_i)\ sin(\varphi_i)]^t \quad (8)$$

Si on considère le gradient df_i suivant le direction de φ_i, en pultipliant par $[cos(\varphi_i)\ sin(\varphi_i)]^t$ l'équation 8 devient:

$$\frac{dv_i}{v_i} = -\frac{\nabla^t a_i dx}{a_i} \quad \text{d'où}$$

$$dln(v_i) = \frac{-sin(\sigma)}{a_i}[-sin(\tau)\ cos(\tau)][dx_i dy_i]^t$$

Nous obtenons finalement que pour une surface homogène et localement stationnaire, l'angle de tilt correspond à la direction du gradient de fréquence et le slant est proportionnel à la norme du gradient du logarithme des fréquences avec la relation:

$$tan(\sigma) = \left(\frac{d}{d+zw0}\frac{1}{|dln(v_i)|} - ([-sin(\tau)\ cos(\tau)][x_i y_i]^t/d)\right)^{-1} \quad (9)$$

4 Algorithme final et résultats

L'algorithme final peut alors se résumer par:

1. Préfiltrage rétinien de l'image et décomposition en imagettes (par exemple: de taille $96X96$ pixels avec un recouvrement de 8 pixels).
2. Extraction des coefficients des filtres log-normaux (par exemple: $7X7$ filtres), normalisation par bande d'orientation et sommation sur l'ensemble des orientations afin d'obtenir les FBF.
3. Combinaison des réponses avec l'équation 2 et combinaison finale avec l'équation 4 afin d'obtenir une estimation de la fréquence moyenne locale sur l'ensemble de l'image.
4. Calcul des gradients locaux de fréquence, extraction du tilt et du slant et moyennage final.

La figure 5 présente les résultats obtenus sur différentes textures et sur des scènes naturelles. Nous obtenons une précision de l'ordre de 3^o pour le tilt et de 5^o pour le slant ce qui est comparable à la précision obtenue par d'autres auteurs tels que [1] et [5]. La méthode donne également des résultats fiables sur des textures irrégulières.

5 Conclusions

Dans ce papier nous avons présenté un nouveau type de filtres log-normaux, à variables séparables, permettant d'obtenir une méthode simple, efficace et robuste d'estimation de la fréquence moyenne locale dans une image. Un algorithme complet d'analyse de la forme par la texture a ainsi pu être présenté et évaluer sur différents types de textures à la fois régulières et irrégulières, de même que sur des scènes naturelles. La précision obtenue est comparable à des méthodes proches mais plus complexes et s'accorde bien avec un traitement bas niveau du système visuel. La précision de cette méthode pourra être notamment améliorée par des méthodes de régularisation robustes.

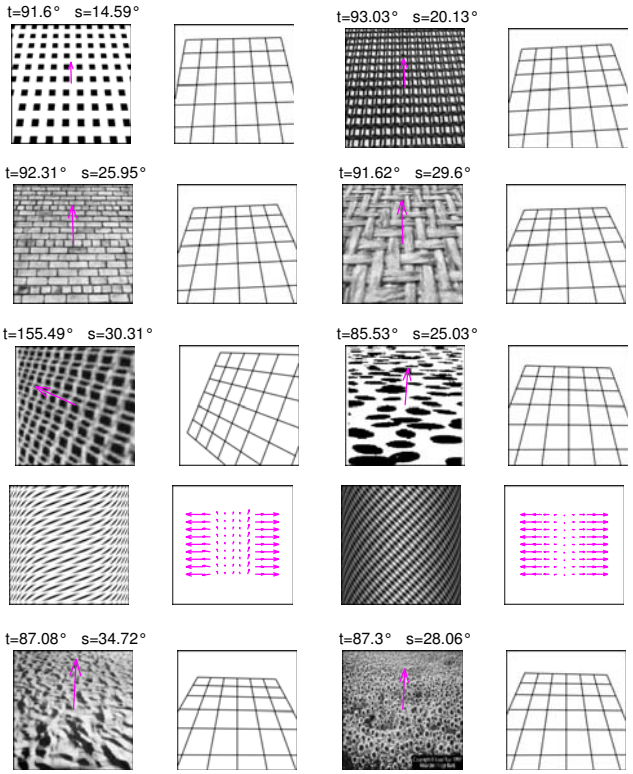


FIG. 5: *Résultats obtenus sur des textures régulières, irrégulières de même que sur des scènes naturelles.*

Références

[1] B.J. Super et A.C. Bovik, *Planar surface orientation from texture spatial frequencies*, Pattern Recognition 28, n°5, pp. 729–743, 1995.

[2] M. Clerc et S. Mallat, *The texture gradient equation for recovering shape from texture*, IEEE Trans. PAMI 24, n°4, 2002.

[3] J. Malik et R. Rosenholtz, *Computing local surface orientation and shape from texture for curved surfaces*, International Journal of Computer Vision 23, n°2, pp. 149–168, 1997.

[4] E. Ribeiro et E.R. Hancock, *Shape from periodic texture using the eigen vectors of local affine distortion*, IEEE Trans. PAMI 23, n°12, 2001.

[5] W.L. Hwang, C.S. Lu et P.C. Chung, *Shape from texture estimation of planar surface orientation throught the ridge surfaces of continuous wavelets transform*, Trans. on Image Processing 7,n°5, 1998.

[6] L. Spillmann et J.S. Werner, *Visual Perception: The Neurophysiological Foundations*, Academic Press, Inc., 1990.

[7] W.H. Beaudot, *The neural information in the vertebra retina : a melting pot of ideas for artificial vision*, PhD thesis, tirf laboratory, Grenoble, France, 1994.

[8] G. Wallis, *Linear models of simple cells: Correspondence to real cell responses and space spanning properties*, Spatial Vision 14, n°3,4, pp. 237–260, 2001.

[9] H. Knutsson, C.F. Westin et G. Granlund, *Local multiscale frequency and bandwidth estimation*, IEEE International Conference on Image Processing (ICIP'94), Austin, Texas, 1994.

Recalage semi-rigide d'images de scanner aéroporté

Claude CARIOU, Kacem CHEHDI

Institut d'Électronique et de Télécommunications de Rennes – UMR CNRS 6164
Équipe Traitement des Signaux et Images Multicomposantes et Multimodales – Université de Rennes 1
6, rue de Kerampont, BP 80518, 22305 Lannion Cedex, France
claude.cariou@univ-rennes1.fr, kacem.chehdi@univ-rennes1.fr

Résumé – Le problème traité est celui du recalage automatique d'images issues d'un scanner multispectral aéroporté (CASI) à l'aide d'ortho-images de référence issues de photographies aériennes. Nous développons le modèle utilisé, qui consiste en l'adjonction, à un modèle de déformation classique RST (*rotation, scale, translation*), d'un processus de décalage relatif des lignes successives d'acquisition du scanner, afin de compenser le mouvement de roulis important de l'avion porteur. L'estimation des paramètres du modèle est menée par une approche variationnelle, par maximisation de l'information mutuelle entre les images source et cible. Nous montrons que l'approche proposée est relativement robuste et utilisable en routine pour le mosaïquage et le géoréférencement de données d'imagerie aérienne.

Abstract – We address the problem of automatic registration of images provided by an airborne multispectral scanner (CASI) onto reference ortho-images obtained by means of aerial photographs. Into the image spatial distortion model, we introduce a scan line shift process together with a classical RST (*rotation, scale, translation*) process, in order to compensate for the prominent roll motion of the aircraft. The computation of the optimal set of model parameters is performed via a variational approach by maximizing the mutual information shared by the source and target images. We show that this approach is rather robust and can be used in a routine fashion for the mosaicking and georeferencing of aerial image data.

1 Introduction

Le recalage automatique est devenu en quelques années un outil indispensable de valorisation de l'information dans les systèmes de traitement d'images multi-modales, particulièrement dans les domaines de l'imagerie médicale et de l'imagerie aérienne et spatiale. Si les premières techniques de recalage développées mettaient en œuvre des opérations d'intercorrélation 2-D simples (en cartésien pour la translation et/ou log-polaire pour la rotation et l'homothétie) entre images issues d'un même capteur ou bien de capteurs similaires, ces techniques assez robustes sont par contre de peu d'intérêt dès lors que l'on souhaite (*i*) recaler automatiquement deux images issues de sources sensiblement différentes (capteurs visible et infrarouge par exemple), et (*ii*) prendre en compte des déformations géométriques complexes qui ne se réduisent pas à des transformations rigides (rotation, homothétie, translation).

Depuis quelques années, des réponses ont été apportées à ces deux problèmes, et parfois de façon conjointe. La théorie de l'information de Shannon a pu répondre au problème de la non homogénéité des sources d'information [4], en adoptant l'information mutuelle entre deux images de natures différentes comme fonction de comparaison. D'autre part, le développement important des techniques variationnelles pour l'analyse et le traitement des images a permis la modélisation de déformations locales par des champs de vecteurs régularisés tenant plus ou moins compte des conditions inhérentes à la physique des objets de la scène et à la géométrie de l'acquisition [1] [3].

Dans cette communication, nous proposons un outil de recalage automatique d'images issues d'un scanner aéroporté. Un tel dispositif (comme CASI ou AVIRIS) permet d'acquérir une même scène dans plusieurs bandes spectrales, grâce à l'acquisition simultanée et répétitive du spectre de chaque site spatial dans la direction normale au déplacement de l'avion. Si l'une des dimensions spatiales de l'acquisition correspond à un balayage (*scan*) supposé instantané, l'autre dimension spatiale est soumise aux mouvements de l'avion, parmi lesquels, en dehors de la translation, le mouvement de roulis est généralement prépondérant (voir figure 1). Le recalage de telles images nécessite donc l'introduction d'un modèle de décalage des lignes successives, ainsi que la prise en compte de la multimodalité des acquisitions.

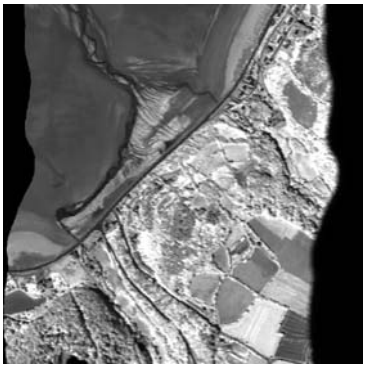


FIG. 1: *Image CASI après correction de roulis.*

Nous présenterons tout d'abord le modèle de déformation adopté, qui utilise la composition d'une transformation rigide RST avec un processus de décalage des lignes successives. Ensuite, nous présenterons le critère d'optimisation des paramètres de la transformation inverse, et sa mise en œuvre par approche

Avec la participation de l'Union Européenne – Ces travaux sont cofinancés par le FEDER et le Conseil Régional de Bretagne au travers du projet Interreg3B PIMHAI.

intégro-différentielle. Nous présenterons finalement quelques résultats de recalage sur des images synthétiques et réelles avant de conclure.

2 Modèle de déformation

Un recalage des images brutes obtenues (non corrigées des variations d'attitude de l'avion porteur) sur une image de référence nécessite une modélisation aussi précise que possible des processus de déformation mis en jeu. Dans le cas présent, si l'on suppose dans un premier temps un mouvement de translation uniforme du vecteur, la première source de déformation de l'image acquise est constituée d'une transformation rigide de type RST : celle-ci est superposable à l'image de référence à une similitude positive près. Dans un second temps, les mouvements d'attitude de l'avion (tangage, roulis, lacet) peuvent être pris en compte. Toutefois, après analyse d'un grand nombre d'images issues du capteur CASI, et afin d'éviter une paramétrisation trop complexe du problème, nous avons considéré que le mouvement de roulis est principalement responsable de ces déformations, affectant d'un décalage relatif les lignes successives du balayage.

Plus précisément, nous avons fait le choix du modèle de déformation suivant :

$$D(\alpha, \rho, \mathbf{t}, \mathbf{l}) = L(\mathbf{l}) \circ T(\mathbf{t}) \circ S(\rho) \circ R(\alpha) \quad , \tag{1}$$

où R, S et T sont respectivement les transformations de rotation d'angle α, d'homothétie de facteur ρ (toutes deux centrées sur l'image par hypothèse) et de translation de vecteur $\mathbf{t}$, et L représente l'application d'un ensemble de décalages relatifs $\mathbf{l} = [l_0 \dots l_{i_{\max}}]^T$ entre les lignes du processus de balayage (voir figure 2). La solution au problème de recalage consiste donc à optimiser le paramètre $\Theta = [\alpha, \rho, \mathbf{t}, \mathbf{l}^T]^T$:

$$\Theta^\star = \arg\min_{\Theta} C(I_r, I_d \circ D_\Theta^{-1}) \quad , \tag{2}$$

où $I_r(\mathbf{s}), \mathbf{s} = (i, j) \in \Omega = \Omega_i \times \Omega_j$ représente l'image de référence (la cible), $I_d(\mathbf{s})$ l'image déformée (la source à rectifier), et $C(.,.)$ représente un critère de comparaison entre deux images, c'est-à-dire une fonction de coût de non alignement.

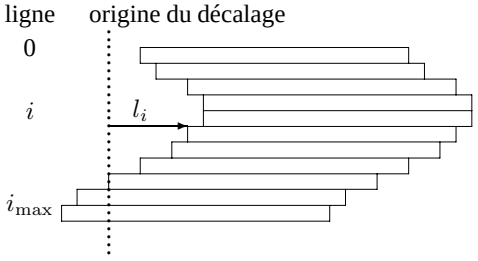


FIG. 2: *Modèle de décalage des lignes dû au roulis.*

3 Optimisation par approche variationnelle

Dans le cas d'un même imageur identique pour la source et la cible, le choix du critère SSD (*Sum of Squared Differences*) [2] est le plus naturel. Toutefois, dans le cas de systèmes imageurs différents, ou d'acquisitions effectuées à des dates éloignées, ou encore de la diversité des images cibles (autres images aériennes, satellitaires, cartes numériques, etc.) potentiellement utilisables pour le recalage, un choix maintenant classique consiste à utiliser le critère d'information mutuelle. Plus précisément, nous avons considéré le problème de la minimisation du critère suivant:

$$\mathcal{J}_{\mathcal{IM}}(\Theta) = -\mathcal{IM}(I_r, I_d \circ D_\Theta^{-1}) \quad , \tag{3}$$

où l'information mutuelle entre les images déformée et référence est définie par:

$$\mathcal{IM}(I_r, I_d \circ D_\Theta^{-1}) = \int_{\mathbb{R}^2} f_{I_r, I_d \circ D_\Theta^{-1}}(i_1, i_2) \log \frac{f_{I_r, I_d \circ D_\Theta^{-1}}(i_1, i_2)}{f_{I_r}(i_1) f_{I_d \circ D_\Theta^{-1}}(i_2)} di_1 di_2 \ , \tag{4}$$

où $f_{I_r, I_d \circ D_\Theta^{-1}}$ représente la f.d.p. conjointe des images référence et rectifiée, f_{I_r} la f.d.p. de l'image référence et $f_{I_d \circ D_\Theta^{-1}}$ la f.d.p. de l'image source rectifiée. Nous avons mis en œuvre l'optimisation conjointe des paramètres RST et du décalage des lignes $\mathbf{l}$ en approchant la solution de l'équation d'Euler-Lagrange à l'aide d'une descente de gradient. Les paramètres RST n'ont fait l'objet d'aucune régularisation particulière de la solution, tandis que la solution sur $\mathbf{l}$ a été régularisée par un filtrage passe-bas spécifié et effectué dans le domaine fréquentiel d'une façon analogue à celle décrite dans [1]. Plus précisément, le flot de gradient obtenu (dérivé de [2]) s'écrit:

$$\begin{cases} \alpha_t &= \int_\Omega W_\Theta(\mathbf{s}) \nabla_\alpha I_d \circ D_\Theta^{-1}(\mathbf{s}) \, d\mathbf{s} \\ \rho_t &= \int_\Omega W_\Theta(\mathbf{s}) \nabla_\rho I_d \circ D_\Theta^{-1}(\mathbf{s}) \, d\mathbf{s} \\ \mathbf{t}_t &= \int_\Omega W_\Theta(\mathbf{s}) \nabla_{\mathbf{s}} I_d \circ D_\Theta^{-1}(\mathbf{s}) \, d\mathbf{s} \\ \mathbf{l}_t &= \left[\int_{\Omega_j} W_\Theta(\mathbf{s}) \nabla_j I_d \circ D_\Theta^{-1}(\mathbf{s}) \, dj \right] \circledast \mathbf{g} \\ \Theta(t=0) &= \Theta_0 \end{cases} \tag{5}$$

où

$$W_\Theta(\mathbf{s}) = -\frac{1}{\mu(\Omega)} \left[\psi \star \frac{\partial L^\Theta}{\partial i_2} \right] (I_r(\mathbf{s}), I_d \circ D_\Theta^{-1}(\mathbf{s})), \tag{6}$$

$$\frac{\partial L^\Theta}{\partial i_2} = \frac{1}{f_{I_r, I_d \circ D_\Theta^{-1}}(i_1, i_2)} \frac{\partial f_{I_r, I_d \circ D_\Theta^{-1}}(i_1, i_2)}{\partial i_2} - \frac{1}{f_{I_d \circ D_\Theta^{-1}}(i_2)} \frac{\partial f_{I_d \circ D_\Theta^{-1}}(i_2)}{\partial i_2} \quad , \tag{7}$$

ψ représente un noyau de lissage pour l'estimation (de type Parzen-Rosenblatt) de la f.d.p. conjointe $f_{I_r, I_d \circ D_\Theta^{-1}}$, $\circledast$ est l'opérateur de convolution circulaire et le vecteur $\mathbf{g}$ contient la réponse impulsionnelle (non-causale) correspondant à un filtre passe-bas fréquentiel idéal:

$$g(i) = 1 + 2 \sum_{k=1}^{K} \cos \frac{2\pi k i}{i_{\max} + 1} \ , \ i = [0 \dots i_{\max}] \tag{8}$$

4 Résultats expérimentaux

4.1 Recalage d'images synthétiques

Afin de tester l'algorithme mis en œuvre, nous avons utilisé un jeu de deux images synthétiques, dont l'une a été obtenue

après l'application sur l'autre du modèle de déformation présenté plus haut (voir figure 4). Les paramètres de la transformation RST appliquée sont donnés dans le tableau 1. Le processus de décalage de lignes est donné par $l_i = 10\cos(2\pi i/256)$.

Nous avons estimé ces paramètres dans deux cas, en appliquant à chaque fois 500 itérations de la descente (5). Le premier cas correspond à une régularisation du processus de décalage des lignes par un filtrage passe-bas fréquentiel optimal, c'est-à-dire dont l'expression dans le domaine spectral est réduite à deux coefficients indépendants ($K = 1$ dans l'équation 8). Le second cas correspond à une surestimation de l'ordre du filtre, avec trois coefficients indépendants ($K = 2$). Les résultats obtenus sont également donnés dans le tableau 1 pour ce qui concerne les paramètres RST. Les estimations du processus de décalage des lignes sont présentées en figure 3. Ces résultats montrent que la technique proposée converge vers la solution attendue, même si l'ordre du filtre dépasse de peu l'ordre optimal. Toutefois, une surestimation trop élevée de cet ordre ($K = 4$ ou 5) conduit comme on peut s'y attendre à une mauvaise estimation des paramètres de la transformation. D'un point de vue pratique, la sélection de l'ordre optimal du filtre peut être effectuée grâce à la mesure de l'information mutuelle $\mathcal{IM}(I_r, I_d \circ D_\Theta^{-1})$, ou encore de l'entropie conjointe $\mathcal{H}(I_r, I_d \circ D_\Theta^{-1})$.

TAB. 1: Estimation des paramètres de la transformation

Paramètres RST	α	ρ	$\mathbf{t}$
original	20^o	0.9	$(0, 0)$
régul. $K = 1$	19.95^o	0.898	$(1.27, 1.00)$
régul. $K = 2$	19.57^o	0.898	$(1.25, 0.80)$

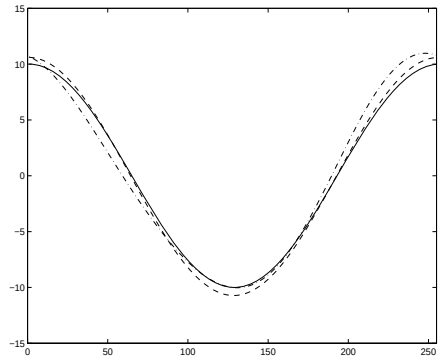


FIG. 3: *Estimation du décalage des lignes: (—) original ; (- -) estimation avec régularisation du processus de décalage pour $K = 1$; (-.-) estimation avec régularisation du processus de décalage par filtrage fréquentiel pour $K = 2$.*

4.2 Recalage d'images réelles

Nous avons testé la robustesse de cette technique dans plusieurs cas expérimentaux, en utilisant des données réelles brutes (non corrigées du roulis et non géoréférencées) issues du capteur multispectral CASI disponible dans notre laboratoire, et acquises en 1998. Ces images brutes ont été alignées sur des ortho-images acquises en 2003 et provenant de la BD-ortho de l'IGN (Institut Géographique National). Les figures 5 et 6

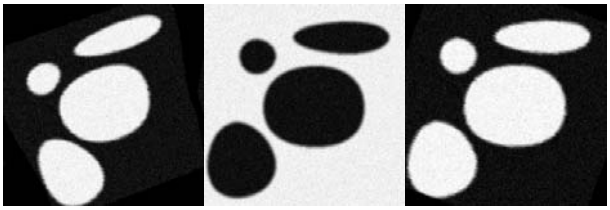


FIG. 4: *Recalage d'images synthétiques: (gauche) image source ; (centre) image cible ; (droite) recalage de l'image source pour $K = 1$.*

montrent les images source et référence, ainsi que les recalages associés. Ces recalages ont été obtenus après 300 itérations du flot de gradient, et en choisissant un filtrage fréquentiel à $K = 19$ coefficients. Les alignements obtenus sont relativement corrects, excepté dans la figure 6, où l'on peut observer dans le haut et le bas de l'image rectifiée un décalage résiduel des lignes. Ce décalage provient de la régularisation adoptée pour le processus de décalage des lignes, qui, après filtrage fréquentiel, se trouve contraint d'être une fonction périodique de la variable i, ce qui n'est pas le cas en général.

5 Conclusion

Dans cette communication, nous avons présenté une technique de recalage d'images issues d'un capteur aéroporté de type *scanner*, en utilisant comme données de référence des ortho-images. Cette technique s'appuie sur un modèle *ad-hoc*, bien que simplifié, du mouvement relatif de la ligne de balayage, et utilise le critère d'information mutuelle pour l'estimation des paramètres de la déformation, estimation mise en œuvre à l'aide d'une approche variationnelle. Nous avons montré que cette technique est relativement robuste et qu'elle est utilisable de façon automatique et avec une paramétrisation légère pour le géoréférencement et/ou le mosaïquage de données d'imagerie aérienne. De plus, elle est aisément extensible à des modèles de déformation plus complexes, prenant en compte par exemple des facteurs d'échelle différents suivant la dimension spatiale considérée (*along track* ou *across track*), ou bien encore des champs de déformation continus, ce qui pourrait permettre de corriger l'image brute des variations dues à la topographie.

Références

[1] G. Christensen, 'Consistent linear-elastic transformations for image matching', *Lecture Notes in Computer Science*, **1613**, pp. 224-237, 1999.

[2] G. Hermosillo, C. Chefd'Hotel, O. Faugeras, 'Variational methods for multimodal image matching', *Int. Journal Computer Vision*, **50**(3), pp. 329-343, 2002.

[3] J. Maintz, M. Viergever, 'A survey of medical image registration', *Medical Image Analysis*, **2**(1), pp. 1-36, 1998.

[4] P. Viola, W. Wells, 'Alignment by maximization of mutual information', *Int. Journal Computer Vision*, **24**, pp. 137-154, 1997.

FIG. 5: *Recalage d'images aériennes: (en haut) Image source CASI ; (au centre) Ortho-image cible IGN ; (en bas) Recalage de l'image source.*

FIG. 6: *Recalage d'images aériennes: (en haut) Image source CASI ; (au centre) Ortho-image cible IGN ; (en bas) Recalage de l'image source.*

Contours actifs pour le suivi d'objet en temps-réel: multi-topologies et multi-résolutions

Sébastien LEFÈVRE[1], Nicole VINCENT[2]

[1] Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection (LSIIT)
Université Louis Pasteur – Strasbourg I
Parc d'Innovation, Bd Brant, BP 10413, 67412 Illkirch Cedex

[2] Centre de Recherche en Informatique de Paris V (CRIP5)
Université René Descartes – Paris V
45 rue des Saints Pères,75270 Paris Cedex 06

lefevre@lsiit.u-strasbg.fr, nicole.vincent@math-info.univ-paris5.fr

Résumé – Dans cet article notre objectif est de présenter des solutions au suivi d'objets multiples dans une séquence d'images avec une contrainte de temps-réel et une caméra mobile. Nous proposons une solution basée sur les modèles de contours actifs, en s'affranchissant de leurs limites biens connues liées à l'initialisation, au réglage optimal des paramètres, au coût de calcul, et à l'incapacité à gérer les changements de topologie. Pour cela, l'algorithme est basé sur des énergies originales, un processus de suivi par boîte englobante, des étapes de scission et fusion, et une analyse multirésolution des images.

Abstract – In this paper our purpose is to present some solutions to multiple object tracking in an image sequence with a real time constraint and a mobile camera. We propose a solution which is built upon active contours models, without their well-known limitations related to initialisation, optimal parameter setting, computational cost, and unability to deal with topology changes. To do so, the algorithm is based on original energies, a box-based tracking process, splitting and merging steps, and multiresolution image analysis.

1 Introduction

Parmi les problèmes qu'il est nécessaire de résoudre lors de l'analyse de séquences vidéo, le suivi d'objet occupe une place prépondérante. Lorsque les objets à suivre sont de nature non-rigide (comme des corps humains), l'utilisation de modèles déformables comme les contours actifs est préconisée. La puissance d'un outil comme les *level sets* [7] n'est plus également à démontrer, mais le temps de calcul nécessaire est généralement prohibitif dans un contexte d'exécution en temps-réel. C'est pourquoi nous considérerons plutôt ici les contours actifs, de complexité plus réduite. Les contours actifs [4] ont été largement utilisés pour résoudre des problèmes tels que la segmentation d'image et le suivi des objets.

Dans cet article, nous montrons qu'il est possible d'utiliser les contours actifs pour suivre des objets en intégrant au modèle des améliorations qui permettent de supprimer leurs principales limites relatives au réglage des paramètres, à la sensibilité à l'initialisation, à l'incapacité de changer de topologie, et au temps de calcul probihibif.

2 Modèles de contours actifs

Dans la définition initale de Kass *et al* [4], un contour actif est représenté par une courbe v, fermée ou non, évoluant au cours du temps. Cette évolution s'effectue selon la minimisation d'une fonctionnelle d'énergie :

$$E(v) = \int_0^1 \left[\alpha_{\text{int}} E_{\text{int}}(v(s)) + \alpha_{\text{ext}} E_{\text{ext}}(v(s))\right] ds \qquad (1)$$

où E_{int} et E_{ext} représentent les énergies internes et externes. Les coefficients α_{int} et α_{ext} sont utilisés pour donner plus ou moins d'influence à chaque énergie. L'énergie interne représente les propriétés physiques intrinsèques du contour (contour globalement homogène et de forme simple) tandis que l'énergie externe est le terme d'attache aux données qui permettra au contour actif de se fixer sur les limites des objets suivis.

A partir de ce modèle initial, différentes implémentations ont été proposées et nous avons adopté l'algorithme glouton (ou *greedy*) [10] connu pour être le plus efficace (cf. l'étude de Denzler et Niemann [3]). Il consiste en une implémentation discrète et locale du contour. La définition de la fonction d'énergie est alors la suivante :

$$E(V) = \sum_{i=1}^{m} E(V^\gamma(i)) \qquad \text{et} \qquad V = \underset{V(i), i\in[1,m]}{\arg\min}\ E \qquad (2)$$

où V représente le contour actif discret et $V(i)$ son i$^{\text{ème}}$ point. Ce contour actif est constitué par un nombre m de points. La déformation est effectuée itérativement et localement en chaque point du contour, V^γ représentant le contour à l'itération γ.

3 Définition des énergies utilisées et description de la méthode de suivi

La méthode de suivi proposée dans cet article est basée sur différentes énergies internes et externes, et se différencie de l'existant par une initialisation par boîte englobante et non par estimation de mouvement.

L'énergie interne est constituée des énergies classiques de continuité et de courbure. Nous y intégrons également une autre énergie non liée au contenu de l'image, l'énergie ballon inversée. Cette dernière permet de réduire le contour actif plutôt que de l'étendre, comme nous le verrons par la suite.

L'énergie externe est quant à elle composée de deux énergies liées respectivement à des informations relatives au gradient et à l'arrière-plan de l'image. Pour la première, nous nous basons sur l'estimateur de Sobel, appliqué de façon marginale sur les trois composantes RGB, et dont la valeur est seuillée pour limiter la sensibilité au bruit. Cette énergie a pour but de fixer le contour actif sur les frontières de l'objet suivi. La seconde énergie permet au contour actif de rester sur les bords de l'objet, évitant ainsi une convergence à l'intérieure de la forme de l'objet suivi. Si l'arrière-plan est homogène, elle utilise des informations a priori sur ses caractéristiques couleur, et est définie en chaque point comme la différence (seuillée également pour limiter la sensibilité au bruit) entre la couleur du point et la couleur moyenne de l'arrière-plan. Dans le cas d'un fond non homogène, cette énergie est obtenue par comparaison non plus avec la couleur moyenne mais avec une image de référence, ce qui nécessite une caméra statique. Nous ne considérons pas le cas simultané d'un fond complexe et d'une caméra en mouvement pour laquelle nous n'avons aucun paramètre.

Puisque le but ici est de suivre en temps réel des objets non-rigides couleur par une caméra en mouvement, nous avons décidé afin de limiter le temps de calcul de ne pas effectuer de prétraitement (tel un filtrage) ni d'estimation du mouvement des objets ou de compensation du mouvement de la caméra. Ces choix différencient clairement notre contribution des travaux précédents de la littérature. Pour effectuer le suivi, nous considérons sur chaque image deux étapes successives : l'initalisation puis la déformation du contour actif.

L'initalisation est obtenue par la création d'un rectangle englobant le résultat fourni à l'image précédente. Ce rectangle est supposé englober également le résultat attendu sur l'image courante. La dilatation du rectangle englobant permet de prendre en compte le mouvement de la caméra et de l'objet.

La déformation amène ensuite le contour actif à se réduire pour se fixer sur les limites de l'objet. Ces deux étapes sont illustrées dans la figure 1 qui nous permet de plus de comparer notre approche originale avec le schéma traditionnel. Une initialisation spécifique est nécessaire pour la première trame d'un plan, où aucune information sur le contour actif n'est disponible *a priori*. Dans ce cas, l'initialisation est effectuée de manière automatique en utilisant le résultat d'une étape de séparation du fond et des objets [5]. La technique de suivi présentée ici ne permettant pas de gérer des objets se croisant dans la scène, nous proposons dans la section suivante un processus pour considérer des topologies multiples.

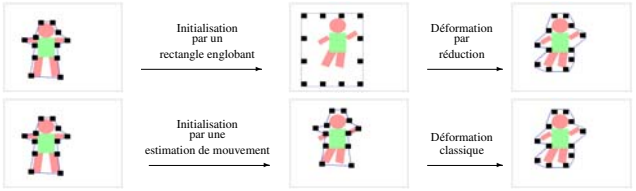


Figure 1: Suivi avec des contours actifs : le processus de suivi débute toujours par le résultat précédent (image de gauche) pour obtenir le résultat courant (image de droite). Cependant, notre contribution ne nécessite pas d'estimation de mouvement (ligne du haut) contrairement aux autres approches par contours actifs (ligne du bas).

4 Multi-topologies

Lorsque plusieurs objets d'intérêt sont présents dans la scène, il est nécessaire d'intégrer au modèle de contour actif une capacité à changer de topologie. D'après la taxonomie introduite par Cheung *et al* [1], notre approche n'est donc pas basée sur une représentation indépendante d'une topologie particulière comme les ensembles de niveaux, mais requiert plutôt l'utilisation explicite d'une technique de *division-et-fusion* [2, 6]. Nous considérons un mécanisme de scission et fusion pour respectivement séparer un contour en deux lorsque deux objets proches sont suivis par un unique contour et s'éloignent l'un de l'autre et pour regrouper deux contours en un seul lorsque deux objets éloignés se rapprochent. Pour plus d'efficacité, cette étape de scission/fusion ne sera effectuée sur chaque image qu'après convergence du ou des contours actifs.

L'objectif de l'étape de scission est de diviser le contour actif en plusieurs composants. D'après l'équation (2), l'énergie obtenue est un minimum. Puisque nous utilisons une implémentation locale et discrète, chacun des points $V(i)$ est un optimum local. Comme nous le verrons plus loin, nous souhaitons donner autant d'importance aux énergies interne et externe. Cependant, les énergies externes ne fournissent pas toujours un unique minimum (ou optimum). En effet, ces termes ont été seuillés pour accroître la robustesse au bruit. Localement, les valeurs de ces énergies peuvent être uniformément égales à 0. Une fois la convergence du contour actif obtenue, certains points du contours peuvent être piégés dans ces zones d'énergie externe constante et ne pas se fixer correctement sur les bords des objets suivis, comme l'illustre la figure 2.

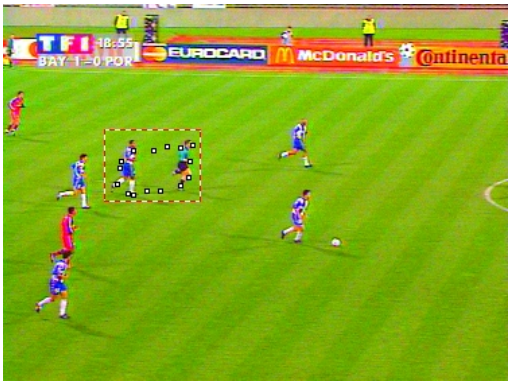


Figure 2: Répartition précise des points du *snake* avant d'effectuer l'étape de scission.

Nous proposons donc de supprimer ces points et de scinder les contours actifs aux positions de ces points incorrects. A partir de la liste des points restantes, chaque séquence de points sucessifs est utilisée pour créer un nouveau contour fermé. La figure 3 illustre ce procédé. La qualité de ces nouveaux contours n'est pas optimale, mais est suffisante pour le suivi dans les images suivantes de la séquence.

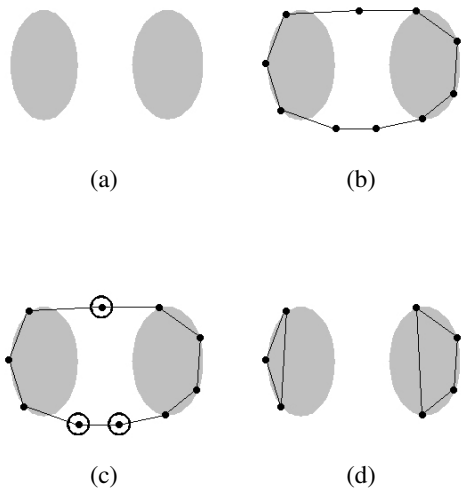


Figure 3: Processus de scission de deux objets (a), associés à un unique contour (b) ; ce contour contient des points incorrects (c), supprimés pour créer les contours finaux (d).

Il peut arriver que les contours créés par l'étape de scission ne soient pas corrects. En effet, certains de ces contours peuvent être associés à des pixels représentant le bruit ou l'arrière-plan présent dans l'image. Une étape de sélection est donc ensuite nécessaire pour valider les contours obtenus et supprimer ceux qui semblent non valides. La validation d'un composant est basée sur la comparaison de l'aire délimitée par le contour avec un intervalle prédéfini, correspondant à la taille des objets suivis.

L'étape de scission décrite précédemment nécessite la définition d'une étape de fusion correspondante. La fusion de deux contours actifs a lieu si les objets suivis par ceux-ci se rapprochent, avant un phénomène d'occlusion. L'heuristique utilisée dans ce cadre consiste à comparer les centres de gravité des contours actifs et à évaluer la proportion de chevauchement.

5 Multi-résolutions

Notre approche intègre également un aspect multirésolution afin d'accélerer le traitement nécessaire au suivi des objets. Afin d'effectuer le suivi dans un cadre multirésolution, nous traitons les images selon une approche pyramidale *coarse-to-fine*. Le processus multirésolution n'est donc pas associé au modèle de contour actif mais plutôt aux images analysées, contrairement aux approches similaires de la littérature [9]. De plus, notre processus multirésolution est stoppé dès qu'une solution correcte a été obtenue, au sens du critère de décision défini dans la section précédente. Le processus peut donc ne pas être réalisé jusqu'à la résolution originale, fournissant ainsi un résultat imprécis en contrepartie d'un temps de calcul plus faible. Cet aspect est particulièrement intéressant lorsqu'un contour obtenu à faible résolution sur une trame t est suffisant pour initaliser l'algorithme sur une trame $t+1$.

L'intégration d'un cadre multirésolution au modèle de contour actif implique naturellement la modification des paramètres de l'algorithme selon la résolution, et nous avons donc été amenés à définir des fonctions d'évolution des paramètres selon la résolution étudiée. Ainsi, la taille du rectangle englobant, le nombre de points composant le contour actif, le nombre d'itérations du processus de déformation, et les seuils utilisés peuvent varier selon la résolution considérée. Les résolutions pour lesquelles l'algorithme peut donner des résultats pertinents sont illustrées par la figure 4.

Figure 4: Représentation d'une image à différentes résolutions, depuis $r = 0$ (en haut à gauche) jusqu'à $r = 5$ (en bas à droite). Les images ont été rééchantillonnées spatialement pour observer les différences de détail.

De plus, nous pouvons remarquer que cette approche permet de bénéficier des codages compressés. En effet, dans le cas de séquences vidéo compressées à l'aide d'une Transformée en Cosinus Discrète, les coefficients DC peuvent être utilisés pour fournir une image faible résolution, de taille 8×8 fois inférieure à la taille originale.

6 Résultats et Conclusion

Nous avons effectué des tests sur des séquences vidéo couleur de matchs de football obtenues à l'aide d'une caméra en mouvement. Les images, de taille 384×284 pixels, ont été acquises à une fréquence de 15 Hz. A la résolution originale, nous avons considéré un voisinage de taille 5×5 pixels, un nombre d'itérations maximal fixé à 30, et nous décidé de supprimer les contours d'aire inférieure à 10 pixels. Le processus multirésolution fonctionne à une résolution de taille $16 \times 16 = 256$ fois inférieure à l'image originale ($r = 4$), comme l'illustre la figure 5. On peut observer le manque de précision de la forme du contour actif, qui est due à la faible résolution à laquelle l'image est analysée.

Le contour actif est également capable de changer de topologie pour s'adapter aux multiples joueurs présents dans la scène. La figure 6 illustre l'intérêt de cette capacité.

Les temps de calcul nécessaires, généralement de 35ms par trame pour un PC Pentium IV cadencé à 1,7 GHz, sont faibles par rapport à d'autres approches récentes [8]. Nous avons également testé notre approche sur des séquences de vidéosurveillance obtenues par une webcam statique avec un fond hétéro-

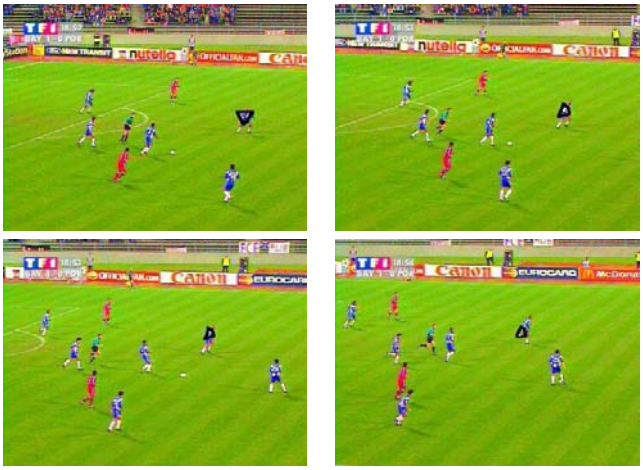


Figure 5: Suivi d'objet non-rigide à une résolution 256 fois inférieure à la résolution originale (images similaire à la 5ème image de la figure 4).

gène, ceci afin de vérifier la généricité de notre approche et de valider la définition de l'énergie d'arrière-plan basée sur une image de référence.

Dans cet article, nous proposons une solution au problème du suivi d'objet multiples dans une séquence vidéo couleur acquise par une caméra mobile. Notre approche s'affranchit des limites habituelles des contours actifs et consiste en deux étapes que sont l'initialisation par un rectangle englobant et la déformation par réduction. Elle se différencie notamment par le fait qu'elle ne nécessite ni prétraitement, ni cacul et utilisation de l'information de mouvement. De plus, pour gérer des objets multiples aux trajectoires croisées, nous avons intégré au modèle une capacité à changer de topologie par un mécanisme de scission et fusion. Nons avons également introduit un cadre multirésolution dans lequel l'analyse des images est effectuée.

Nous souhaitons maintenant intégrer à notre modèle des énergies plus robustes, puis le comparer aux techniques émergentes de *level sets* en considérant simultanément des critères de qualité et d'efficacité. Nous devons également intégrer au modèle des caractéristiques calculées sur les régions (telles que la couleur ou la texture) pour assurer une correspondance des objets après un phénomène d'occlusion.

Références

[1] K.W. Cheung, D.Y. Yeung, and R.T. Chin. On deformable models for visual pattern recognition. *Pattern Recognition*, 35(7):1507–1526, Juillet 2002.

[2] H. Delingette and J. Montagnat. Shape and topology constraints on parametric active contours. *Computer Vision and Image Understanding*, 83(2):140–171, Août 2001.

[3] J. Denzler and H. Niemann. Evaluating the performance of active contours models for real-time object tracking. In *Asian Conference on Computer Vision*, volume 2, pages 341–345, Singapour, 1995.

[4] M. Kass, A . Witkins, and D. Terzopoulos. Snakes: Active contour models. *International Journal of Computer Vision*, 1(4):321–331, Janvier 1988.

[5] S. Lefèvre, L. Mercier, V. Tiberghien, and N. Vincent. Multiresolution color image segmentation applied to background extraction in outdoor images. In *IS&T European Conference on Color in Graphics, Image and Vision*, pages 363–367, Poitiers, France, Avril 2002.

[6] T. McInerney and D. Terzopoulos. T-snakes: Topology adaptive snakes. *Medical Image Analysis*, 4(2):73–91, Juin 2000.

[7] N. Paragios and R. Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(3):266–280, Mars 2000.

[8] F. Precioso. *Contours actifs paramétriques pour la segmentation d'images et vidéos*. Thèse de doctorat, Université de Nice - Sophia Antipolis, Septembre 2004.

[9] N. Ray, B. Chanda, and J. Das. A fast and flexible multiresolution snake with a definite termination criterion. *Pattern Recognition*, 34(7):1483–1490, Juillet 2001.

[10] D.J. Williams and M. Shah. A fast algorithm for active contours and curvature estimation. *Computer Vision, Graphics and Image Processing: Image Understanding*, 55(1):14–26, Janvier 1992.

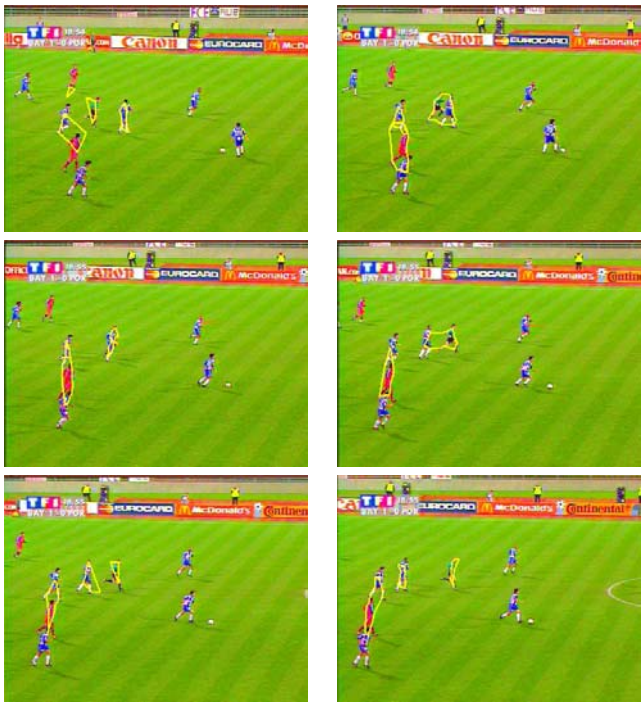


Figure 6: Intérêt des étapes de scission et fusion dans le cas d'objets proches et d'occlusions temporaires.

Influence des Vecteurs Caractéristiques en Stéganalyse par Séparateurs à Vastes Marges

Benoit ROUE[1],Patrick BAS[2], Jean-Marc CHASSERY[1]

[1]Laboratoire des Images et des Signaux INPG/ENSIEG, Grenoble, France

[2]Laboratory of Computer Science, Helsinky University of Technology, Finland

First.Name@lis.inpg.fr

Résumé − Le but de l'analyse stéganographique est de prouver la présence d'une information cachée dans un signal hôte. Cette étude se focalise sur une analyse stéganographique aveugle des images numériques utilisant des Séparateurs à Vastes Marges (SVM).
Dans un premier temps nous décrivons cette méthode de stéganalyse puis nous étudions plusieurs prédicteurs utilisés dans la phase d'extraction de caractéristiques. Enfin nous dégageons des propriétés statistiques et texturelles qui permetteront d'améliorer les performances des classificateurs à vaste marge.

Abstract − The goal of steganalysis is to search for the presence of hidden information in numerical contents.This paper focuses on blind steganalysis using Support Vectors Machine (SVM).
First we describe this algoritm, then we study several predictors used during the caracteristics extraction. Finally some attributes are outlined in order to improve steganalysis accuracy.

1 Introduction

La stéganographie consiste en l'insertion d'une information dans un support hôte sans que celle-ci ne puisse être décelable analytiquement ou visuellement. L'intérêt porté à cette discipline s'est accru depuis que l'on soupçonne son utilisation dans les documents multimédias [2] [5] [4] (images numériques marquées, fichiers audio...).
La stéganalyse (ou analyse stéganographique) a pour objectif de détecter l'éventuelle présence d'un message inséré et d'en estimer, si possible, la taille. Plusieurs familles de techniques d'analyse stéganographique ont été développées dans la littérature, mais nous nous focalisons ici sur des méthodes aveugles (*i.e.* détection de la présence d'un message sans connaître l'algorithme d'insertion) basées sur l'apprentissage par Séparateurs à Vastes Marges (SVM).

2 Méthode de stéganalyse par SVM

L'analyse stéganographique par SVM permet de séparer, à l'aide d'un classifieur, les images "propres" (images non marquées) des stego-images (images contenant un message). La construction du classifieur se fait en deux étapes :

- une phase d'extraction des caractéristiques qui permettront de discriminer les deux classes d'images
- une phase d'entraînement du classifieur par Séparateur à Vaste Marge à l'aide de ces caractéristiques.

2.1 Extraction des caractéristiques

Les caractéristiques extraites de l'image doivent faire ressortir au mieux les propriétés de la marque. Lyu *et al* [6] proposent une méthode basée sur les hautes fréquences de l'image. Dans ce travail cette information est extraite à partir d'une décomposition multirésolution de l'image (*e.g.* décomposition en ondelettes).
Cette décomposition est effectuée via des filtres miroirs en quadrature qui décomposent l'image en sous-bandes d'orientations et de fréquences différentes (Figure 1) : une sous-bande horizontale H, verticale V, diagonale D et basse fréquence L. En itérant le processus sur la sous-bande basse fréquence on obtient une décomposition multirésolution de l'image (Figure 2). Les différentes sous-bandes à l'échelle $=1$ sont notées $_i$, $_i$ et $_i$.
Une première série de caractéristiques est alors extraite de chaque sous-bande aux échelles 1 à -1 : les moments centraux normalisés d'ordre 1 à 4 (*i.e.* la *moyenne* , la *variance* , l'*asymétrie* et le *kurtosis*). Pour une sous-bande B cela donne :

$$\begin{array}{rl} = & \frac{1}{N_x N_y} \sum_{x,y} \quad (\quad) \\ = & \frac{1}{N_x N_y} \sum_{x,y} (\quad (\quad) - \quad)2 \\ = & \frac{1}{N_x N_y \sigma 3} \sum_{x,y} (\quad (\quad) - \quad)3 \\ = & \frac{1}{N_x N_y \sigma 4} \sum_{x,y} (\quad (\quad) - \quad)4 - 3 \end{array} \qquad (1)$$

où la somme s'effectue sur tous les coefficients de la sous-bande B de $_x$ pixels de largeur et de $_y$ pixels de hauteur. Cette opération permet de construire un vecteur caractéristique de $4 \times 3 \times (\quad -1)$ éléments.
Ensuite des caractéristiques tenant compte des propriétés intrinsèques des images sont calculées. En effet dans une

image naturelle les pixels ne varient pas aléatoirement, il est donc alors possible de prédire la valeur d'un pixel grâce à celle de ses voisins (dans le domaine spatial ou multi-résolution).
Cependant, lorsqu'une image a été stéganographiée la corrélation locale est perturbée. L'erreur de prédiction est donc discriminante et peut être utilisée afin de construire le vecteur caractéristique.
Pour prédire les valeurs $_i(\quad)$, $_i(\quad)$ et $_i(\quad)$ (= 1), plusieurs prédicteurs peuvent être utilisés, et la section 3 les étudiera plus en détail.
En combinant les moments issus des sous-bandes elles-mêmes et les moments provenant de l'erreur de prédiction les auteurs obtiennent donc enfin un vecteur caractéristique de $24 \times (\quad - 1)$ composantes qui décrit les statistiques de l'image analysée. C'est ce vecteur qui sera utilisé dans l'étape suivante : la classification par apprentissage.

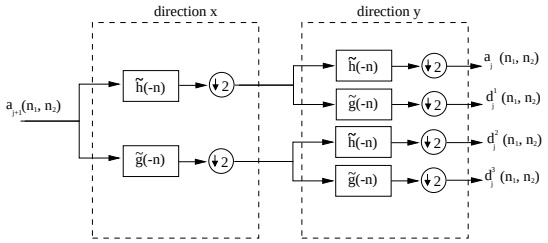


FIG. 1 – Principe de décomposition multirésolution.

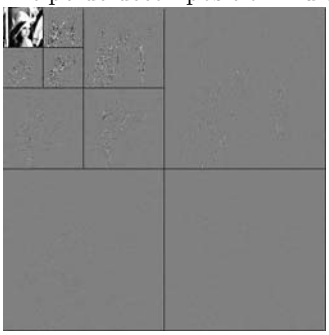


FIG. 2 – L'image Léna et sa décomposition sur 3 échelles.

2.2 Classification par apprentissage

L'étape de classification est faite par apprentissage supervisé et permet de séparer en deux classes (images propres et stégo-images) toutes les images analysées (ou plus formellement leurs vecteurs caractéristiques). Ce dernier point est le coeur de l'analyse stéganographique aveugle.
L'apprentissage supervisé requiert un entraînement du classifieur qui sera construit à l'aide d'échantillons classés au préalable. Dans la littérature, il existe plusieurs méthodes d'apprentissage supervisé (réseaux de neurones, discriminants linéaires de FISCHER, SVM linéaires ou non-linéaires) mais Farid et Lyu [6] proposent l'utilisation de Séparateurs à Vaste Marge non linéaires qui s'avèrent être, dans le cadre de l'analyse stéganographique, les plus efficaces.
Dans le cas d'une SVM linéaire nous allons chercher un hyperplan de $\mathbb{R}^d$ (de dimension $- 1$) qui séparent les échantillons propres des échantillons marqués (l'hyperplan optimal est obtenu par maximisation de lagrangien), avec, si les échantillons ne sont pas tous séparables, une fonction de coût de mauvaise classification (Figure 3).
En analyse stéganographique il est difficile de séparer les échantillons avec un hyperplan linéaire (Figure 4), il faut donc utiliser une SVM non-linéaire. Les données des vecteurs caractéristiques sont projetées dans un espace euclidien $\mathbb{H}$ de dimension supérieure à (éventuellement infinie) à l'aide d'une application injective et on applique une SVM linéaire dans cet espace. Définir serait coûteux en tant de calcul ou même théoriquement impossible, cependant, comme la résolution de l'hyperplan optimal revient à une maximisation de lagrangien, on peut résoudre le problème sans connaître . Il suffit de trouver une fonction $: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ telle que $(\overrightarrow{\quad}\ \overrightarrow{\quad}) = (\overrightarrow{\quad})\ (\overrightarrow{\quad})$. De telles fonctions sont appelées noyaux de MERCER et le noyau utilisé dans cette méthode est le *noyau gaussien* :

$$(\overrightarrow{\quad}\ \overrightarrow{\quad}) = \quad (-\frac{||\overrightarrow{\quad} - \overrightarrow{\quad}||^2}{2\ ^2}) \qquad (2)$$

où est la variance de la gaussienne, valeur à déterminer lors de l'apprentissage.

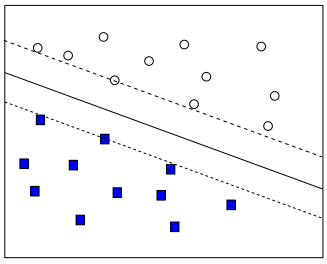


FIG. 3 – SVM linéaire en dimension 2.

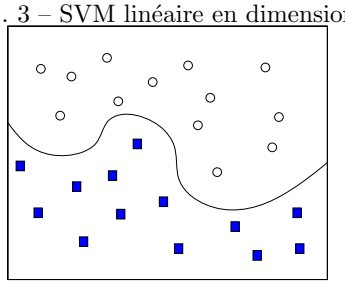


FIG. 4 – SVM non-linéaire en dimension 2

3 Etape de prédiction

Le choix des caractéristiques (Section 2.1) qui composent les vecteurs d'entrée du système d'apprentissage est une étape cruciale qui déterminera la qualité du classifieur. Le prédicteur utilisé lors de l'acquisition des statistiques doit donc être le plus efficace possible.

3.1 Prédicteur linéaire

Pour prédire les valeurs $_i(\quad)$, $_i(\quad)$ et $_i(\quad)$ (= 1), FARID *et al* [3] proposent un prédicteur linéaire [1] :

[1] * désigne la valeur prédite

$$
\begin{aligned}
{}^{*}_{i}(\quad) = {} & {}_{1}\ {}_{i}(\ -1\ ) + {}_{2}\ {}_{i}(\ +1\ ) \\
& + {}_{3}\ {}_{i}(\quad -1) + {}_{4}\ {}_{i}(\quad +1) \\
& + {}_{5}\ {}_{i+1}(\ \ 2\quad 2) + {}_{6}\ {}_{i}(\quad) \\
& + {}_{7}\ {}_{i+1}(\ \ 2\quad 2) \\
{}^{*}_{i}(\quad) = {} & {}_{1}\ {}_{i}(\ -1\ ) + {}_{2}\ {}_{i}(\ +1\ ) \\
& + {}_{3}\ {}_{i}(\quad -1) + {}_{4}\ {}_{i}(\quad +1) \\
& + {}_{5}\ {}_{i+1}(\ \ 2\quad 2) + {}_{6}\ {}_{i}(\quad) \\
& + {}_{7}\ {}_{i+1}(\ \ 2\quad 2) \\
{}^{*}_{i}(\quad) = {} & {}_{1}\ {}_{i}(\ -1\ ) + {}_{2}\ {}_{i}(\ +1\ ) \\
& + {}_{3}\ {}_{i}(\quad -1) + {}_{4}\ {}_{i}(\quad +1) \\
& + {}_{5}\ {}_{i+1}(\ \ 2\quad 2) + {}_{6}\ {}_{i}(\quad) \\
& + {}_{7}\ {}_{i+1}(\ \ 2\quad 2)
\end{aligned}
\tag{3}
$$

Les $_i$ sont les paramètres du prédicteur qu'il faut ajuster pour minimiser l'erreur de prédiction. On montre que l'erreur sur chaque sous bande est donnée, pour le prédicteur optimal, par[2] :

$$\overrightarrow{\quad} = \mathtt{log}_2(\overrightarrow{\quad}) - \mathtt{log}_2(|\ \overrightarrow{\quad}|) \tag{4}$$

où $\overrightarrow{\quad}$ contient les coefficients des sous-bandes, $\overrightarrow{\quad}$ contient les paramètres du prédicteur, et les amplitudes des coefficients voisins. A partir de $\overrightarrow{\quad}$ sont calculées sur chaque sous- bande les mêmes statistiques que dans le paragraphe 2.1 (*moyenne* , *variance* , *asymétrie* et *kurtosis*). Dans l'optique d'améliorer la détection, nous proposons d'utiliser d'autres prédicteurs.

3.2 Prédicteur DCT

L'idée est d'utiliser un prédicteur plus performant utilisé dans la norme MPEG4, qui travaille sur les coefficients DCT de l'image. Il s'agit de prédire les coefficients DCT qui ont une forte probabilité d'être modifiés par un schéma de stéganograhie : les coefficients supérieurs gauche de chaque bloc DCT (*e.g.* les 10 premiers coefficients lors d'une lecture en *zig-zag*). Ensuite les mêmes moments centraux (Section 2.1) sont extraits de l'image reconstruite et ces statistiques s'ajoutent à celles déduites de l'image originale pour former un nouveau vecteur caractéristique à utiliser lors de l'entraînement du classifieur.

3.3 Prédicteur basé B-splines

Ici on utilise des B-splines pour prédire les coefficients des sous bandes de la décomposition multirésolution c'est à dire que les coefficients à prédire sont interpolés grâce aux B-splines construites par les coefficient voisins (ligne ou colonne). Ensuite les statistiques sont extraites comme dans la méthode de FARID (paragraphe 3.1) et s'ajoutent aux caractéristiques de la section 2.1 pour former un autre vecteur caractéristique.

4 Résultats

4.1 Mise en œuvre

Les méthodes décrites dans cette étude ont été testées sur plusieurs algorithmes de stéganographie existant dans la littérature :

- Outguess, développé par N.Provos en 1999. Cet algorithme modifie les bits de poids faibles de certains coefficients DCT, puis modifie les coefficients DCT afin d'obtenir un histogramme identique à celui de l'image originale.
- Jphide, développé par A.Latham en 1998. Cet algorithme modifie les bits de poids faibles de certains coefficients DCT choisis aléatoirement.
- F5, développé par A.Westfeld et A.Pfitzmann. Comme les deux autres cet algorithme travaille sur les coefficients DCT mais décrémente leur valeur absolue et utilise une matrice de codage afin de minimiser le nombre de coefficients à modifier.

Ensuite le classifieur a été entraîné sur une base d'images en niveaux de gris (2700 images) contenant le même nombre d'images marquées et non marquées. Les images ont toutes la même taille (512×512) et les messages insérés via les algorithmes décrits sont de taille 48×48. Pendant cette phase, les paramètres et (respectivement la largeur de gaussienne et la fonction de coût pour mauvaise classification) sont déterminés par validation croisée.
Enfin une même quantité d'images marquées par les algorithmes ci-dessus et d'images propres sont utilisées pour tester les performance du classifieur SVM obtenu.

4.2 Performances

Les résultats obtenus sur des bases de 1500 images par algoritme d'insertion sont présentés sur les Figures 5 (pour le taux de bonne classifications des images propres) et 6 (pour le taux de bonne classification des images contenant un message).
La méthode de Farid a une très bonne précision en ce qui concerne les vrais positifs (*images marquées et reconnues comme telles*) et nos résultats avec le prédicteur linéaire (Figure 6) sont équivalents (seulement équivalents car la qualité d'un classifieur dépend de la base d'image utilisée lors de la phase d'entraînement, on ne peut donc raisonnablement s'attendre à avoir exactement les mêmes résultats). Les résultats obtenus avec les prédicteurs DCT et B-spline en ce qui concerne les vrais positifs sont bons aussi, voire très bons pour Outguess car les coefficient DCT sont modifiés pour restaurer l'histogramme original. On peut donc imaginer un classifieur SVM ayant des caractéristiques extraites à l'aide de ce prédicteur dans le cadre d'une stéganalyse non aveugle (où on connait l'algorithme d'insertion).
Cependant nos erreurs de classification, comme celles de Farid (Figure 5), sont très importantes pour les taux de faux positifs (30% des images non marquées sont mal classées) et surtout pour des images texturées (hautes fréquences, beaucoup de contours). Dans ce cas précis le SVM va considérer les images texturées comme marquées, en effet les erreurs de prédiction, avec tous les prédicteurs, sont importantes dans ce type d'image. Pour ces images il est donc préférable de ne pas utiliser de prédicteur.

[2] le `log` est calculé pixel par pixel (ou coefficient par coefficient).

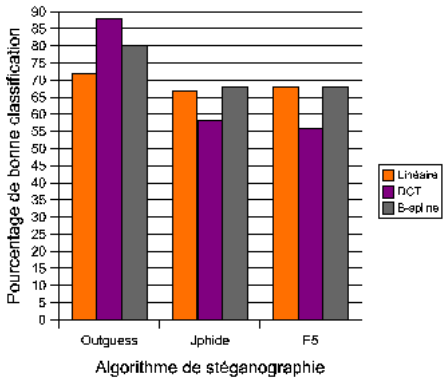


FIG. 5 – Taux de bonne classification des images non marquées.

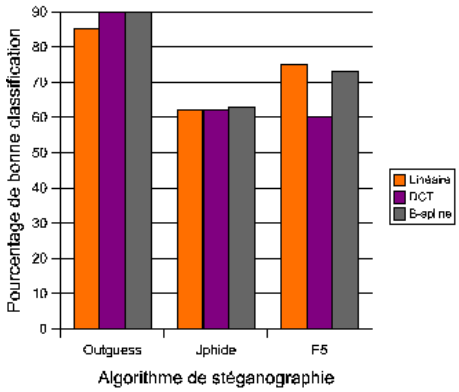


FIG. 6 – Taux de bonne classification des images marquées.

4.3 Analyse

Les performances obtenus avec les trois classifieurs sont bonnes dans l'ensemble, mais nous pouvons d'ors et déjà tirer des enseignements à partir des erreurs de prédictions obtenues. Deux causes principales des ces imprécisions se dégagent de notre analyse :

- Les erreurs sur les images texturées nous poussent à développer des solutions pour réduire l'erreur de classification par l'utilisation d'outils d'analyse de texture comme les matrices de cooccurrenceen s'inspirant de la stéganalyse LSB [1].
- Nous avons, pour entrainer les classifieurs, des vecteurs de dimension 72. c'est beaucoup trop pour une base de 2700 images. Effectivement en très grande dimension il faut un nombre quasi-infini d'échantillons pour définir un hyperplan. Des méthodes de sélection des de variable peuvent nous permettre de réduire significativement ce nombre de variables.

5 Conclusion et Perspectives

L'extraction des caractéristiques est une étape primordiale dans l'entraînement du classifieur. Nous avons voulu améliorer les performances du classifieur en tenant compte, via de nouveaux prédicteurs, des propriétés des images analysées et des algorithmes d'insertion.

Nous pouvons utiliser d'autres prédicteur optimaux (Filtre de Kalman) et il sera également important de réduire l'erreur de classification relative aux images texturées par l'utilisation d'outils de texture comme les matrices de cooccurrence, comme nous l'avons fait pour la stéganalyse LSB [1].

De plus, plutôt que d'obtenir une base d'images très grande afin d'obtenir un hyperplan séparateur plus fiable, nous pouvons réduire le nombre de coefficients des vecteurs caractéristiques, en effet ces coefficients tels qu'ils sont décrits dans cette étude sont extrémement corrélés. Une étude en cours vise à obtenir des vecteurs caractéristiques de dimension 8 en utilisant les mêmes procédés d'extractions de caractéristiques.

Références

[1] B.Roue, P.Bas, and J-M.Chassery. Improving lsb steganalysis using marginal and joint probabilistic distributions. In *Multimedia and Security Workshop*, Magdeburg, 2004.

[2] H. Farid. Detecting hidden messages using higher-order statistical models. In *International Conference on Image Processing*, NY, 2002.

[3] H.Farid and S.Lyu. Higher-order wavelet statistics and their applications to digital forensic. In *IEEE Workshop on Statistical Analysis in Computer Vision*, Wisconsin, 2003.

[4] J.Fridrich, M.Goljan, and R.Du. Detecting LSB Steganogeraphy in color and grayscale images. In *Magazine of IEEE Multimedia Special Issue on Security*, pages 22–28, October 2001.

[5] S.Dumitrescu, X.Wu, and Z.Wang. Detection of LSB steganography via sample pair analysis. In *IEEE transactions on Signal Processing*, pages 1995–2007, 2003.

[6] S.Lyu and H.Farid. Detecting hidden message using higher-order statistics and support vector machine. In 5^{th} *International Workshop on Information Hiding*, Pays-Bas, 2002.

Kernel machines : une nouvelle méthode pour l'optimisation de l'alignement des noyaux et l'amélioration des performances

Jean-Baptiste POTHIN, Cédric RICHARD

Institut des Sciences et Technologies de l'Information de Troyes (ISTIT, FRE CNRS 2732)
12 rue Marie Curie, BP 2060
10010 Troyes cedex, France
jean_baptiste.pothin@utt.fr, cedric.richard@utt.fr

Résumé – On propose ici une méthode visant à optimiser, selon le critère d'alignement maximum, une combinaison linéaire de noyaux donnés. L'algorithme proposé s'inspire de la méthode SMO dédié à l'optimisation des SVMs : il segmente le problème d'optimisation initial en sous-problèmes pour lesquels on dispose d'une solution analytique, évitant ainsi par exemple le problème de Hessien mal conditionné. Son implémentation ne nécessite pas l'utilisation d'une bibliothèque d'optimisation particulière. Enfin, la solution obtenue s'avère en pratique plus parcimonieuse que celle trouvée par l'approche de type programmation quadratique, tout en offrant des performances similaires.

Abstract – In this paper, we propose a method for optimizing the alignment of a linear combination of kernels. The proposed algorithm is based on the well-known SMO idea : it solves a quadratic optimization problem by analytically solving subproblems. Ill-conditioned Hessian can thus been avoided. Its implementation does not require any optimization toolbox. Finally, we show that solutions obtained with our strategy are sparser than those obtained by quadratic programming, for equivalent performance.

1 Introduction

Dans le foisonnant domaine des méthodes dites à *noyau*, dont les SVMs sont le fer de lance, on constate que les efforts de recherche ont été essentiellement portés sur les applications, délaissant quelque peu l'étude des noyaux eux-mêmes. Cet état de fait est d'autant plus surprenant que ces derniers conditionnent largement l'efficacité des traitements, par leur aptitude à révéler les similarités existant au sein des données traitées. L'adaptation d'un noyau à l'intelligence des données relève de trois problématiques. La première concerne la définition de critères de choix d'un noyau, par exemple une mesure d'efficacité statistique maximale. La seconde vise à circonscrire une famille de noyaux apte à répondre au problème posé, et la dernière est consacrée à la détermination du meilleur noyau de la famille exhibée en procédant par exemple à l'optimisation de ses paramètres caractéristiques. On propose ici une méthode originale visant à répondre à ces deux dernières questions. L'article est organisé comme suit. Nous commençons par préciser le contexte de l'étude avant de passer à la présentation de l'algorithme proposé. Puis des résultats expérimentaux viennent illustrer les performances de notre approche, suivis d'une conclusion et de quelques perspectives.

1.1 Contexte

Soit $\mathcal{X}$ un espace métrique compact. On appelle *noyau de Mercer* une fonction κ symétrique de $\mathcal{X} \times \mathcal{X}$ dans $\mathbb{R}$ vérifiant

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0 \qquad (1)$$

pour tout $n \in \mathbb{N}$, $\boldsymbol{x}_1, .., \boldsymbol{x}_n \in \mathcal{X}$ et $a_1, .., a_n \in \mathbb{R}$. En d'autres termes, la matrice carrée $\boldsymbol{K}$ de terme général $K_{ij} = \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j)$, appelée *matrice de Gram*, est symétrique, semi-définie positive pour tout $\{\boldsymbol{x}_i\}_{i=1}^n \subseteq \mathcal{X}$. Dans ces conditions, on montre qu'il existe une fonction $\phi : \mathcal{X} \longrightarrow \mathcal{F}$ telle que :

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle_{\mathcal{F}}, \qquad (2)$$

où $\langle \cdot, \cdot \rangle_{\mathcal{F}}$ exprime le produit scalaire dans l'espace caractéristique $\mathcal{F}$, de dimension éventuellement infinie [1]. Un noyau fournit ainsi un moyen simple d'évaluer la similitude entre les images $\phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \in \mathcal{F}$ pour toute paire $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{X}$ sans avoir à expliciter ϕ ou $\mathcal{F}$.

Supposons maintenant que l'on dispose d'un ensemble de n observations $\boldsymbol{x}_i \in \mathcal{X}$ divisées en 2 classes, la variable $y_i = \pm 1$, stockée dans le vecteur colonne $\boldsymbol{y}$, désignant la classe d'appartenance de $\boldsymbol{x}_i$. Selon l'approche SVM standard [2], l'hyperplan séparateur optimal est à marge maximale. Il est trouvé dans l'espace caractéristique $\mathcal{F}$ engendré par un noyau donné, par exemple le noyau gaussien, polynomial ou encore tangente hyperbolique :

$$\begin{array}{ll} \kappa_1(\boldsymbol{x}_i, \boldsymbol{x}_j) = & \exp \frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2} \\ \kappa_2(\boldsymbol{x}_i, \boldsymbol{x}_j) = & (\gamma \boldsymbol{x}_i^t \boldsymbol{x}_j + b)^d, \\ \kappa_3(\boldsymbol{x}_i, \boldsymbol{x}_j) = & \tanh(a\, \boldsymbol{x}_i^t \boldsymbol{x}_j + b). \end{array} \qquad (3)$$

Ici, $\boldsymbol{\Theta}_1 = \sigma$, $\boldsymbol{\Theta}_2 = (\gamma, b, d)$ et $\boldsymbol{\Theta}_3 = (a, b)$ sont appelés *hyperparamètres*. Ils sont habituellement fixés *a priori*, c'est-à-dire avant apprentissage.

1.2 Noyau adapté par critère d'alignement

L'objectif de l'article est de trouver un noyau adapté aux données dans le cadre d'une approche de type *boosting*. On rappelle à ce propos que l'idée de base du *boosting* est de combiner des solutions de moindre qualité afin d'aboutir à une solution

plus performante. C'est notamment ainsi que l'on peut obtenir un classifieur compétitif étant donné des classifieurs aux performances modestes [3]. Ce principe, reformulé dans le contexte de l'ingénierie des noyaux, vise à combiner des noyaux génériques tels que ceux cités précédemment afin d'améliorer les performances [4]. Grâce aux propriétés algébriques des fonctions définies positives, il est facile de construire de nouveaux noyaux de Mercer à partir de noyaux donnés. Par exemple :

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \alpha_1 \kappa_1(\boldsymbol{x}_i, \boldsymbol{x}_j) + \alpha_2 \kappa_2(\boldsymbol{x}_i, \boldsymbol{x}_j), \quad (4)$$

est un noyau valide pour α_1, α_2 positifs. De même, le produit de deux noyaux donne un noyau valide [5] :

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \kappa_1(\boldsymbol{x}_i, \boldsymbol{x}_j)\kappa_2(\boldsymbol{x}_i, \boldsymbol{x}_j). \quad (5)$$

Le modèle adopté pour la combinaison de noyaux est communément linéaire pour lequel il s'agit d'identifier les paramètres α_i :

$$\kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{i=1}^{M} \alpha_i \, \kappa_i(\boldsymbol{x}_i, \boldsymbol{x}_j). \quad (6)$$

En remarquant que la matrice de Gram cible est définie par $\boldsymbol{K}^* = \boldsymbol{y}\boldsymbol{y}^t$, Cristianini *et al.* proposent de quantifier la qualité *a priori* de l'espace de représentation $\mathcal{F}$ à l'aide de la mesure d'*alignement* suivante [6] :

$$0 \leq \mathcal{A}(\boldsymbol{K}, \boldsymbol{K}^*) = \frac{\langle \boldsymbol{K}, \boldsymbol{K}^* \rangle_F}{\|\boldsymbol{K}\|_F \|\boldsymbol{K}^*\|_F} \leq 1, \quad (7)$$

où $\langle \cdot, \cdot \rangle_F$ désigne le produit scalaire de Frobenius. Si d'autres critères existent, on peut citer par exemple le *class separability criterion* [7] et le *data centering criterion* [8], l'avantage de l'alignement est que son optimisation relativement au modèle linéaire (6) se ramène à un problème d'optimisation convexe [9] :

$$\max_{\boldsymbol{\alpha}} - \sum_{i,j} \alpha_i \, \alpha_j \left(\langle \boldsymbol{K}_i, \boldsymbol{K}_j \rangle_F + \lambda \, \delta_{ij} \right) + \sum_i \alpha_i \, \boldsymbol{y}^t \boldsymbol{K}_i \, \boldsymbol{y} \quad (8)$$

sous les M contraintes $\alpha_i \geq 0$ afin de garantir la semi-définie positivité de la matrice de Gram $\boldsymbol{K}$ associée au noyau (6). Notons que dans cette expression, δ_{ij} désigne le symbole de Kronecker et $\lambda > 0$ un paramètre de régularisation. En outre, on montre que l'alignement possède un biais nul et une variance faible pouvant être bornée [6]. Son lien avec l'erreur de généralisation a quant à lui été établi théoriquement et validé expérimentalement [9, 10, 11].

2 Méthode proposée

Dans le même esprit que la méthode SMO [12] dédiée à l'élaboration des SVMs, nous suggérons de segmenter le problème d'optimisation (8) en sous-problèmes pour lesquels on est en mesure d'exhiber une solution analytique.

2.1 Solution analytique dans le cas $M = 2$

D'après (8), la maximisation de l'alignement d'une combinaison linéaire de deux noyaux κ_1 et κ_2 revient à maximiser la fonction

$$\begin{aligned} &W(\alpha_1, \alpha_2) = \\ &\quad -\alpha_1^2(\|\boldsymbol{K}_1\|_F^2 + \lambda) - \alpha_2^2(\|\boldsymbol{K}_2\|_F^2 + \lambda) \\ &\quad -2\,\alpha_1\,\alpha_2 \langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle_F + \alpha_1 \, \boldsymbol{y}^t \boldsymbol{K}_1 \, \boldsymbol{y} + \alpha_2 \, \boldsymbol{y}^t \boldsymbol{K}_2 \, \boldsymbol{y} \end{aligned} \quad (9)$$

sous les contraintes $\alpha_1, \alpha_2 \geq 0$. Intéressons-nous dans un premier temps au problème sans contraintes. Le point stationnaire de (9) étant obtenu en annulant ses dérivées premières, on obtient pour conditions d'optimalité :

$$-2(\|\boldsymbol{K}_1\|_F^2 + \lambda)\alpha_1 - 2\alpha_2 \langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle_F + \boldsymbol{y}^t \boldsymbol{K}_1 \boldsymbol{y} = 0, \quad (10)$$

$$-2(\|\boldsymbol{K}_2\|_F^2 + \lambda)\alpha_2 - 2\alpha_1 \langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle_F + \boldsymbol{y}^t \boldsymbol{K}_2 \boldsymbol{y} = 0. \quad (11)$$

On aboutit ainsi au système suivant :

$$\begin{array}{l} (\|\boldsymbol{K}_1\|_F^2 + \lambda)\alpha_1 + \langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle_F \alpha_2 = \frac{\boldsymbol{y}^t \boldsymbol{K}_1 \boldsymbol{y}}{2} \\ \langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle_F \alpha_1 + (\|\boldsymbol{K}_2\|_F + \lambda)\alpha_2 = \frac{\boldsymbol{y}^t \boldsymbol{K}_2 \boldsymbol{y}}{2} \end{array} \quad (12)$$

qui admet pour solution :

$$\alpha_1^* = \frac{1}{2} \frac{\boldsymbol{y}^t \boldsymbol{K}_1 \, \boldsymbol{y} - 2\langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle_F \, \alpha_2^*}{\|\boldsymbol{K}_1\|_F^2 + \lambda} \quad (13)$$

$$\alpha_2^* = \frac{1}{2} \frac{(\|\boldsymbol{K}_1\|_F^2 + \lambda)\boldsymbol{y}^t \boldsymbol{K}_2 \, \boldsymbol{y} - \langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle_F \, \boldsymbol{y}^t \boldsymbol{K}_1 \, \boldsymbol{y}}{(\|\boldsymbol{K}_1\|_F^2 + \lambda)(\|\boldsymbol{K}_2\|_F^2 + \lambda) - \langle \boldsymbol{K}_1, \boldsymbol{K}_2 \rangle_F^2}. \quad (14)$$

D'après (13), α_1^* et α_2^* ne peuvent être nuls simultanément. De même, un raisonnement par l'absurde montre que α_1^* et α_2^* ne peuvent être strictement négatifs simultanément. Supposons pour cela $\alpha_1^*, \alpha_2^* < 0$. La combinaison linéaire des matrices $\boldsymbol{K}_1$ et $\boldsymbol{K}_2$ peut alors s'écrire $-(|\alpha_1^*|\boldsymbol{K}_1 + |\alpha_2^*|\boldsymbol{K}_2) = -\boldsymbol{K}$. En utilisant les propriétés $\mathcal{A}(-\boldsymbol{K}, \boldsymbol{K}^*) = -\mathcal{A}(\boldsymbol{K}, \boldsymbol{K}^*)$ et $0 \leq \mathcal{A}(\boldsymbol{K}, \boldsymbol{K}^*) \leq 1$, on trouve qu'une solution de la forme $\alpha_1^*, \alpha_2^* < 0$ est toujours associée à un alignement négatif. Or, $(\alpha_1 = 1, \alpha_2 = 0)$ amène à un alignement positif. On en conclut que l'hypothèse de départ $\alpha_1^*, \alpha_2^* < 0$ ne peut être vraie.

En dehors de cela, deux scénarios peuvent être observés :

- Cas 1 : $\alpha_1^*, \alpha_2^* > 0$
  La solution (13)-(14) est admissible puisqu'elle vérifie les contraintes de positivité. En outre, cette solution est toujours associée à un alignement maximum, le Hessien de (9) étant défini négatif.

- Cas 2 : $\alpha_2^* < 0$, à un échange près de α_1^* et α_2^*.
  Le couple (α_1^*, α_2^*) n'est pas admissible. Partant d'une solution réalisable du plan (α_1, α_2), on montre, par la méthode des directions réalisables [13], que l'optimum contraint est de la forme $(\alpha_1^+ > 0, \alpha_2^+ = 0)$. En d'autres termes, le noyau κ_2 est écarté. Or, l'alignement est invariant par changement d'échelle, signifiant que l'on a $\mathcal{A}(\alpha_1^+ \boldsymbol{K}_1, \cdot) = \mathcal{A}(\boldsymbol{K}_1, \cdot)$, $\forall \alpha_1^+ > 0$. On peut donc attribuer arbitrairement la valeur 1 à α_1^+, indépendamment de la constante de régularisation λ.

En résumé, la combinaison linéaire de deux noyaux est d'alignement maximum pour :

$$(\alpha_1^+, \alpha_2^+) = \begin{cases} (\alpha_1^*, \alpha_2^*) & \text{si } \alpha_1^*, \alpha_2^* > 0 \\ (1, 0) & \text{si } \alpha_2^* \leq 0 \\ (0, 1) & \text{si } \alpha_1^* \leq 0. \end{cases} \quad (15)$$

Dans la section suivante, on exploite ce résultat pour traiter le problème (8) lorsque M est supérieur à 2.

Table 1: Optimisation de l'alignement par algorithme BB

Initialisations :
$k = 0$, $I = \emptyset$, $\hat{\boldsymbol{K}} = \boldsymbol{0}$, $\mathcal{A}_{BB} = 0$, $\boldsymbol{\alpha} = \boldsymbol{0}$
$j^* = \arg\max_j \mathcal{A}(j)$, $\alpha^+_{1,j^*} = 0$, $\alpha^+_{2,j^*} = 1$.
Tant que $\mathcal{A}_{BB} < \mathcal{A}(j^*)$
$k = k + 1$, $I = I \cup \{j^*\}$, $\mathcal{A}_{BB} = \mathcal{A}(j^*)$
$\hat{\boldsymbol{K}}^{(k)} = \alpha^+_{1,j^*} \hat{\boldsymbol{K}}^{(k-1)} + \alpha^+_{2,j^*} \boldsymbol{K}_{j^*}$
$\boldsymbol{\alpha}^{(k)} = \alpha^+_{1,j^*} \boldsymbol{\alpha}^{(k-1)}$
$\alpha^{(k)}_{j^*} = \alpha^+_{2,j^*}$
Pour tout $j \notin I$
Calculer $\alpha^+_{1,j}$, $\alpha^+_{2,j}$ et $\hat{\boldsymbol{K}}^{(k+1)}(j)$
Calculer $\mathcal{A}_{BB}(j) = \mathcal{A}(\hat{\boldsymbol{K}}^{(k+1)}(j), \boldsymbol{K}^*)$
Choisir j^* tel que $j^* = \arg\max_{j \notin I} \mathcal{A}(j)$
Retourner $\boldsymbol{\alpha}$

2.2 Traitement du cas $M > 2$

Pour traiter le cas plus fréquent où $M > 2$, nous adoptons une stratégie de type *Branch and Bound* (BB) [14]. L'idée consiste à partir du meilleur noyau à disposition puis de sélectionner, parmi tous les noyaux restants, celui qui maximise l'alignement après combinaison. Ce choix est fait itérativement sur un horizon égal à 1 tant que l'alignement peut-être amélioré.

Supposons qu'à l'itération k de l'algorithme, on dispose de :

$$\hat{\boldsymbol{K}}^{(k)} = \sum_{i \in I} \alpha_i^{(k)} \boldsymbol{K}_i, \tag{16}$$

où I est un sous-ensemble de $\{1, \ldots, M\}$ et $\boldsymbol{K}_i$ la matrice de Gram associée au noyau κ_i. L'objectif est d'atteindre, à l'itération $k+1$, le meilleur alignement possible. D'après le paragraphe précédent, on est en mesure de déterminer la combinaison optimale

$$\hat{\boldsymbol{K}}^{(k+1)}(j) = \alpha^+_{1,j} \left(\sum_{i \in I} \alpha_i^{(k)} \boldsymbol{K}_i \right) + \alpha^+_{2,j} \boldsymbol{K}_j \tag{17}$$

au sens du critère d'alignement pour tout j. On détermine ainsi l'optimum local atteint à l'itération $k+1$ par :

$$\hat{\boldsymbol{K}}^{(k+1)} = \hat{\boldsymbol{K}}^{(k+1)}(j^*), \tag{18}$$

où

$$j^* = \arg\max_{j \notin I} \mathcal{A}(\hat{\boldsymbol{K}}^{(k+1)}(j), \boldsymbol{K}^*). \tag{19}$$

L'algorithme proposé Table 1 se résume alors ainsi : le noyau courant $\hat{\kappa}^{(k)}$, composé d'une combinaison linéaire des $\kappa_{i, i \in I}$, est amélioré en le combinant avec le noyau κ^*_j qui maximise l'alignement parmi tous les noyaux restants $\kappa_{j, j \notin I}$. En répétant cette stratégie jusqu'à ce que l'alignement ne progresse plus, on obtient, à défaut de l'optimum global, une combinaison (6) mieux adaptée aux données de l'ensemble d'apprentissage que chacun des noyaux κ_i pris individuellement.

Remarquons qu'un moyen simple d'accélérer l'algorithme et de précalculer les réels $\langle \boldsymbol{K}_i, \boldsymbol{K}_j \rangle_F$, $\|\boldsymbol{K}_i\|^2_F + \lambda$ et $\boldsymbol{y}^t \boldsymbol{K}_i \boldsymbol{y}$ intervenant lors du calcul de (13) et (14). De même, notons que le critère d'arrêt peut facilement être modifié de manière à intégrer une contrainte sur le nombre $M_{max} \geq 1$ de α_i non nul souhaité. Enfin, on peut également envisager, grâce au paramètre $0 \leq \eta \leq 1$, de mettre fin prématurément à l'algorithme si l'amélioration n'est pas significative :

Tant que $(k < M_{max})$ et $(\frac{\mathcal{A}_{BB}}{\mathcal{A}(j^*)} < \eta)$.

Table 2: Alignements individuels

	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)
I	.201	.208	.188	.151	.056	.136	.094	.080
B	**.079**	.207	.218	.189	.090	**.604**	.446	.095

I = ION, B = BCW

3 Résultats expérimentaux

Afin d'illustrer l'algorithme précédent, nous l'avons appliqué aux données[1] Ionosphere (ION) et Breast Cancer Wisconsin (BCW). Les noyaux considérés ont été, dans un premier temps, les noyaux polynômiaux de degré 1 (a), 2 (b), 3 (c) et 4 (d), et les noyaux gaussiens de paramètre σ égal à 0.1 (e), 5 (f), 10 (g) et 100 (h). Le paramètre de régularisation λ a quant à lui été fixé à 1 et l'on a considéré trois seuils η en deçà desquels l'amélioration de l'alignement n'a plus été jugée suffisante, provoquant l'arrêt prématuré de l'algorithme. Le tableau 2 fournit la valeur de l'alignement pour chacun des huit noyaux considérés individuellement.
L'alignement maximum pour les données ION, trouvé en résolvant (8) selon l'approche classique (QP : Quadratic Programming),[2] est $\mathcal{A}_{QP} = .2105$, une combinaison linéaire optimale associée étant $.0041(a) + .0008(b) + .2443(e)$. Pour BCW, l'optimum est $\mathcal{A}_{QP} = .6277$, obtenu pour $3.4 * 10^{-09}(a) + .5675(f)$. Le tableau 3 récapitule les résultats obtenus selon la stratégie proposée. Dans le premier cas (I), on remarque que l'alignement maximum $\mathcal{A}_{QP} = .2105$ est sensiblement identiques à l'alignement du noyau (b) pris individuellement. On peut donc écarter les noyaux (a) et (e) sans dégradation notable (voir résultats avec $\eta = .990$). Dans le second cas (B), la solution $\mathcal{A}_{BB} = .6277$ que nous obtenons coïncide avec $\mathcal{A}_{QP}$. On notera également la faible contribution de (a) comparativement à (f) et leur combinaison qui semble indiquer que l'amélioration de l'alignement s'effectue en combinant des noyaux qui ne sont pas alignés entre eux.

Nous avons ensuite procédé à un certain nombre de simulations sur le jeu de donnée BCW en particulier (voir Table 4). Pour chaque simulation, nous avons sélectionné aléatoirement 30 vecteurs paramétrant les noyaux générique suivant : le noyau polynomial, gaussien et exponentiel. L'optimisation par combinaison linéaire des 30 noyaux générés s'est ensuite déroulée comme précédemment. La solution trouvée par l'algorithme proposé apparaît plus parcimonieuse que celle obtenue selon l'approche QP. On constate également que notre programme est plus stable dans le sens où l'approche QP est sensible au

[1] ftp://ftp.ics.uci.edu/pub/machine-learning-databases/
[2] implémentation réalisée sous MATLAB avec *quadprog*

Table 3: Solutions trouvées par l'algorithme BB

	η	$\mathcal{A}_{BB}$	noyaux retenus
I	.990	.2081	(b)
	.995	.2095	$.0011(b) + .2266(e)$
	1	.2103	$.0036(a) + .0009(b) + .175(e)$
B	.990	.6277	$3.4 * 10^{-09}(a) + .5675(f)$
	.995	.6277	$3.4 * 10^{-09}(a) + .5675(f)$
	1	.6277	$3.4 * 10^{-09}(a) + .5675(f)$

I = ION, B = BCW

Table 4: Résultats pour 10 simulations comprenant chacune 30 noyaux paramétrés aléatoirement - données BCW

$\mathcal{A}_{min}$	$\mathcal{A}_{max}$	$\mathcal{A}_{BB}$	$\mathcal{A}_{QP}$	n_{BB}	n_{QP}
0.0785	0.5953	0.6094	0.2081	4	30
0.0785	0.4952	0.4963	0.4963	2	6
0.0785	0.1433	0.1433	0.1433	1	1
0.0785	0.1872	0.1872	0.1872	1	1
0.0785	0.1975	0.1975	0.1975	1	1
0.0785	0.3225	0.3630	0.2081	3	30
0.0785	0.6040	0.6040	0.6040	1	1
0.0785	0.3225	0.3450	0.3450	2	2
0.0785	0.2938	0.3002	0.3002	2	6
0.0785	0.2290	0.2290	0.2290	1	1
0.0785	0.3390	0.3475	0.2919	1.8	7.9

$\mathcal{A}_{min(max)}$ est l'alignement individuel le plus petit (grand) parmi les 30 noyaux de la collection considérée, n_{BB} est le nombre de $\alpha_i \neq 0$ obtenu par l'approche proposée et n_{QP} le nombre de $\alpha_i \neq 0$ obtenu par l'approche QP.

conditionnement du Hessien régissant le problème, signifiant que le choix du paramètre λ peut être critique. On remarque d'ailleurs que $\mathcal{A}_{QP}$ est quelques fois inférieur à $\mathcal{A}_{max}$ (voir résultats pour lesquels $n_{QP} = 30$). Cela correspond à des situations où le programme *quadprog* échoue numériquement.

4 Conclusion et perspectives

Nous avons mis en œuvre une stratégie simple pour l'optimisation, par combinaison linéaire, d'une collection de noyaux donnés. L'algorithme proposé permet de contrôler facilement la parcimonie de la solution, tout en restant stable numériquement. Les exemples choisis dans cet article sont restreints au domaine d'application qu'est la classification, même si une extension à la régression peut dores et déjà être envisagée [10]. Nous planifions également de comparer la méthode proposée ici avec la technique d'apprentissage utilisant le principe d'optimisation SDP (Semi-Definite Programming) [11]. Un autre axe de recherche est d'adopter un modèle additif/multiplicatif.

References

[1] N. ARONSZAJN. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, vol. 68, pages 337-404, 1950.

[2] C. CORTES ET V. VAPNIK. Support Vector Networks. *Machine Learning*, 20 : 1-25, 1995.

[3] M. COLLINS, R. E. SCHAPIRE ET Y. SINGER. Logistic regression, adaboost and bregman distances. *Machine Learning*. vol. 47, no 2-3, pages 253-285, 2002.

[4] K. CRAMMER, J. KESHET ET Y. SINGER. Kernel design using boosting. *Advances in Neural Information Processing Systems 15*. Cambridge, MA : MIT Press, 2003.

[5] B. SCHÖLKOPF ET A.SMOLA. Learning with kernels. MIT Press, Cambridge, MA, 2002.

[6] N. CRISTIANINI, J. SHAWE-TAYLOR, A. ELISSEEFF ET J. KANDOLA. On kernel-target alignment. *Advances in Neural Information Processing Systems 14*, Cambridge, MA : MIT Press, pages 367-373, 2002.

[7] L. WANG ET K. LUK CHAN. Learning kernel parameters by using class separability measure. *Neural Information Processing Systems (NIPS)*, Canada, 2002.

[8] M. MEILA. Data centering in feature space. *Proceedings of Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[9] J. KANDOLA, J. SHAWE-TAYLOR ET N. CRISTIANINI. Optimizing kernel alignment over combinations of kernels. Technical Report 121, Department of Computer Science, University of London, http://www.neurocolt.org, 2002.

[10] J. KANDOLA, J. SHAWE-TAYLOR ET N. CRISTIANINI. On the extensions of kernel alignment. Technical Report 120, Department of Computer Science, University of London, http://www.neurocolt.org, 2002.

[11] G. LANCKRIET, N. CRISTIANINI, P. BARTLETT, L. EL GHAOUI ET M. I. JORDAN. Learning the kernel matrix with semi-definite programming. *Proceedings of the 19th International Conference on Machine Learning*, pages 323-330, 2002.

[12] J. PLATT. Fast training of Support Vector Machines using Sequential Minimal Optimization. *Advances In Kernel Methods - Support Vector Learning*, chapitre 12, pages 185-208, MIT Press, 1999.

[13] G. ZOUTENDIJK. Methods of feasible directions: a study in linear and non-linear programming. Elsevier, 1970.

[14] E.L LAWLER ET D.E WOOD. Branch and Bound methods: a survey. *Operations Research*, vol. 149, pages 699-719, 1966.

Reconstruction par noyaux polynomiaux

M. BERAR[1], M. DESVIGNES[1], B. ROMANIUK[2], G. BAILLY[3], Y. PAYAN[4]

[1]LIS, 961, rue de la Houille Blanche, 38402 St Martin d'Hères Cedex
[2] Laboratoire CResTIC LERI, URCA, rue des Crayères BP 1035, 51687 Reims cedex 2,France
[3] Institut de la Communication Parlée (ICP), UMR CNRS 5009, INPG/U3, 46,av. Félix Viallet, 38031 Grenoble, France
[4] Techniques de l'Imagerie, de la Modélisation et de la Cognition (TIMC), Faculté de Médecine, 38706 La Tronche, France

berar@lis.inpg..fr desvignes@lis.inpg.fr barbara.romaniuk@leri.univ-reims.fr Yohan.Payan@imag.fr

Résumé – Le problème traité est celui de la reconstruction de données partiellement connues. Nos données sont modélisées dans un espace caractéristique issu d'une fonction de projection non-linéaire et nous proposons une solution directe et un formalisme issus de l'astuce noyau. L'originalité de ce travail se situe dans la formalisation de cette reconstruction. La présence de données partiellement connues reformulant entièrement le problème connu de la recherche d'antécédents dans le cadre des méthodes à noyaux traditionnelles. Nous présentons un formalisme nouveau ainsi que son application aux noyaux polynomiaux, déduisant simultanément les paramètres de variabilité caractéristique non-linéaire et les données inconnues. Ce formalisme est ensuite appliqué à des données synthétiques et à l'estimation de points céphalométriques.

Abstract *– In this paper, we deal with the problem of partially observed objects. These objects are defined by a set of points and their shape variations are represented by a statistical model. We present two models in this paper: a linear model based on PCA and a non-linear model based on KPCA. The present work attempts to localize of non visible parts of an object, from the visible part and from the model, using the variability represented by the models. Both are applied to synthesis data and to cephalometric data with good results*

1. Introduction

Les modèles statistiques de variabilité de formes sont couramment utilisés en analyse ou synthèse d'image [1] à l'aide de modèles de formes et d'apparence [2]. Un ensemble d'apprentissage permet de construire un modèle compact par Analyse en Composantes Principales (ACP) qui est ensuite utilisé pour des applications de compression [3], débruitage [4] ou de segmentation par contour actif. Dans ce papier, nous nous plaçons dans le cadre où une partie des données ou de l'objet sont connues et où il s'agit de retrouver la partie inconnue à l'aide du modèle statistique.

Dans ce cadre, l'ACP permet de modéliser les relations linéaires entre les composantes de l'objet, à travers des paramètres de variabilité. Ces relations linéaires peuvent être assimilées à des relations spatiales. Quand certaines composantes sont manquantes ou occultées, nous montrons que ces relations permettent de reconstruire la forme.

Dans le cas d'une modélisation statistique non-linéaire obtenue par Analyse en Composantes Principales à Noyau, les paramètres de variabilité se situent alors dans un espace caractéristique différent de l'espace d'origine des données. Le passage de l'espace d'origine à l'espace caractéristique est rendu implicite par l'emploi de l'astuce noyau. La difficulté est ici de retrouver les composantes dans l'espace d'origine connaissant celles de l'espace caractéristique. Ce problème est celui de la pré-image. De même que dans le cas de l'ACP, nous montrons dans ce papier que cette formulation peut être étendu à notre problématique et nous la développons pour certains noyaux. Si une partie de la forme est occultée, la reconstruction de la forme peut être réalisé par une minimisation utilisant les relations entre composantes. Nous avons appliqué cette méthode à des données de synthèse de caractère non-linéaire ainsi qu'à des données céphalométriques en utilisant des noyaux polynomiaux et obtenons de bons résultats.

1. Modèles linéaires

La reconstruction avec données manquantes dans le cadre d'un modèle linéaire coïncide avec la reconstruction. En effet, soit un modèle linéaire construit par ACP sur un jeu de n+m coordonnées pour l > n+m individus :

$$\begin{bmatrix} C_1 \\ \vdots \\ C_n \\ X_1 \\ \vdots \\ X_m \end{bmatrix} = \begin{bmatrix} \overline{C}_1 \\ \vdots \\ \overline{C}_n \\ \overline{C}_{n+1} \\ \vdots \\ \overline{C}_m \end{bmatrix} + \begin{bmatrix} \Phi_{1,1} & \cdots & \Phi_{1,n+m} \\ \vdots & \ddots & \vdots \\ \Phi_{n+m,1} & \cdots & \Phi_{n+m,n+m} \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_n \\ b_{n+1} \\ \vdots \\ b_{n+m} \end{bmatrix} \quad (1)$$

où $\overline{C}$ est la forme moyenne des individus, $\Phi = (\phi_1 | ... | \phi_t)$ est une matrice $(n+m)\times t$ composée des t premiers vecteurs propres de la matrice de covariance des données centrées $(t \le n+m)$et b un vecteur de dimension t tel que $b = \Phi^t(T_i - \overline{T})$. Le vecteur b est alors une bonne approximation du jeu de données originale et n'importe lequel des n+m points peut être représenté ou retrouvé avec les valeurs du vecteur b.

Dans le cas où les *m* dernières coordonnées sont inconnues, ($z = (c_1,\ldots,c_n,x_1,\ldots,x_m)$) on a alors un système composé de t+m inconnues pour n+m équations. Pour t >n, le système est sous-déterminé et ne peut être résolu, pour t=n, il existe une solution exacte, et pour t <n le système est surdéterminé et la solution est donnée par une minimisation au sens des moindres carrés. Par ailleurs, le système peut s'exprimer sous la forme suivante :

$$\left\| \begin{bmatrix} C_1-\overline{C}_1 \\ \vdots \\ C_n-\overline{C}_n \\ -\overline{C}_{n+1} \\ \vdots \\ -\overline{C}_{n+m} \end{bmatrix} - \begin{bmatrix} \phi_{1,1} & \cdots & \phi_{1,t} & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi_{n,1} & \cdots & \phi_{n,t} & 0 & \cdots & \cdots & 0 \\ \phi_{n+1,1} & \cdots & \phi_{n+1,t} & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ \phi_{n+m,1} & \cdots & \phi_{n+m,t} & 0 & \cdots & 0 & -1 \end{bmatrix} \begin{bmatrix} b_1 \\ \vdots \\ b_t \\ X_1 \\ \vdots \\ X_n \end{bmatrix} \right\|^2 \quad (2)$$

2. Modèles non linéaires

L'ACP à Noyaux peut être considérée comme une généralisation naturelle de l'ACP linéaire et est particulièrement appropriée pour extraire des caractéristiques non linéaires de données. Proches des méthodes appliquées dans le domaine des Machines à Support de Vecteurs, l'ACP à Noyaux projette tout d'abord les données de l'espace d'entrée $\mathcal{I}$ dans un espace haute dimension $\mathcal{F}$ via une fonction habituellement non-linéaire. Une ACP est ensuite appliquée sur ces données projetées (que l'on suppose centrées dans $\mathcal{F}$).

2.1 ACP à Noyaux

Soit l'ensemble d'apprentissage $\{x_i$ i=1 à L$\}$Soit $\varphi : \mathbb{R}^n \to \mathcal{F}$ notre fonction de projection, un noyau de Mercer permet d'exprimer directement le produit scalaire de la fonction de projection $\varphi(x) \bullet \varphi(y) \equiv k(x,y)$. Parmi ces noyaux de Mercer, on peut citer les noyaux gaussiens $k(x,y) = \exp(-\|x-y\|^2 / c)$ ($\mathcal{F}$ de dimension infinie avec φ inconnue) et les noyaux polynomiaux $k(x,y) = (1 + x \bullet y)^d$ ($\mathcal{F}$ de dimension finie avec φ explicite). La matrice de noyaux K de taille NxN est définie par $K_{ij} \equiv \varphi(x_i) \bullet \varphi(x_j)$.

Le problème revient alors à déterminer les vecteurs propres de [5] :

$$N\lambda\alpha = K\alpha \quad (3)$$

La k$^{\text{ième}}$ composante non-linéaire de la φ-image d'un individu x est la paramètre de variabilité β_k :

$$\beta_k = (V^k \bullet \varphi(x)) = \sum_{i=1}^{L} \alpha_i^k k(x, x_i) \quad (4)$$

Si les observations ne sont pas centrées dans $\mathcal{F}$, on utilise la matrice de Gramm à la place de la matrice à noyau:

$$\tilde{K}_{ij} = K_{ij} - \frac{1}{N}\sum_{p=1}^{N} K_{ip} - \frac{1}{N}\sum_{q=1}^{N} K_{qj} + \frac{1}{N^2}\sum_{p,q=1}^{N} K_{pq} \quad (5)$$

2.2 Reconstruction

La reconstruction sur les premières composantes pose le problème de l'antécédent *f* d'un point de $\mathcal{F}$, parfois inexistant. Le problème est souvent abordé [6] en minimisant la distance entre φ(f), et la projection de notre forme réduite à ses premières composantes

$$\begin{aligned} &\min_{f \in R^n} \left\| \varphi(f) - P_N \varphi(z) \right\|^2 \\ &= \min_{f \in R^n} \left(k(f,f) - 2\sum_{k=1}^{N} \beta_k \left(\sum_{i=1}^{L} \alpha_i^k k(f, x_i) \right) + \Omega \right) \end{aligned} \quad (6)$$

Cette minimisation dépend de la forme du noyau choisi. On peut remarquer que le dernier terme de la quantité à minimiser est constant et que les paramètres de variabilité sont connus a priori.

2.3 Reconstruction des données manquantes

Soit la forme $z = (c_1,\ldots,c_n,x_1,\ldots,x_m)$ dont on connaît les *n* premières coordonnées $(c_1,\ldots,c_n)$ et appartenant au modèle. Trouver la partie inconnue de z $(x_1,\ldots,x_m)$, c'est déterminer conjointement les N premiers paramètres de variabilité β et la partie inconnue de z, c'est-à-dire minimiser la fonction suivante :

$$\begin{aligned} &\rho(z = (c_1,\cdots,c_n,x_1,\cdots,x_n)) = \left\| \varphi(z) - P_N \varphi(z) \right\|_2^2 \\ &= k(z,z) - \sum_{k=1}^{N} \left(\sum_{i=1}^{L} \alpha_i^k k(z, x_i) \right) \\ &\text{avec } \beta_k = \sum_{i=1}^{L} \alpha_i^k k(z, x_i) \end{aligned} \quad (7)$$

Nous développons la suite de l'article dans le cadre des noyaux polynomiaux. La fonction à minimiser prend alors la forme :

$$\rho(z)_{\min} = (1 + x_c \bullet x_c + z_x \bullet z_x)^d - \sum_{k=1}^{N} \left(\sum_{i=1}^{L} \alpha_i^k (1 + x_c \bullet x_{ci} + z_x \bullet x_{xi}) \right)^d \quad (8)$$

2.3.1 Polynôme de degré un

Pour des observations centrées, on a $k(x,y) = (x \bullet y)$, φ correspond à l'identité. Minimiser ρ en annulant le gradient de ρ mène à la solution :

$$x = \frac{\sum_{k=1}^{N} \left(\sum_{i=1}^{L} \alpha_i^k \left(c \bullet x_{ci} \right) \right) \left(\sum_{i=1}^{L} \alpha_i^k x_{xi} \right)}{1 - \sum_{k=1}^{N} \left(\sum_{i=1}^{L} \alpha_i^k x_{xi} \right)^2} \quad (9)$$

2.3.2 Polynôme de degré 2

La fonction de projection est $\varphi(x) = (x_1,\ldots,x_n,x_1^2,\sqrt{2}x_1x_2,\ldots,\sqrt{2}x_1x_n,x_2^2,\sqrt{2}x_2x_3,\ldots,x_n^2)$. L'équation (3) devient le polynôme de degré 4 suivant :

$$\rho(z)_{\min} = C_{00} + \left\|\tilde{\varphi}_c(z)\right\|^2 - \sum_{k=1}^{N}\left(C_{0k} + C_{1k} \bullet \tilde{\varphi}_x(z)\right)^2$$

$$C_{0k} = \sum_{i=1}^{L} \alpha_i^k \left(\tilde{\varphi}_c(z) \bullet \tilde{\varphi}_c(x_i)\right), \qquad (10)$$

$$C_0 = \left\|\tilde{\varphi}_c(z)\right\|^2,$$

$$C_{1k} = \sum_{i=1}^{L} \alpha_i^k \tilde{\varphi}_x(x_i)$$

Les solutions $(x_1, \ldots, x_m)$, partie inconnue de z, sont obtenues par une minimisation par descente de gradiant.

$$\nabla_x \rho(z) = 2\left(\tilde{\varphi}_x(z) \bullet \dot{\tilde{\varphi}}_x(z)\right)$$

$$-2\left(\sum_{k=1}^{N}\left(C_{0k}C_{1k} \bullet \dot{\tilde{\varphi}}_x(z) + \left(C_{1k} \bullet \dot{\tilde{\varphi}}_x(z)\right)\left(C_{1k} \bullet \tilde{\varphi}_x(z)\right)\right)\right) \qquad (11)$$

En effet, le gradiant est un polynôme de degré 3 et il n'y a pas de solution analytique directe. L'initialisation de la minimisation par descente de gradiant est donnée par la forme moyenne du modèle $X = \left(\overline{c}_{n+1}, \cdots, \overline{c}_{n+m}\right)$ pour la partie inconnue.

3. Résultats

3.1 Données synthétiques

La reconstruction des données manquantes a été appliquée à des jeux de données synthétiques présentant un caractère non-linéaire (fig.1).

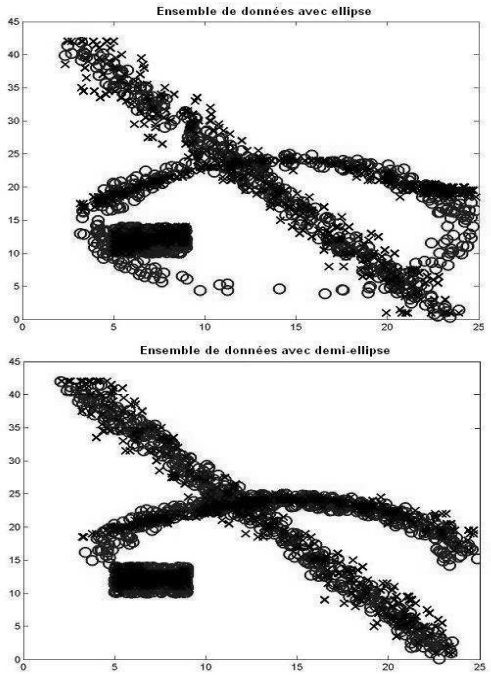


FIG. 1 **:** Jeux de données synthétiques. 3 points avec des relations linéaires et non-linéaires. Les cercles correspondent aux données initiales, les croix aux données reconstruites quand 3 valeurs sont manquantes.

Dans ce premier test, un ensemble de données composées de trios points (c.-a.-d.. six valeurs) est généré (fig 1). Trois paramètres sont nécessaires à la description de ces données, le nombre optimal de paramètres de variabilité est donc de 3 pour l'ACP et l'ACP à noyaux.

1 Parmi les données synthétiques, un point est une constante. Ce point devrait être prédit facilement de par sa valeur moyenne.

2 Les variations d'un autre point sont linéaires. L'ACP linéaire comme l'ACP à noyaux devrait le prédire avec une bonne précision.

3 Les variations du dernier point sont non-linéaires. Ce point décrit une ellipse. Un premier ensemble de données décrit une ellipse complète avec donc une indétermination sur l'axe y quand x est connu. Un deuxième ensemble de données décrit une demi-ellipse où il n'y a plus d'indétermination.

Un bruit indépendant uniforme est ajouté à chacun des points. Les modèles ACP et ACP à noyaux sont appris sur une base de données de 50 exemples et testés sur des ensembles de 300 exemples. Dans ce test de reconstruction, la seconde coordonnée de chaque point est supprimée et les données manquantes sont estimées par le modèle.

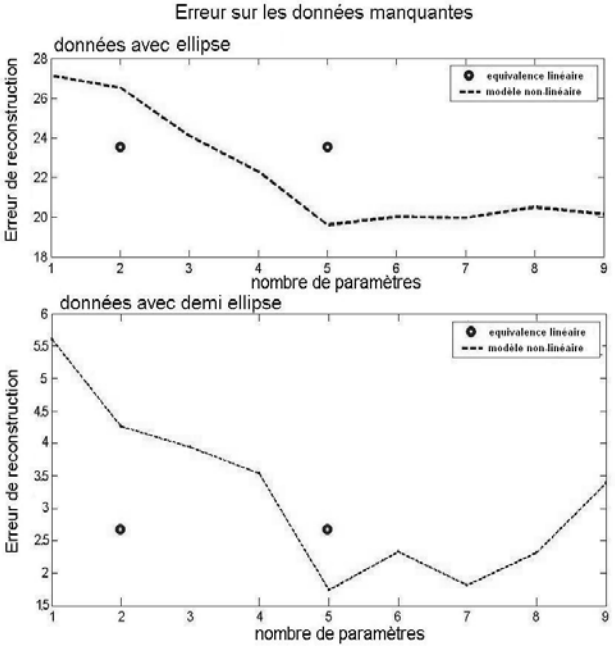


FIG. 2 : Erreur de reconstruction pour 3 points manquants

Tout d'abord, l'expression à minimiser a bien un minimum et la largeur de ce minimum correspond à la largeur du bruit ajouté. L'erreur sur les inconnues estimées est résumée dans la figure 2. Les résultats obtenus montrent l'avantage de la méthode non linéaire. Le caractère non-linéaire des données est pris en compte dans le modèle polynomial et permet un gain de 4 mm sur l'erreur de reconstruction du jeu avec l'ellipse complète. Un gain est aussi observé pour le jeu de données avec la demi ellipse.

Par contre en termes computationnels, il faut étendre le nombre de paramètre pour obtenir ce résultat. En effet, de par la projection dans le noyau polynomial, une seule donnée manquante affecte toutes les composantes qui en sont tirées (x, x^2, x_1x …). Ainsi il ne reste que 9 neufs paramètres de variabilités libres pour 3 valeurs manquantes (5 pour 4 valeurs manquantes et 2 pour 5 valeurs manquantes). Pour le modèle linéaire équivalent, il existe 3 paramètres de variabilité pour 3 données manquantes (2 paramètres pour 4 valeurs et 1 paramètres pour 5 valeurs).

3.2 Données réelles

L'un des objets d'étude de la céphalométrie est la croissance du crane des jeunes enfants afin d'améliorer la thérapie en orthodontie [2,8]. La céphalométrie est basée sur le placement d'amers, les points céphalométriques, sur des téléradiographies, images deux dimensions aux rayons X de la vue sagittale du crâne (FIG.3). Ces points sont utilisés pour le calcul de traits, tels que la longueur d'un segment ou l'angle entre deux lignes. L'interprétation de ces traits est utilisée pour diagnostiquer la déviation de la forme du patient de la forme idéale. Cette interprétation est aussi utilisée lors de l'évaluation d'un traitement orthodontique. Les points céphalométriques sont liés à la forme du contour crânien. Dans ce contexte le contour crânien est échantillonné et les amers sont apprises avec le contour crânien [9].

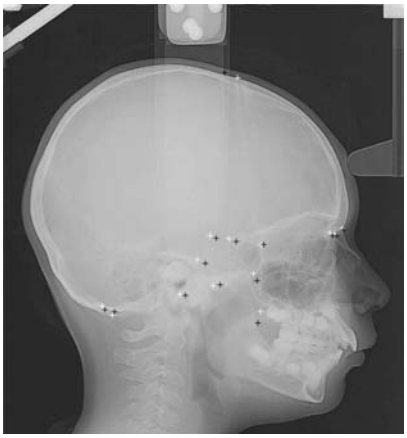


FIG. 3 **:** Céphalogramme, contour crânial, points réels (blancs) et estimé (noirs).

Placer les amers d'un nouveau céphalogramme, connaissant le contour revient à déterminer la partie inconnue (les amers céphalométriques) d'une détermination du modèle à l'aide du modèle statistique et de la partie connue (le contour crânien).

La base d'apprentissage est constituée de 70 individus pour seulement 8 exemples de test. Sur ces données réelles, l'erreur de prédiction des sur les points céphalométriques est 3,88 mm pour l'ACP linéaire et de 2.25 mm pour l'ACP à Noyau polynomial de degré 2. De plus, Sur certains points, l'erreur de reconstruction est équivalente à la variabilité inter-expert de 1 mm.

4. Conclusion

Dans cet article, un modèle de forme basé sur un noyau polynomial est présenté. Ce modèle non-linéaire est utilisé pour résoudre un problème de données manquantes dans une image.

L'équation de minimisation trouvée peut être résolue numériquement dans le cas générale. Les paramètres de forme et les données manquantes sont alors estimés conjointement. Dans le cadre des noyaux polynomiaux, il s'agit de trouver les racines d'une équation polynomiale et les solutions les plus robustes.

Ce modèle basé sur des noyaux polynomiaux a été comparé à un modèle ACP linéaire classique sur des données synthétiques et réelles. Quand des relations non-linéaires sont présentes entre les données, celles-ci sont représentées avec une meilleur précision par le modèle à noyau que par le modèle linéaire, au prix d'un coût computationnel plus grand.

Références

[1] T.F.Cootes, G.J. Edwards, C.J.Taylor. *Active Appearance Models*, IEEE PAMI, Vol. 23 (6), pp. 681-685, 2001.

[2] T.J. Hutton, S. Cunningham, P. Hammond. *An Evaluation of Active Shape Models for the Automatic Identification of Cephalometric Landmarks*. European Journal of Orthodontics, Vol. 22(5), pp. 499-508, 2000.

[3] S. Mika, B. Schölkopf, A.J. Smola, K.-R. Müller, M. Scholz, and G. Rätsch. *Kernel PCA and de-noising in feature spaces.* In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, Advances in Neural Information Processing Systems 11, pages 536-542. MIT Press, 1999.

[4] B. Schölkopf, A. Smola et K. Müller. *Non linear component Analysis as a Kernel Eignevalue Problem.* Neural Computation, 10(5): 1299-1319, 1998.

[5] J. T. Kwok et Ivor W. Tsang. *The Pre-Image Problem in Kernel Methods.* Proceedings of ICML 2003 : pp.408-415, 2003.

[6] Bakir, G.H., J. Weston and B. Schölkopf: *Learning to Find Pre-Images.* Advances in Neural Information Processing Systems, 16, 449-456. (Eds.) Thrun, S., L. Saul and B. Schölkopf, MIT Press, Cambridge, MA, USA (2004)

[7] S. Sclaroff, AP..Pentland, *Modal Matching for Correspondence and Recognition.* IEEE Transactions on Pattern Recognition and Machine Intelligence, 17(6):545-561, 1995.

[8] B. Romaniuk, M. Desvignes, M. Revenu, M.J. Deshayes *Linear and Non-Linear Model for Statistical Localization of Landmarks*, ICPR, Vol. 4, pp. 393-396, 2002

[9] B. Romaniuk, M. Desvignes, *Contour Tracking by Minimal Cost Path Approach. Application to Cephalometry.* International Conference on Image Processing ICIP, Singapore, October 2004

Classification de végétaux par SVM

M. CHAPRON[1], L. ASSEMAT[2],
[1]ETIS-ENSEA-UCP, UMR 8051, 6 avenue du ponceau, 95014 Cergy-Pontoise
[2]INRA, UMR Biologie et Gestion des Adventices, BP 86510, Dijon

chapron@ensea.fr

Résumé – Le travail décrit ici consiste à présenter les méthodes mises en œuvre pour classer les pixels des images couleur dans les classes maïs, adventices et sol dans le cadre de l'agriculture de précision. Les différentes méthodes de classification SVM ont été décrites et testées. Les méthodes de classification linéaires se sont avérées insuffisantes pour classifier les pixels en sol/végétation, les méthodes non linéaires et à classes non séparables se sont montrées nettement meilleures tant pour classer les pixels en sol/végétation, puis les pixels de végétation en maïs et en adventices.

Abstract – *This paper is concerned with the classification of pixels of color images into maize, weeds and soil in the framework of precision agriculture. The different SVM based classification methods have been described and tested. The linear techniques provide poor results during the classification of pixels into soil and vegetation, the non linear and non separable ones give much better results for classifying the pixels into soil and vegetation as well as classifying the pixels of vegetation into maize and weeds.*

1.Introduction

La reconnaissance ou la classification de végétaux [2, 6] est un problème difficile et important en agriculture de précision et particulièrement pour la diminution des intrants chimiques qui créent de plus en plus de pollutions d'origine chimique dans les nappes phréatiques (atrazine). Le développement technologique des caméras, ordinateurs et méthodes de classification/reconnaissance permet d'envisager des applications réelles de réduction d'intrants chimiques tels que les herbicides. En effet, les herbicides peuvent n'être répandus que si la présence des adventices (mauvaises herbes) est nocive pour le rendement de la culture céréalière lors de la compétition entre végétaux. Dans la suite, il sera étudié la densité des mauvaises herbes dans une culture de maïs.

Les étapes du traitement proposé sont d'abord la classification des pixels en deux classes : sol et végétation, puis les pixels de la végétation sont classés à nouveau en deux classes : maïs et adventices. On comparera les résultats obtenus par les méthodes SVM (support vector machine) linéaires, non linéaires et donnera les résultats de classification maïs/adventices avec les SVM non linéaires, non séparables. Sur les images présentées [1, 3, 4, 7, 8, 9, 10], les caractéristiques utilisées pour la classification des végétaux sont les trois composantes rouge, verte et bleu normalisées ou pas. Le formalisme et les méthodes seront d'abord décrits dans le cas général puis adaptés au cas des végétaux.

2. SVM linéairement séparable

Soit un ensemble d'apprentissage A formé de N couples (x_i,y_i) de données et d'étiquettes où x_i est une donnée et y_i l'étiquette associée, x_i appartient à R^d et y_i à C={-1,1}, l'ensemble des deux classes. L'ensemble A s'écrit comme suit : A = $\{(x_i,y_i), i=1..N\}$. L'objectif des SVM dans le cas linéaire est de calculer un hyperplan qui sépare au mieux les échantillons des deux classes. Cet hyperplan f(x) est défini par :

$$f(x)= <w,x> + b = \sum_{j=1}^{d} w_j x_j + b = 0 \qquad (1)$$

avec $(w,x) \in R^d \times R^d$ et $b \in R$,

Les SVM cherchent à maximiser la distance entre l'hyperplan et les points les plus proches de lui de chaque classe, cette distance est égale à $2/\|w\|$. En conséquence, le problème d'optimisation consiste à maximiser cette distance ou à minimiser son inverse sous les contraintes : $y_i(\langle w,x_i\rangle+b)\geq 1 \quad i=1..N$

Nous ne décrivons pas l'ensemble des équations qui régissent ce modèle qui ne s'est pas avéré le plus performant ; les modèles ci-dessous ont donné des résultats bien meilleurs.

En phase opérationnelle, pour une donnée nouvelle x, f(x) est calculée et en fonction du signe de f(x), x est classée -1 si f(x) est négative et +1 dans le cas contraire. Nous pouvons remarquer que la solution ne dépend pas de la dimension des données (d) mais du nombre d'échantillons et en fait du nombre de vecteurs supports.

3. SVM non linéaires

Dans le cas fréquent où les données ne sont pas linéairement classifiables, le formalisme des SVM propose de dépasser cette difficulté en projetant les données dans un espace P de dimension supérieure en général à d (dimension de l'espace des données brutes) grâce une fonction non linéaire $\Phi(x)$.

La même procédure d'optimisation que précédemment est appliquée sur les données projetées par $\Phi(x)$ dans P. Dans ce cas, (4) et (5) deviennent respectivement (8) et (9) :

$$\mathrm{Min}_{\alpha*} \frac{1}{2} \sum_{l,k=1}^{N} \alpha_l \alpha_k y_l y_k \langle \Phi(x_l), \Phi(y_k) \rangle - \sum_{i=1}^{N} \alpha_i \qquad (2)$$

et $\sum_{i=1}^{N} \alpha_i y_i = 0$ et $\alpha_i \geq 0$, i=1..N

$$f(x)= \sum_{i=1}^{N} \alpha_i^{*} y_i \langle \Phi(x_i), \Phi(x) \rangle + w_0^{*} \quad (3)$$

La même technique d'optimisation quadratique que ci-dessus a permis d'obtenir α_i^{*} et w_0^{*}.

La difficulté ici vient des calculs des produits scalaires $\langle \Phi(x_i), \Phi(x) \rangle$ qui deviennent très longs. Pour outrepasser cette difficulté, il existe des fonctions bilinéaires symétriques positives K(x,y) appelées fonctions à noyau faciles à calculer et qui correspondent à des produits scalaires dans un espace à grande dimension ; cette propriété correspond à la condition de Mercer. Le problème d'optimisation (2) dans ce cas devient (4) :

$$\text{Min}_{\alpha^*} \frac{1}{2} \sum_{l,k=1}^{N} \alpha_l \alpha_k y_l y_k K(x_l, y_k) - \sum_{i=1}^{N} \alpha_i \quad (4)$$

et $\sum_{i=1}^{N} \alpha_i y_i = 0$ et $\alpha_i \geq 0$, i=1..N

La solution est l'hyperplan séparateur d'équation :

$$f(x)= \sum_{i=1}^{N} \alpha_i^{*} y_i K(x_l, y_k) + w_0^{*} \quad (5)$$

α_i^{*} et w_0^{*} sont obtenus par la même technique d'optimisation quadratique que ci-dessus. Les fonctions à noyau les plus utilisées sont les fonctions polynomiales, à base radiale et sigmoïdes. L'équation correspondant à la fonction à base radiale est :

$$K(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \quad (6)$$

σ est donné par l'utilisateur.

Le choix de la fonction noyau à base radiale permet dans nos applications de régulariser les fonctions de décision obtenues.

4. SVM non séparables

Si l'ensemble d'apprentissage A n'est pas séparable, le problème d'optimisation (2) n'a pas de solution, il faut introduire des variables ξ_i de relâchement des contraintes et un hyper paramètre C dans la fonction de coût à minimiser. Le problème d'optimisation devient (7) :

$$\text{Min}_{w,b,\xi_i} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i \quad (7)$$

avec $y_i f(x_i) \geq 1 - \xi_i$ et $\xi_i \geq 0$ pour i=1..N

Cette optimisation sous contraintes peut être effectuée en utilisant les multiplicateurs de Lagrange, elle s'écrit :

$$L(w,b,\xi,\alpha,\beta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i [\xi_i - 1 + y_i(w^T x_i + b)] - \sum_{i=1}^{N} \beta_i \xi_i \quad (8)$$

Les dérivées partielles de L par rapport à w, b, et ξ_i donnent :

$$\frac{\partial L(w,b,\xi,\alpha,\beta)}{\partial \omega} = w - \sum_{i=1}^{N} \alpha_i y_i x_i = 0 \quad (9)$$

$$\frac{\partial L(w,b,\xi,\alpha,\beta)}{\partial b} = \sum_{i=1}^{N} \alpha_i y_i = 0 \quad (10)$$

$$\frac{\partial L(w,b,\xi,\alpha,\beta)}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \quad \text{pour i=1..N} \quad (11)$$

(9) donne $w= \sum_{i=1}^{N} \alpha_i y_i x_i$ (12)

(10) et (12) dans (8) donnent :

$\text{Min}_{w,b,\xi} L(w,b,\xi,\alpha,\beta) =$

$$-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j x_i^T x_j + \sum_{i=1}^{N} \alpha_i \quad (13)$$

(13) doit être maximisé par rapport à α *et* β, β n'apparaît pas dans (13), donc maximisons (13) par rapport à α et qu'il existe β qui vérifie (11), c'est le cas quand $\alpha_i \leq C$ car $\beta_i \geq 0$.

Le problème initial d'optimisation se réduit à minimiser le critère $Cr(\alpha)$ suivant :

$$Cr(\alpha) = -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} y_i y_j \alpha_i \alpha_j x_i^T x_j + \sum_{i=1}^{N} \alpha_i \quad (14)$$

Sous les contraintes :

$\sum_{i=1}^{N} \alpha_i y_i = 0$ et $0 \leq \alpha_i \leq C$ pour i=1..N

Remarquons qu'une fois α trouvé, β l'est aussi par (11) :

$C - \alpha_i = \beta_i \quad \forall i \in \{1..N\}$

Pour déterminer b, il est utilisé le fait que les conditions de Karash-Kuhn-Tucker correspondant aux multiplicateurs de Lagrange non nuls sont atteintes au point sel du Lagrangien.

Cela implique que les contraintes $\xi_i = 0$ et

$\xi_i - 1 + y_i(w^T x_i + b) = 0$ sont encore vraies, elles donnent avec (12) :

$$b = y_i - w^T x_i = y_i - \sum_{j=1}^{N} y_j \alpha_j x_j^T x_i \quad (15)$$

La classification non linéaire pour une donnée d'entrée x se fait à partir de l'équation :

$$f(x) = \sum_{i=1}^{N} y_i \alpha_i k(x_i, x) + b \quad (16)$$

Selon le signe de f(x), x reçoit l'étiquette +1 si f(x)>0 et l'étiquette -1 dans le cas contraire.

5. Résultats

D'abord, les images 1 et 2 présentent les résultats de la séparation linéaire pour classer les pixels en sol et végétation. Les résultats ne sont pas de bonne qualité, ils sont nettement améliorés par les SVM non linéaires image 3. Cette méthode qui utilise en entrée des données vectorielles x composée des valeurs des pixels rouge, vert et bleu

divisées par 255 ainsi leurs valeurs normalisées du rouge et du vert dans un voisinage 7x7 autour du pixel courant. Cela fait 245 composantes dans le vecteur x. En apprentissage, deux échantillons ont été pris sur le sol, deux autres sur les cailloux pour la classe sol ; deux échantillons ont été prélevés sur le maïs et deux autres sur les adventices pour la deuxième classe. La fonction noyau utilisée est la fonction à base radiale. Avec très peu d'échantillons en apprentissage, les résultats sont excellents car ils mettent en évidence des petites plantules uniquement visibles après zoom. Les résultats sur la végétation (classification maïs/mauvaises herbes) apparaissent sur les images 4, 5 et 6 en utilisant le cas non séparable : pour obtenir l'image 5 une SVM non linéaire a été utilisée, pour avoir l'image 6 une SVM non linéaire et non séparable a été mise en œuvre. Dans ce cas aussi très peu d'échantillons ont été pris : 6 sur le maïs autant sur les adventices. Les voisinages utilisés sont aussi 7x7 et les rouge et vert normalisés n'ont pas été pris en compte, les vecteurs données ont donc 147 composantes. Les résultats sont bons dans une végétation complexe à cause de ses multiples chevauchements de feuilles de végétaux. Ces résultats peuvent être améliorés par une méthode segmentation qui reprend ces résultats de classification et fusionne les plus petites régions aux plus grandes qui leur sont connexes. L'utilisation d'espaces colorimétriques différents (teinte, luminance et saturation), Yuv et Lab n'ont pas améliorés les résultats de classification. Dans ce genre de scènes naturelles, les méthodes non linéaires donnent de meilleures résultats que les méthodes linéaires. A partir de l'image 1, des essais ont été conduits pour distinguer les trois espèces végétales (maïs et deux adventices), les résultats obtenus avec la technique précédente utilisant les mêmes descripteurs et une SVM multi-classe étaient médiocres, ce qui a conduit à ajouter les caractéristiques de texture sur les feuilles des végétaux pour mieux les classer. Les matrices de coocurrence calculées sur les trois composantes chromatiques et selon les directions horizontale et verticale [5] permettent d'intégrer les signatures texturales des végétaux. Les 4 paramètres utilisés à partir des matrices de coocurrence ont été l'énergie, l'entropie de Shannon, le max et le contraste. Le tableau ci-dessous synthétise les résultats obtenus sur ces 4 critères, les 3 premières colonnes ont été calculées à partir des 3 composantes rouge, verte et bleue sur les 12 échantillons d'apprentissage avec la direction horizontale pour le calcul des matrices de coocurrence, les trois colonnes suivantes ont été construites sur le même modèle mais avec la direction verticale pour le calcul des matrices de coocurrence.

Chardon	moyenne					
Energie	0,052	0,043	0,045	0,061	0,053	0,052
Entropie	3,015	3,155	3,119	2,861	2,961	2,966
Max	0,091	0,056	0,070	0,109	0,075	0,060
Contraste	4,432	7,104	7,429	3,352	6,775	7,065
	Variance					
Energie	0,008	0,004	0,005	0,008	0,003	0,003
Entropie	0,106	0,056	0,079	0,106	0,038	0,051
Max	0,031	0,034	0,034	0,037	0,027	0,022
Contraste	4,203	5,638	6,901	2,300	2,725	3,641
Maïs	moyenne					
Energie	0,054	0,045	0,050	0,063	0,056	0,058
Entropie	2,978	3,123	3,056	2,830	2,906	2,890
Max	0,090	0,063	0,070	0,100	0,084	0,093
Contraste	2,764	3,688	4,318	2,733	3,859	4,111
	Variance					
Energie	0,012	0,004	0,009	0,010	0,005	0,008
Entropie	0,173	0,064	0,139	0,138	0,073	0,101
Max	0,017	0,023	0,034	0,032	0,025	0,038
Contraste	1,652	1,718	3,045	1,563	1,105	1,257
Advent1	moyenne					
Energie	0,060	0,050	0,046	0,062	0,053	0,054
Entropie	2,895	3,037	3,111	2,832	2,944	2,938
Max	0,105	0,085	0,069	0,101	0,068	0,075
Contraste	2,167	2,959	3,688	2,096	2,812	3,100
	Variance					
Energie	0,006	0,004	0,004	0,010	0,004	0,005
Entropie	0,089	0,070	0,068	0,135	0,051	0,068
Max	0,024	0,002	0,022	0,032	0,025	0,027
Contraste	0,541	0,522	0,807	0,651	1,062	0,585

TAB. 1 comparaisons des résultats sur les différents critères

La comparaison des nombres de TAB. 1 indique que le critère le plus discriminant est le contraste. Les résultats de classification des végétaux ont été améliorés mais peuvent l'être encore en prenant en compte aussi des descripteurs de formes. Faute de place cette image n'est pas présentée.

6. Conclusion

La méthode SVM non linéaire pour classes non séparables donnent les meilleurs résultats avec très peu d'échantillons de chaque classes (de 4 à 6). L'apprentissage est simple. Les méthodes proposées se sont avérées robustes par rapport aux changements d'échelle. Ces méthodes peuvent être améliorées lors de la classification des végétaux en utilisant en plus des descripteurs de texture et de forme dans les composantes des vecteurs x de données d'entrée utilisés en phases d'apprentissage et opérationnelle. Ces méthodes présentent en outre l'avantage de travailler sur des données de grande dimension.

Références

[1] C. Burges, a tutorial on support vector machines for pattern recognition, Data mining and knowledge discovery, 2, 121-167, 1998

[2] M. Chapron, L. Assemat, P. Boissard, P. Huet, Weed and corn recognition using 2D and 3D data fusion, Third ECPA, 2001.

[3] A. Cornuéjols, une nouvelle méthode d'apprentissage : les SVM. Séparateurs à vaste marge, bulletin de l'AFIA, no 51, juin 2002

[4] N. Cristianini, J. Shawe-Taylor, an introduction to support vector machines, Cambridge 2000,

Herbrich (02) : *Learning kernel classifiers*. MIT Press, 2002.
[5] R.M. Harralick et al, textural features for image classification, IEEE transactions on systems, man and cybernetics, vol 3, nb 6, p 610-621
[6] A.G. Manh, G. Rabatel, L. Assemat, M.J. Aldon, In field classification of weed leaves by machine vision using deformable templates, Third ECPA, 2001.
[7] B. Schölkoft , C. Burges, A. Smola, advances in kernel methods, MIT press, 1998
[8] B. Schölkoft, A. J. Smola, learning with kernels, MIT press,
2002
[9] Smola, Bartlett, Schölkopf & Schuurmans (00) : *Advances in large margin classifiers*. MIT Press, 2000.
[10] Vapnik *The nature of statistical learning.* Springer-Verlag, 1995.

Image 1- Image originale

Image 2- Classification sol/végétation

Image 3- Classification sol/végétation

Image 4- Image originale

Image 5- Classification sol/végétation

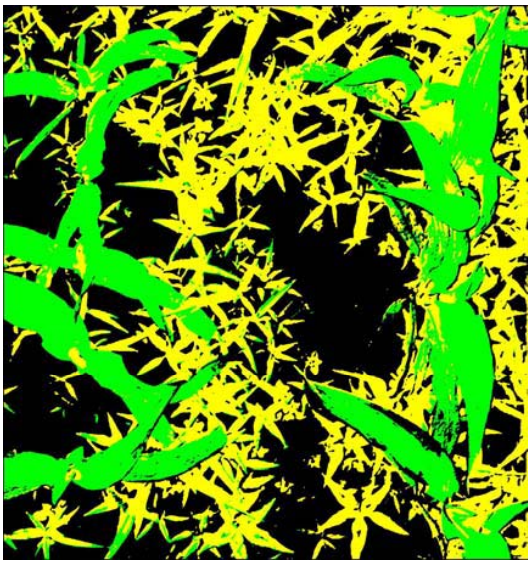


Image 6- Classification maïs/adventices

Classification avec contraintes : problématique et apprentissage d'une règle de décision par SVM

Edith Grall-Maës, Pierre Beauseroy, Abdenour Bounsiar
Institut des Sciences et Technologies de l'Information de Troyes (CNRS FRE 2732)
Équipe Modélisation et Sûreté des Systèmes
Université de Technologie de Troyes
12, rue Marie Curie - BP 2060 -10010 Troyes cedex - FRANCE
{Edith.Grall,Pierre.Beauseroy,Abdenour.Bounsiar}@utt.fr

Résumé – Le travail présenté porte sur la détermination d'une règle de décision pour un problème avec deux classes et deux contraintes qui fixent des bornes supérieures pour les probabilités d'erreur conditionnelles aux classes. Dans le cas où il existe des règles de décision satisfaisant conjointement les contraintes, la règle choisie sera celle qui minimise un coût combinant les probabilités de décision conditionnelles. Dans le cas contraire, il est nécessaire de définir une règle qui ajoute une classe de rejet. La règle optimale recherchée est alors est celle qui minimise la probabilité de rejet. Dans un premier temps, la règle de décision est définie lorsque les densités de probabilités de chacune des classes sont connues : elle consiste à comparer le rapport de vraisemblance à un ou deux seuils selon que du rejet est nécessaire ou pas. Dans un second temps, une méthode basée sur les SVM est proposée pour élaborer une règle de décision lorsque le processus est uniquement décrit par un ensemble d'apprentissage. La règle consiste à comparer la sortie du SVM avec un ou deux seuils. Ceux-ci sont déterminés à partir des sorties du SVM pour un ensemble d'apprentissage, soit directement avec les valeurs de sortie, soit à partir d'estimation des densités de probabilités. Le biais et la variance des probabilités de rejet et d'erreurs conditionnelles inhérents à la taille de l'ensemble d'apprentissage sont étudiés.

Abstract – The paper deals with the definition of a decision rule for a two classes and two constraints problem. The considered constraints bound the conditional error probabilities. When decision rules verifying both constraints exist, the rule that minimizes a cost combining the conditional decision probabilities is chosen. If such rule does not exist, it is necessary to introduce a rejection class and the optimal rule is the one that minimizes the reject probability. Firstly the rule is defined assuming that the conditional density functions are known. It consists in comparing the likelihood ratio with one or two thresholds depending on the use of rejection or not. Secondly, a method based on SVM is proposed to design decision rules for processes described by sample sets. The rule consists in comparing the function of the SVM output with thresholds. They are determined using outputs of a training set, either directly, either by estimating the output densities. The bias and the standard deviation of the reject and of the error probabilities arising from the size of the training set are studied.

1 Introduction

Les approches théoriques de la classification entre deux classes ont permis d'établir différentes règles de décision optimales au sens de certains critères, par exemple la règle de Bayes, le test de Neyman-Pearson, le minimax [1]. Elles peuvent être utilisées pour autant que le problème posé vérifie les hypothèses d'établissement de ces règles.
L'introduction de contraintes dans ces problèmes peut conduire à l'absence de solutions dans un cadre classique. Pour traiter ces situations, la notion de rejet a été développée notamment dans le cas de contraintes portant sur la probabilité d'erreur totale [2, 3, 4]. Toutefois un certain nombre de problèmes réels peuvent impliquer un ensemble de contraintes plus complexe. Les contraintes peuvent s'exprimer selon un ou plusieurs coûts qui combinent différentes probabilités d'erreur (totale ou conditionnelle). Elles peuvent impliquer des conditions d'inégalité ou d'ordre. Pour tous ces scenarii, deux cas se présentent : soit les contraintes peuvent être satisfaites en appliquant une règle de décision classique, alors cette solution est adoptée ; soit pour satisfaire les contraintes il est nécessaire d'introduire du rejet, c'est-à-dire une non décision pour certaines observations. Dans ce cas la règle vise à minimiser la probabilité de rejet. Pour qualifier ce type de règle il convient de mesurer les performances mais aussi de s'assurer du respect des contraintes. Dans les cas d'apprentissage à partir d'exemples étiquetés la qualification pose un problème particulier compte tenu de la variabilité inhérente à l'ensemble d'échantillons considéré.
Le travail proposé est une illustration de ce type de problème au cas particulier où les probabilités d'erreur conditionnelles à chacune des classes sont bornées. Le paragraphe 2 décrit précisément le problème de classification posé. Le paragraphe 3 est consacré au développement de la règle de décision lorsque les densités de probabilités sont connues. Le paragraphe 4 traite des problèmes d'estimation lorsque le processus est seulement connu à l'aide d'un ensemble de données étiquetées. Dans le paragraphe 5, une méthode est proposée pour construire un classifieur dans le cas d'un processus décrit par un ensemble d'apprentissage ; elle s'appuie sur les machines à vecteurs de support (SVM). Son application à un exemple simulé est présentée dans le paragraphe 6, qui précède la conclusion.

2 Classification avec contraintes

Soient deux classes C_1 et C_2. On considère le problème de classification binaire avec les deux contraintes suivantes :

$$\begin{cases} P(D_1/C_2) \leq e_{12} \\ P(D_2/C_1) \leq e_{21} \end{cases} \quad \text{où } e_{12} \text{ et } e_{21} \in [0,1] \tag{1}$$

avec $P(D_i/C_j)$ la probabilité de décider la classe i conditionnellement à la classe j et e_{ij} la borne supérieure pour la probabilité $P(D_i/C_j)$.

Le problème à résoudre diffère selon qu'il existe ou non au moins une partition (Z_1, Z_2) de $\mathbb{R}^n$ telle que les contraintes

soient vérifiées. S'il existe au moins un couple (Z_1, Z_2), le problème consiste à rechercher celui qui minimise un critère, qui de façon générale est un coût $\bar{c}$ qui s'exprime selon :

$$\begin{aligned}\bar{c} = \quad & c_{11}P(D_1/C_1)P_1 + c_{12}P(D_1/C_2)P_2 \\ & + c_{21}P(D_2/C_1)P_1 + c_{22}P(D_2/C_2)P_2.\end{aligned}$$

où P_i est la probabilité *a priori* de la classe i. En prenant $c_{ii} = 0$, $c_{12} = c_{21} = 1$, le critère $\bar{c}$ est égal à la probabilité d'erreur P_e.

S'il n'existe pas de partition (Z_1, Z_2) telle que les contraintes soient vérifiées, il est nécessaire, pour satisfaire les contraintes, d'introduire une zone de rejet, dans laquelle aucune décision sur les réalisations n'est prise. Le problème consiste alors à rechercher la partition (Z_1, Z_2, Z_R) qui minimise la probabilité de rejet P_R.

La qualification de la règle de décision nécessite d'estimer des grandeurs caractéristiques (moyenne, écart-type) du critère minimisé ($\bar{c}$ ou P_R selon le cas) et également des probabilités intervenant dans la définition des contraintes.

3 Élaboration d'une règle de décision à partir des densités de probabilités

Lorsque les densités de probabilités conditionnelles à chacune des classes $P(x/C_1)$ et $P(x/C_2)$ et les probabilités *a priori* P_1 et P_2 sont connues, la règle de décision théorique peut être élaborée.

Tout d'abord, il est nécessaire de déterminer si les contraintes peuvent être conjointement satisfaites. À cette fin, il est possible de considérer chaque contrainte indépendamment puis d'analyser la compatibilité des solutions obtenues.

Soit le domaine Z_1^* tel que la contrainte $P(D_1/C_2) = e_{12}$ soit vérifiée et le coût $\bar{c}$ soit minimum. Pour déterminer la solution à ce problème, la démarche est similaire à celle pour le test de Neymann-Pearson. Z_1^* est défini à partir d'un seuil λ_1^* sur le rapport de vraisemblance $l(x)$ défini par :

$$l(x) = \frac{P(x/C_1)}{P(x/C_2)}. \tag{2}$$

Z_1^* est donné par :

$$Z_1^* = \{x | l(x) \geq \lambda_1^*\} \text{ avec } \int_{Z_1^*} P(x/C_2)dx = e_{12}. \tag{3}$$

Par symétrie, on définit le domaine Z_2^* qui satisfait la contrainte $P(D_2/C_1) = e_{21}$ et qui minimise $\bar{c}$. On montre que Z_2^* est défini à partir d'un seuil λ_2^* sur le rapport de vraisemblance $l(x)$ selon :

$$Z_2^* = \{x | l(x) \leq \lambda_2^*\} \text{ avec } \int_{Z_2^*} P(x/C_1)dx = e_{21}. \tag{4}$$

Dans le cas où $\lambda_1^* \leq \lambda_2^*$, l'intersection entre les domaines Z_1^* et Z_2^* est non nulle, et il est donc possible de trouver une loi qui satisfasse les deux contraintes sans introduire de rejet. Toute règle de décision qui engendre une partition à partir d'un seuil $\lambda \in [\lambda_1^*, \lambda_2^*]$ sur le rapport de vraisemblance vérifie les contraintes. La règle de décision qui minimise le coût $\bar{c}$ et vérifie les contraintes est :

$$l(x) \underset{D_2}{\overset{D_1}{\gtrless}} \lambda \quad \text{avec } \lambda = \begin{cases} \lambda_0 \text{ si } \lambda_1^* \leq \lambda_0 \leq \lambda_2^* \\ \lambda_1^* \text{ si } \lambda_0 \leq \lambda_1^* \leq \lambda_2^* \\ \lambda_2^* \text{ si } \lambda_1^* \leq \lambda_2^* \leq \lambda_0. \end{cases} \tag{5}$$

où λ_0 est le seuil sur le rapport de vraisemblance qui minimise $\bar{c}$ sans contraintes :

$$\lambda_0 = \frac{P_2(c_{12} - c_{22})}{P_1(c_{21} - c_{11})}. \tag{6}$$

Dans le cas où $\lambda_1^* > \lambda_2^*$, il est nécessaire de faire appel à du rejet pour satisfaire les deux contraintes. La probabilité de rejet minimum est obtenue lorsque les contraintes sont saturées. La règle de décision est donc :

$$l(x) \quad \begin{cases} \geq \lambda_1^* \text{ alors } D_1 \\ \leq \lambda_2^* \text{ alors } D_2 \\ \notin [\lambda_2^*; \lambda_1^*] \text{ alors } D_R. \end{cases} \tag{7}$$

4 Performances d'une règle de décision élaborée à partir d'échantillons

Généralement les probabilités théoriques ne sont pas connues et le processus est uniquement décrit par un ensemble d'apprentissage. Le classifieur doit alors être élaboré à partir de cet ensemble d'échantillons étiquetés. Une famille d'approches pour construire la règle de décision consiste à déterminer une fonction représentative du rapport de vraisemblance et à déterminer les seuils à appliquer sur la fonction. Les performances du classifieur alors sont liées aux erreurs d'estimation de ces deux grandeurs.

Il est assez aisé de mesurer l'effet de la taille de l'ensemble d'apprentissage sur l'erreur d'estimation des seuils. Considérons que le rapport de vraisemblance est parfaitement connu mais pas les densités conditionnelles. la détermination des seuils λ_1^* et λ_2^* se fait sans biais mais avec une certaine variance dépendant du nombre d'échantillons. En effet, considérons une loi donnée et un domaine D, avec p la probabilité théorique qu'un échantillon appartienne au domaine D et N le nombre d'échantillons tirés de cette loi. Soit la variable aléatoire P_N mesurant la proportion d'échantillons dans le domaine D ; sa loi de probabilité est donnée par :

$$P\left[P_N = \frac{k}{N}\right] = C_N^k p^k (1-p)^{N-k}. \tag{8}$$

On peut montrer que la moyenne et l'écart-type de la la variable aléatoire P_N sont :

$$E[P_N] = p \quad \text{et} \quad \sigma[P_N] = \sqrt{\frac{p(1-p)}{N}}. \tag{9}$$

Par conséquent, même si l'on est capable de disposer une estimation parfaite du rapport de vraisemblance, l'élaboration d'un classifieur à partir d'un ensemble de N échantillons permettra d'obtenir un classifieur sans biais mais avec une variance inversement proportionnelle au nombre d'échantillons.

5 Règle de décision basée sur les SVM

Lorsque le processus est uniquement décrit par un ensemble d'apprentissage, le rapport de vraisemblance n'est pas connu et difficilement estimable avec précision. Il est alors nécessaire de construire une règle de décision à l'aide d'une autre méthode. Les SVM ont été utilisés en raison de leurs aptitudes à trouver des frontières non linéaires et leur robustesse vis à vis de l'ensemble d'apprentissage.

Considérons un ensemble de vecteurs d'apprentissage appartenant à deux classes :

$$\mathcal{A} = \{(x_1, y_1), ...(x_l, y_l)\} \quad x_i \in \mathbb{R}^n, y_i \in \{-1, 1\} \quad (10)$$

Le principe des vecteurs à support machine (SVM) [5, 6] repose sur la transformation d'un vecteur x dans un espace de grande dimension et la construction d'un hyperplan optimal dans cet espace. Le produit scalaire dans l'espace transformé est effectué à l'aide d'un noyau K défini par un vecteur de paramètres θ. Le noyau gaussien est couramment utilisé ; il est défini par :

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\rho^2}\right). \quad (11)$$

La fonction de décision du SVM est la solution de la minimisation de :

$$\frac{1}{2}\|w\|^2 + C\sum_i \xi_i \quad (12)$$

où C est une constante de régularisation qui détermine le compromis entre l'erreur empirique et le terme de complexité, et w est le vecteur normal à l'hyperplan séparateur dans l'espace transformé.

Pour le problème avec contraintes, par analogie avec la règle obtenue sur le rapport de vraisemblance, il est possible d'élaborer une règle de décision en exploitant la dynamique de la sortie du SVM :

$$\widetilde{f}_\theta(x) = \sum_{i=1}^{l} a_i y_i K_\theta(x_i, x) + b \quad (13)$$

où les a_i sont les multiplicateurs de Lagrange du problème dual de (12).

La règle est alors définie comme suit :

$$\widetilde{f}_\theta(x) \quad \begin{cases} \geq \xi_1^* \text{ alors on décide } D_1 \\ \leq \xi_2^* \text{ alors on décide } D_2 \\ \notin [\xi_2^*; \xi_1^*] \text{ alors on décide } D_R \end{cases} \quad (14)$$

L'estimation des seuils ξ_1^* et ξ_2^* doit être faite en utilisant un ensemble d'apprentissage $\mathcal{A}'$ indépendant de l'ensemble $\mathcal{A}$ utilisé pour construire la fonction de décision du SVM.

Notons $z_j^{(i)}$ (où (i) indique la classe et j indique l'échantillon) les valeurs des $\widetilde{f}_\theta(x)$ de l'ensemble d'apprentissage $\mathcal{A}'$. Pour déterminer les seuils, le procédé le plus simple est d'estimer l'erreur empirique pour différents seuils et de sélectionner les seuils permettant de respecter les contraintes :

$$\frac{1}{N_2}\sum_{i \in C_2} I_1(i) = e_{12} \qquad \tfrac{1}{N_1}\textstyle\sum_{i \in C_1} I_1(i) = e_{21}$$

$$I_1(i) = \begin{cases} 1 \text{ si } \widetilde{f}_\theta(z_j^{(2)}) > \xi_1^* \\ 0 \text{ sinon} \end{cases} \qquad I_2(i) = \begin{cases} 1 \text{ si } \widetilde{f}_\theta(z_j^{(1)}) < \xi_2^* \\ 0 \text{ sinon.} \end{cases}$$

Une autre approche consiste à utiliser les valeurs $z_j^{(i)}$ pour estimer les densités de probabilité conditionnelles aux classes. L'estimation pour chacune des classes peut être obtenue à l'aide de l'estimateur de Parzen :

$$\hat{p}(z^{(i)}) = N_i^{\alpha-1}\frac{(2\pi)^{1/2}}{\widehat{\sigma}} \sum_{j=1}^{N_i} \exp\left(-\frac{1}{2}N_i^{2\alpha}\frac{(z^{(i)} - z_j^{(i)})^2}{\widehat{\sigma}}\right)$$

où $\widehat{\sigma}$ est une estimation de l'écart-type des valeurs, N_i est le nombre d'échantillons dans la classe considérée, et α est un coefficient de lissage.

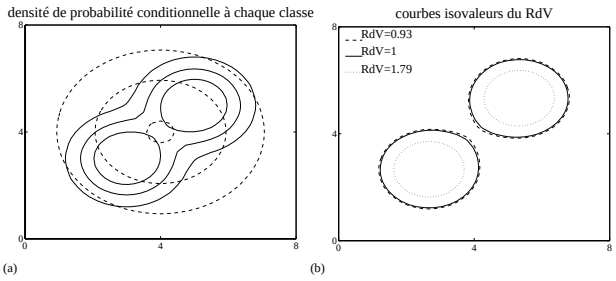


FIG. 1 – courbes isovaleurs de (a) la densité de probabilité conditionnelle à chacune des classes (b)du rapport de vraisemblance

6 Application à un problème simulé

6.1 Description du problème simulé

On recherche une règle de décision qui satisfasse les deux contraintes présentées ci-dessus avec $e_{12} = e_{21} = 0.2$ pour le processus observé dans $\mathbb{R}^2$ décrit ci-après. Les classes 1 et 2 sont équiprobables et respectivement composées du mélange de deux distributions normales, et d'une distribution normale seule. Elles sont représentées par des courbes isovaleurs sur la figure 1a.

Différentes probabilités théoriques ont été calculés numériquement. Pour un rapport de vraisemblance égal à 1, la probabilité d'erreur totale est minimum et égale à 0.312, $P(D_2/C_1) = 0.228$ et $P(D_1/C_2) = 0.396$. Pour les valeurs de e_{12} et e_{21} choisies les seuils obtenus sont $\lambda_2^* = 0.93$ et $\lambda_1^* = 1.79$. Les contraintes ne peuvent donc être satisfaites qu'en introduisant du rejet. Les courbes isovaleurs pour ces valeurs de rapport de vraisemblance (RdV) sont représentées sur la figure 1b.

L'expérimentation a été effectuée avec 50 ensembles d'apprentissage comptant chacun par classe 100 échantillons pour la construction du classifieur et 100 échantillons pour la détermination du seuil. Pour la règle de décision optimale donnée par (7), les moments théoriques et estimés à partir des 50 ensembles sont :

$$\begin{array}{ll} E[P(D_1/C_2)] = 0.2, & \widehat{E}[\widehat{P}(D_1/C_2)] = 0.205 \\ \sigma[P(D_1/C_2)] = 0.04, & \widehat{\sigma}[\widehat{P}(D_1/C_2)] = 0.036, \\ E[P(D_2/C_1)] = 0.2, & \widehat{E}[\widehat{P}(D_2/C_1)] = 0.204 \\ \sigma[P(D_2/C_1)] = 0.04, & \widehat{\sigma}[\widehat{P}(D_2/C_1)] = 0.038, \\ E[P_R] = 0.248, & \widehat{E}[\widehat{P}_R] = 0.247 \\ \sigma[P_R] = 0.030, & \widehat{\sigma}[\widehat{P}_R] = 0.031 \end{array}$$

Ces valeurs mettent en évidence qu'avec 100 échantillons dans chaque classe, le problème peut être résolu au mieux avec un écart-type de l'ordre de 20% ($\sigma = 0.04$) de l'erreur de consigne ($e_{12} = 0.2$).

6.2 Performances du classifieur

Le classifieur a été construit en prenant un noyau gaussien défini par (11). Les paramètres ont tout d'abord été optimisés en utilisant les méthodes proposées dans [7] et [8]. La figure 2 montre les courbes isovaleurs de la sortie du SVM obtenu à partir d'un ensemble d'apprentissage pour les deux seuils correspondant aux contraintes d'erreur e_{12} et e_{21}.

La figure 3 montre les densités de probabilité conditionnelles des sorties du SVM pour un ensemble d'échantillons donné,

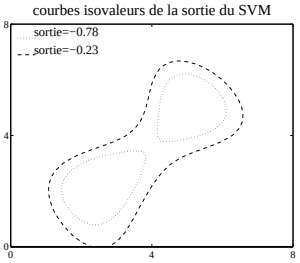


FIG. 2 – courbes isovaleurs de la sortie du SVM

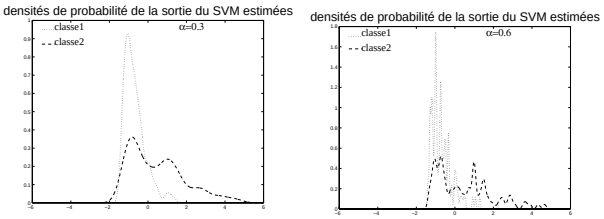


FIG. 3 – densités de probabilité conditionnelles des sorties du SVM pour $\alpha = 0,3$ et $\alpha = 0,6$

dans le cas de $\alpha = 0.3$ et $\alpha = 0.6$.

Pour chaque ensemble d'apprentissage, le SVM est construit et une règle de décision est définie selon (14). Les performances de chaque règle de décision sont estimées en mesurant les grandeurs $\widehat{P}(D_1/C_2)$, $\widehat{P}(D_1/C_2)$ et $\widehat{P_R}$. Pour cette estimation, plutôt que d'utiliser un ensemble test, nous avons utilisé les densités de probabilité théoriques de chacune des deux classes : le plan a été fragmenté en régions égales suivant une grille, la probabilité conditionnelle associée à chaque région a été calculée à partir des densités théoriques, puis la décision pour chaque région a été déterminée et pondérée par les probabilités conditionnelles pour obtenir les grandeurs estimées. L'opération répétée pour les 50 ensembles d'apprentissage a permis d'estimer la moyenne et l'écart-type des grandeurs mesurées. Les résultats sont représentés sur la figure 4.

Cette figure montre que la valeur de α doit être judicieusement choisie : moins le lissage est important, plus l'écart-type est grand, mais un lissage trop important déforme la densité de probabilité. L'estimateur de Parzen introduit un produit de convolution qui a pour effet d'étaler la densité de probabilité. Pour un seuil donné, il existe donc un biais entre la probabilité théorique et la probabilité obtenue par estimateur de Parzen.

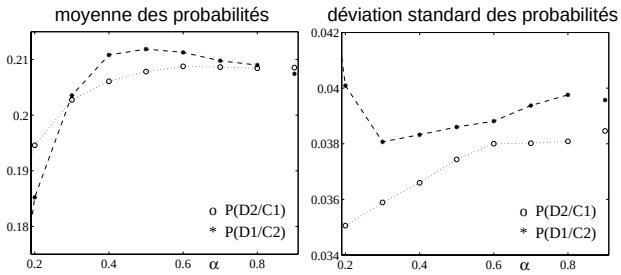


FIG. 4 – Moyenne et écart-type des probabilités $P(D_1/C_2)$ et $P(D_2/C_1)$. L'axe des abscisses indique la valeur de utilisée pour estimer les densités, sauf pour la dernière valeur, qui correspond à α très grand.

En prenant $\alpha = 0.4$ la valeur de l'écart-type est proche de la valeur théorique.

7 Conclusion

Le problème porte sur la classification avec contraintes. Dans le cas particulier de deux classes et où deux contraintes sont à satisfaire (probabilités d'erreur conditionnelles à chacune des classes bornées), la règle de décision théorique consiste à comparer le rapport de vraisemblance à un ou deux seuils selon que les contraintes peuvent être satisfaites sans ou avec rejet.

Une solution pour l'élaboration d'une règle de décision pour ce problème lorsque le processus est uniquement décrit à l'aide d'une base d'exemples est proposée. Elle est basée sur l'utilisation des SVM. Les densités de probabilité conditionnelles des sorties du SVM sont estimées à partir de la base d'apprentissage, à l'aide d'un estimateur de Parzen. Les résultats expérimentaux montrent qu'un lissage adéquat des densités estimées permet de limiter la variance des probabilités d'erreur conditionnelles intervenant dans les contraintes. Ce principe permet d'améliorer la robustesse de la règle de décision apprise.

Le problème de classification traité dans ce papier constitue un cas particulier. Les méthodes peuvent être généralisées à d'autres types de contraintes et à plus de deux classes.

Références

[1] K. Fukunaga. *Introduction to statistical pattern recognition*. Academic Press, Boston, 1990.

[2] C.K. Chow, *On Optimum Reject Error and Reject Tradeoff*, IEEE Transactions on Information Theory, Vol. IT-16, N°.1, pp. 41-46, January 1970.

[3] B. Dubuisson and M. Masson, *A statistical decision rule with incomplete knowledge about classes*, Pattern recognition, Vol. 26, N°.1, pp. 155-165, 1993.

[4] G. Fukmera, F. Roli and G. Giacinto, *Reject option with multiple thresholds*, Pattern recognition, Vol. 33, N°.12, pp. 2099-2101, 2000.

[5] C. Cortes and V. Vapnik. *Support Vector Networks*. Machine Learning, 20 :273-279, 1995.

[6] C. J. C. Burges. *A tutorial on support vector machines for pattern recognition.* Data Mining and Knowledge Discovery, 2(2) :121-167, 1998.

[7] T. Hastie, S. Rosset, R. Tibshirani and J. Zhu *The Entire Regularization Path for the Support Vector Machine.*, Journal of Machine Learning Research, 5 :131-159, 2004.

[8] O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukhrjee. *Choosing kernel parameters for support vector machines.* Machine Learning, 46 :131 :159, 2002.

Guidage du filtrage particulaire par une étude du mouvement

Elodie BICHOT, Laurent MASCARILLA, Pierre COURTELLEMONT

Laboratoire d'Informatique – Image – Interaction
Université de La Rochelle
Av. Michel Crépeau, 17042 La Rochelle Cedex, FRANCE
elodie.bichot@univ-lr.fr, laurent.mascarilla@univ-lr.fr
pierre.courtellemont@univ-lr.fr

Résumé – Nous proposons d'exploiter le résultat d'une segmentation par étude du mouvement afin d'améliorer le suivi par filtrage particulaire. Nous générons d'abord des hypothèses du filtrage particulaire qui correspondent aux régions dont le mouvement est similaire au déplacement de la cible à l'instant précédent. Ensuite, après avoir identifié la région de mouvement correspondant à la cible, nous apprenons son modèle qui est intégré comme référence lors de la prochaine itération du suivi. Cette approche améliore les performances du suivi dans des situations complexes en comparaison avec un filtre Bootstrap conventionnel comme le montrent les expérimentations menées sur des séquences d'images d'aquarium.

Abstract – In this article, we propose to use results of a motion based segmentation to improve tracking by particle filtering. We generate firstly particle filtering hypotheses corresponding to regions of similar motion to target displacement at previous time step. Then, after identifying target motion region, target model is updated and integrated as reference for the next tracking iteration. This approach improves performances in complex situations compared with Bootstrap filter as shown our experiments on fish tank sequences.

1 Introduction

Dans le cadre d'une application automatique de reconnaissance de poissons appelée projet Aqu@thèque [5], nous devons suivre des poissons en mouvement jusqu'à ce que le système de reconnaissance puisse les identifier. Notre problème consiste donc à suivre des objets non rigides et animés de mouvements complexes en milieu perturbé. Pour ce faire, nous proposons d'exploiter le résultat d'une segmentation par analyse du mouvement en propageant les particules du filtre particulaire vers les régions de mouvement similaire à celui de la cible et en apprenant un modèle qui est ensuite intégré comme référence lors de la prochaine itération du suivi. Les résultats obtenus montrent que le suivi est maintenu dans des situations complexes d'occultations et de variations d'illumination en comparaison avec un filtre Bootstrap conventionnel. Nous présentons l'approche proposée dans la partie 2 puis analysons les résultats obtenus dans la partie 3 et finalement dressons une conclusion dans le dernier paragraphe.

2 L'approche proposée

2.1 Vue générale de l'approche de suivi proposée

Le filtrage particulaire consiste en une exploration de l'espace d'état de la cible par des particules pondérées. Les particules représentent des hypothèses sur l'état de la cible et les poids associés les probabilités des hypothèses. Dans le filtre Bootstrap, elles sont propagées suivant le modèle dynamique de la cible, pondérées selon la vraisemblance des observations conditionnellement à l'état prédit et finalement, sélectionnées proportionnellement à leur poids. La somme des fonctions de Dirac pondérées par les poids des particules et centrées en les particules constitue une approximation de la distribution de probabilté *a posteriori* sur l'état de la cible [3]. Si la prédiction est précise et si le modèle de représentation permet de discriminer la cible de son environnement, alors cette méthode donne des résultats satisfaisants. Cependant, dans le cas de scènes réelles où il est difficile de déterminer un modèle dynamique précis et où la représentation de la cible varie en fonction de son orientation et de l'illumination, elle peut être mise en échec. Ceci justifie notre approche : nous utilisons le résultat de la segmentation pour générer des particules du filtre particulaire correspondant aux régions de mouvement similaire au déplacement de la cible et améliorer ainsi la prédiction. Ensuite, nous apprenons, à partir de la région de mouvement identifiée comme étant la cible, un modèle précis et représentatif de la cible qui est intégré comme référence dans l'étape de correction à la prochaine itération du suivi. Le schéma proposé est représenté sur la figure 1.

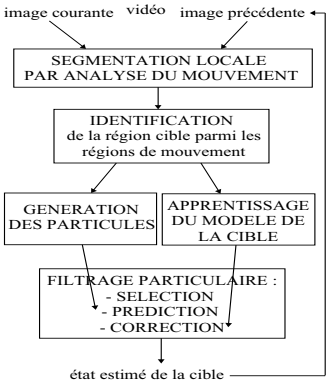


FIG. 1 – Schéma de suivi proposé.

2.2 Segmentation par analyse du mouvement

Afin de segmenter l'image au sens du mouvement, le flot optique a d'abord été estimé par une procédure de mise en correspondance sur des images couleurs. Ensuite, nous avons appliqué l'approche de classification relationnelle des C-médoïds flous sur le champ des vecteurs de mouvement et pour cela introduit une mesure de similarité entre vecteurs géométriques fondée sur l'égalité de deux vecteurs [1]. Nous obtenons une liste de $Nb_regions_mvt$ régions homogènes en mouvement :

$$Region_mvt_t^{(k)} = \left\{ BB_t^{(k)}, M_t^{(k)}, v_t^{(k)} \right\}_{k=1}^{Nb_regions_mvt}$$

- $BB_t^{(k)}$ est le rectangle optimal englobant la région k,
- $M_t^{(k)}$ désigne le masque des pixels appartenant à la forme pour la région k,
- $v_t^{(k)}$ est le vecteur déplacement de la région k.

Parmi ces régions, nous retenons, pour l'étape d'échantillonnage d'importance du filtrage particulaire, celles dont le mouvement maximise la similarité avec le déplacement de la cible à l'instant $t-1$ par

$$s = \arg\max_k R(v_cible_{t-1}, v_t^{(k)}) \quad (1)$$

où v_cible_{t-1} est le déplacement de la cible à l'instant $t-1$ et $R(.,.)$ la mesure de similarité entre deux vecteurs mouvements introduite dans [1].

2.3 Filtrage particulaire fondé sur le mouvement et la couleur

Nous renforçons un filtre Bootstrap conventionnel par la conduite des particules vers les régions d'intérêt de l'espace d'état et l'apprentissage du modèle de référence de la cible à partir du résultat de la segmentation. Notre approche suppose que d'une image à l'autre le mouvement et l'apparence de la cible n'évoluent que très peu.

2.3.1 Espace d'état

L'espace d'état est défini par le rectangle englobant l'objet d'intérêt. $X_t = \{x_t, y_t, largeur_t, hauteur_t\}$ correspond à une particule, c'est-à-dire à une hypothèse de rectangle englobant. Nous désignons par

$$\left\{ X_t^{(i)}, w_t^{(i)} \right\}_{i=1}^{Nb_total} \quad (2)$$

la liste des Nb_total particules $X_t^{(i)}$ et leur poids associé $w_t^{(i)}$.

2.3.2 Modèle dynamique

Le modèle dynamique s'appuie sur les positions et vitesses estimées de la cible dans les images précédentes, la vitesse étant mise à jour à partir des positions précédentes de la cible.

$$x_t^{(i)} = x_{t-1}^{(i)} + \dot{x}_{t-1}^{(i)} + \sigma_{dyn}\omega \quad (3)$$

$$\dot{x}_t = \bar{x}_{t-1} - \bar{x}_{t-2} \quad (4)$$

$$largeur_t^{(i)} = largeur_{t-1}^{(i)} \pm 10\% largeur_{t-1}^{(i)} \quad (5)$$

où σ_{dyn} est l'écart-type du bruit de dynamique et ω un bruit blanc Gaussien de moyenne nulle. $\bar{x}_t$ est la moyenne des positions $x_t^{(i)}$ à l'instant t. Les composantes y et $\dot{y}$ suivent respectivement les mêmes lois que x et $\dot{x}$. La $hauteur$ évolue de même que la $largeur$ mais indépendamment de celle-ci. Une partie des particules est propagée suivant ce modèle.

2.3.3 Echantillonnage d'importance

L'autre partie des particules est générée selon le principe d'échantillonnage d'importance. Il s'agit de diriger la recherche en utilisant une connaissance supplémentaire décrivant les parties de l'espace d'état qui contiennent le plus d'informations sur la distribution de probabilité *a posteriori* de la cible. Dans notre cas, nous utilisons le résultat de la segmentation par étude du mouvement. Plus précisément, nous générons les particules à partir des régions de mouvement cohérent avec le déplacement de la cible (et retenues après l'étape de segmentation par l'équation (1)) et d'une hypothèse de bruit blanc Gaussien pour les positions et de bruit uniforme pour les largeur et hauteur (équation 5). Nous limitons ainsi la recherche aux régions de forte vraisemblance de l'espace d'état et améliorons la prédiction. Toutes les particules, qu'elles soient propagées par le modèle dynamique ou générées selon l'échantillonnage d'importance, sont pondérées suivant le modèle de mesures que nous décrivons maintenant.

2.3.4 Modèle de mesures

Le modèle d'observation s'appuie sur la couleur dans l'espace rgb normalisé. Le choix de cet espace se justifie par son invariance à l'intensité d'illumination. L'objectif de l'étape de correction est de favoriser les histogrammes candidats similaire à l'histogramme modèle de la cible. Nous définissons alors le modèle de la vraisemblance couleur par

$$w_t^{(i)} \propto \exp(-D^2(h_{X_t^{(i)}}, h_{cible_{t-1}})/2\sigma_{obs}^2) \quad (6)$$

- $h_{X_t^{(i)}}$ est l'histogramme calculé sur le rectangle donné par l'hypothèse $X_t^{(i)}$,
- $h_{cible_{t-1}}$ est l'histogramme de la cible à l'instant $t-1$,
- $D(h1, h2) = (1 - \sum_{j=1}^{b} \sqrt{h1(j)h2(j)})^{1/2}$ est la distance de Bhattacharyya entre les histogrammes $h1$ et $h2$,
- b désigne le nombre de classes de l'histogramme,
- σ_{obs} est l'écart-type du bruit d'observation.

Une fois pondérées, les particules sont rééchantillonnées proportionnellement à leur poids. L'état de la cible est, finalement, estimé par la moyenne des particules

$$\hat{X}_t = \frac{1}{Nb_total} \sum_{i=1}^{Nb_total} X_t^{(i)} \quad (7)$$

2.3.5 Mise à jour du mouvement et de l'apparence de la cible

Afin de préparer l'étape suivante de sélection des régions de mouvement d'intérêt, nous mettons à jour le déplacement de la cible selon l'équation

$$v_cible_t = v_t^{(s)} \quad (8)$$

où s désigne les régions retenues par l'équation (1). Puis, nous recherchons parmi les régions de mouvement la région dont

l'apparence maximise la similarité avec celle de la cible à l'instant $t-1$:

$$u = \arg\min_k D(h_{cible_{t-1}}, h_{M_t}^{(k)}) \quad (9)$$

où D désigne la distance de Bhattacharyya et $h_{M_t}^{(k)}$ l'histogramme construit à partir du masque de la région k. Nous mettons alors à jour l'histogramme couleur de la cible à partir du masque de la forme identifiée ci-dessus :

$$h_{cible_t} = h_{M_t}^{(u)} \quad (10)$$

En construisant l'histogramme à partir du masque de la forme, nous n'intégrons pas de faux pixels (pixels appartenant au fond de l'image ou à d'autres objets) et bénéficions ainsi d'un modèle représentatif de la cible pour l'étape suivante.

3 Expérimentations

Nous appliquons notre approche et un filtre Bootstrap conventionnel sur des séquences d'aquarium. L'initialisation du suivi s'effectue manuellement selon le projet Aqu@thèque [5] à partir des images 219 pour la première cible et 163 pour la seconde. L'écart-type du bruit de dynamique σ_{dyn} est de 5 pixels et celui du bruit σ_{obs} du modèle d'observation fondé sur la distance de Bhattacharyya est égal à 0.1. Dans le filtre Bootstrap, 150 particules sont utilisées. Dans notre approche, 100 particules sont propagées suivant le modèle dynamique et 50 générées à partir de la carte de segmentation. Différentes vidéos résultats sont disponibles à l'adresse *http ://l3iexp.univ-lr.fr/magnoliatrico/aquathequedemo/prb_scientifique/suivi2.html*.

Pour la première séquence représentée en figure 3, nous nous intéressons au poisson porc encadré dans l'image 223. Ce poisson est occulté par un chirurgien bleu de l'image 243 à l'image 265. Au moment de l'occultation se trouve dans son voisinage un autre poisson porc de même espèce donc présentant une distribution couleur similaire. Le suivi effectué par le filtre Bootstrap conventionnel dérive du poisson porc cible, au moment de l'occultation par le chirurgien bleu, pour se concentrer sur l'autre poisson porc. La cible est définitivement perdue comme le montrent les images de la première ligne figure 3. En revanche, avec notre approche, les particules générées à partir de la segmentation et représentées par les hypothèses de rectangles dessinés en blanc (dernière ligne figure 3) dirigent la recherche vers la région de mouvement cohérent à l'évolution de la cible. Et, le suivi du poisson d'intérêt est maintenu pendant l'occultation. De plus, son état est correctement estimé tout au long de la séquence. L'information de couleur ne discrimine pas les deux poissons de la même espèce comme l'illustrent les deux modes de l'image de la vraisemblance couleur de la figure 2. Par ailleurs, une segmentation par détection du mouvement comme dans [2] ou [4] ne nous permet pas de suivre le poisson d'intérêt car elle ne lève pas l'ambiguïté entre les régions de mouvement correspondant aux poissons porcs qui se déplacent (image de droite de la figure 2). En revanche, l'étude du mouvement offre la possibilité de distinguer ces trois régions car elles sont animées de mouvements différents (seconde ligne figure 3).

L'objectif de la seconde expérimentation (figure 4) est de suivre le chirurgien bleu qui évolue d'abord rapidement puis se retourne et finalement se déplace plus lentement dans une zone faiblement illuminée (empêchant en outre sa détection par différence d'images (image de droite figure 2)). Avec un filtre Bootstrap conventionnel, l'état de la cible n'est pas toujours estimé correctement comme sur les images 208 et 264 même si elle n'est jamais définitivement perdue. En revanche, avec notre approche, les positions et tailles sont évaluées de manière satisfaisante quelles que soient les conditions d'illumination grâce à la mise à jour du modèle de référence. La trajectoire estimée par notre approche, de par sa régularité, illustre également l'amélioration apportée en comparaison avec celle donnée par le filtre Bootstrap.

FIG. 2 – De gauche à droite : image 251 de la séquence Aqu@thèque analysée avec le poisson d'intérêt encadré en blanc. Vraisemblance couleur représentée sur une grille sous échantillonnée en fonction de la position : les régions ayant une forte vraisemblance de correspondre à la cible sont représentées en blanc. Détection du mouvement par différence d'images avec une image de référence [5].

4 Conclusion

Notre contribution consiste à guider le suivi en exploitant le résultat d'une segmentation par analyse du mouvement en deux temps : nous générons des particules dans les régions de forte vraisemblance de l'espace d'état et nous mettons à jour le modèle d'apparence de la cible afin qu'il reflète les variations liées à l'environnement. L'approche proposée améliore le suivi en présence d'occultations et de variations d'illumination comparé à un filtre Bootstrap conventionnel. Dans le contexte du projet Aqu@thèque et en cas d'échec de la reconnaisssance, nous envisageons de suivre le poisson d'intérêt et de ne lancer l'identification que lorsque l'orientation et l'illumination du poisson sont favorables.

Références

[1] E. Bichot, L. Mascarilla and P. Courtellemont. Motion based Segmentation to improve Tracking of Non Rigid Objects, *Proceedings of the 3rd IEEE Conference on Sciences of Electronic, Technologies of Information and Telecommunications*, Tunisie, 2005.

[2] D. J. Bullock and J. S. Zelek. Real-Time Tracking for Visual Interface Applications in Cluttered and Occluding Situations, *Image and Vision Computing 22*, pages 1083-1091, 2004.

[3] C. Hue. Méthodes séquentielles de Monte-Carlo pour le filtrage non linéaire multi-objets dans un environnement bruité. Applications au pistage multi-cibles et à la trajectographie d'entités dans des séquences d'images 2D, *Thèse*, 2003.

[4] P. Pérez, J. Vermaak and A. Blake. Data Fusion for Visual Tracking With Particles, *Proceedings of the IEEE*, vol. 92, no 3, 2004.

[5] D. Semani, T. Bouwmans, C. Frélicot and P. Courtellemont. Automatic Fish Recognition in Interactive Live Videos, *Int. Workshop on Interactive Video between Research challenges and Industrial Applications (IVRCIA)*, XIV :94-99, 2002.

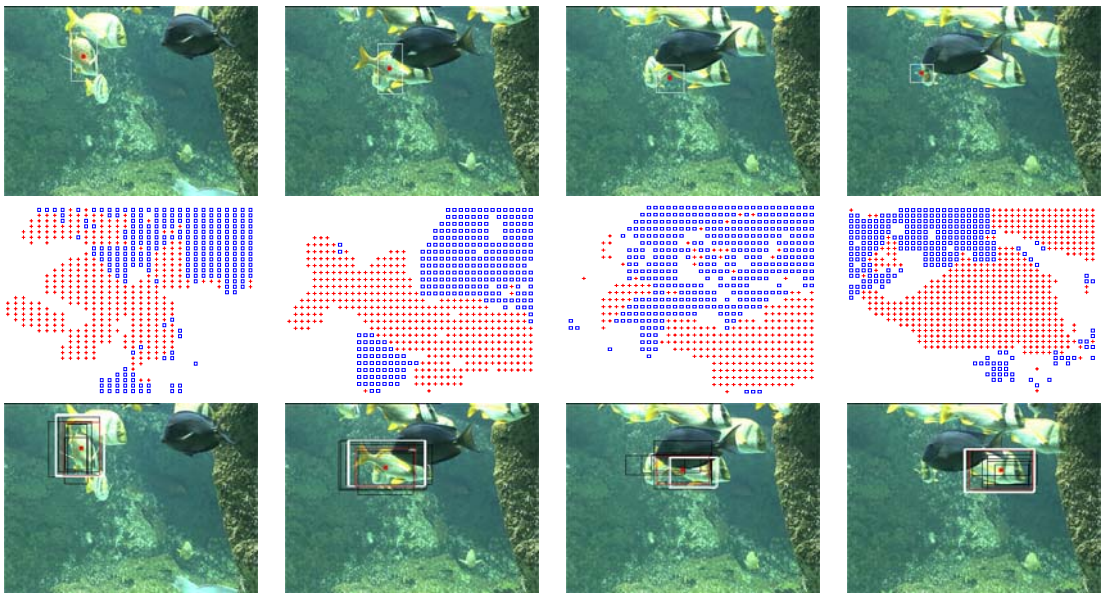


FIG. 3 – Images 223, 243, 251, 258 d'une séquence Aqu@thèque. Première ligne : résultats du suivi avec un filtre Bootstrap conventionnel. Le rectangle blanc représente l'état estimé de la cible et le carré rouge la position estimée. Deuxième ligne : carte de segmentation par étude du mouvement. La cible est représentée par des croix rouges et les autres objets en mouvement par des carrés bleus. Dernière ligne : résultat de notre schéma de suivi. En blanc : rectangles correspondant aux hypothèses générées à partir de la carte de segmentation. En noir : rectangles correspondant aux hypothèses propagées suivant le modèle dynamique. En rouge : rectangle correspondant à l'état estimé de la cible. Les particules générées à partir de la carte de segmentation dirigent la recherche vers la région cible (image 251). Le suivi est maintenu et l'état de la cible correctement estimé. Pour faciliter l'interprétation, nous n'avons représenté que quelques hypothèses sélectionnées aléatoirement.

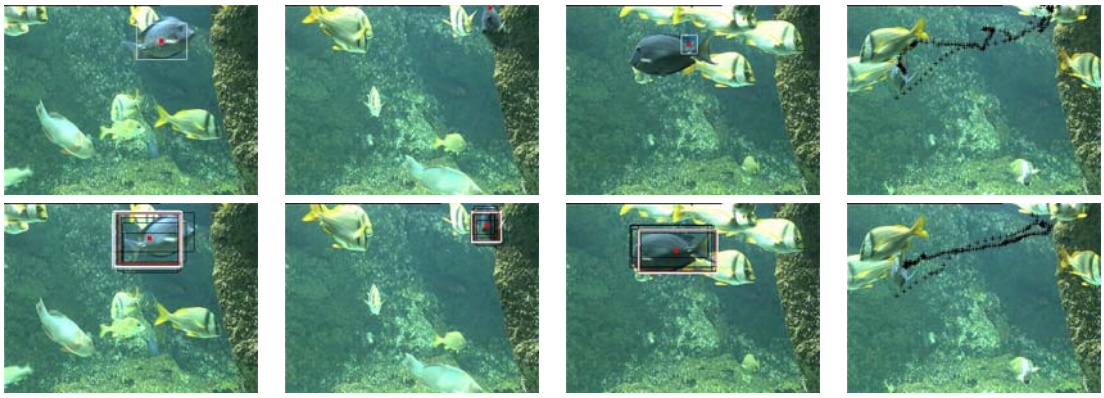


FIG. 4 – Images 180, 208, 264, 306. Première ligne : résultats du suivi avec un filtre Bootstrap conventionnel. Dernière ligne : résultat de notre schéma de suivi. A droite : trajectoires du chirurgien bleu estimées avec l'approche Bootstrap (en haut) et notre approche (en bas).

Recalage variationnel non rigide statistique référencé région

Nicolas ROUGON, Antoine DISCHER, Françoise PRÊTEUX

GET / INT - Unité de Projets ARTEMIS
9, Rue Charles Fourier - 91011 Evry Cedex, France
Nicolas.Rougon@int-evry.fr, Antoine.Discher@int-evry.fr, Francoise.Preteux@int-evry.fr

Résumé – Les mesures de similarité statistiques classiquement utilisées en recalage non rigide iconique exploitent des statistiques de luminance globales n'intégrant aucune information géométrique. Cette absence de prise en compte du contexte spatial peut conduire dans certains cas à un appariement imprécis voire incohérent de structures géométriques homologues, rendant ces critères mal adaptés à la segmentation par propagation d'atlas. Nous proposons de résoudre cette limitation en injectant dans la mesure de similarité un a priori spatial sous forme d'un modèle étiqueté de scène associé à l'image cible. Ce dernier induit un critère régionalisé fondé sur des densités de probabilité régionales, estimées via des noyaux de Parzen spécifiques sur chaque région du modèle. Nous formalisons ainsi les notions de rapport de corrélation et de f-information régionalisés, dont nous calculons les flots de gradient sur des espaces de transformations non paramétriques et paramétriques. L'application au modèle de scène de la transformation optimale inverse induit de facto une segmentation de l'image source. Nous illustrons les performances de cette approche pour la compensation de mouvements respiratoires et cardiaques complexes lors du transit d'un produit de contraste en IRM de perfusion.

Abstract – Classical statistical similarity measures used in intensity-based non rigid registration rely on global luminance statistics and do not incorporate geometrical information. Ignoring spatial correlations may lead to inaccurate or geometrically inconsistent alignment of homologous geometric structures, making these criteria unreliable for atlas-based segmentation purposes. This paper addresses this limitation and presents a region-driven approach to statistical non rigid registration based on regional non-parametric estimates of luminance distributions. The latter are derived from a regional segmentation of the target image which is used as a fixed object/scene template and induces regionalized statistical similarity measures. We provide their expressions in the case of f-informations and correlation ratio, and derive the corresponding gradient flows over parametric and non-parametric transforms spaces. The relevance of this approach is illustrated for the joint respiratory/cardiac motion artifacts compensation and heart segmentation in short-axis cardiac perfusion MR sequences. using a bi-ventricular heart template.

1 Introduction

Le recalage non rigide d'images a suscité de nombreux travaux en imagerie médicale [1] et en vision par ordinateur. Parmi ceux-ci, les approches iconiques, qui visent à maximiser la similarité inter-image au niveau pixel, ont reçu une attention particulière en raison de leur capacité à générer des cartes d'appariement denses sans pré-segmentation [4, 15]. Dans ce contexte, les mesures de similarité statistiques ont démontré leur pertinence pour la mise en correspondance robuste de données non stationnaires, hétérogènes (multimodales) ou incommensurables. Cette classe de mesures inclut notamment le rapport de corrélation, dont l'optimalité pour des signaux présentant une dépendance fonctionnelle non-linéaire a été suggérée [10], et les f-informations [6], applicables dans l'hypothèse plus faible d'une dépendance statistique et dont l'information mutuelle (IM) constitue un cas particulier très étudié [14, 5, 12, 7, 3, 9].

Dans ces approches, les images sont modélisées comme des variables aléatoires (VA) et manipulées via leurs densités de probabilité marginales et conjointes, généralement estimées de manière globale sur l'intégralité du domaine image. Le recalage s'effectue en maximisant itérativement la similarité statistique entre distributions de luminance, sans exploiter aucune information géométrique. L'absence de prise en compte du contexte spatial peut conduire dans certains cas à un appariement imprécis voire incohérent (bien que visuellement satisfaisant) de structures géométriques analogues, rendant ces critères mal adaptés à la segmentation par propagation d'atlas [11]. En outre, s'ils se révèlent très performants pour des densités conjointes présentant des classes individualisées ou correspondant à des mélanges simples, ces critères peuvent atteindre leurs limites dans le cas de mélanges complexes, tels que ceux générés par des objets texturés sur un arrière-plan complexe.

Afin de résoudre ces limitations, diverses approches visant à enrichir la description statistique par l'apport d'informations spatiales ont été suggérées dans la littérature :

- Une première voie consiste à pondérer le critère statistique par un terme mesurant la similarité des dérivées spatiales des images à recaler. Ce programme a été appliqué à l'IM et à l'IM normalisée [12] avec prise en compte du gradient de luminance, et résolu par maximisation directe sur un espace de transformations paramétriques [8]. L'optimisation variationnelle du critère composite, requise pour traiter le cas de transformations non paramétriques, se révèle complexe.

- Une autre piste consiste à adapter spatialement l'estimation des densités de probabilité impliquées dans le critère statistique. Cette idée peut être mise en œuvre : (i) par le biais d'une partition du domaine image associée à une transformation paramétrique par morceaux [2]; limitée au seul cadre paramétrique, cette approche présente l'inconvénient de recourir à une géométrie extrinsèque aux données, à savoir une grille rectangulaire irrégulière non adaptée aux régions de l'image; (ii) en utilisant des estimées de Parzen locales [3]; outre le fait d'être mono-échelle, cette approche bas-niveau souffre d'une complexité numérique élevée du fait de son caractère local.

– Une dernière possibilité consiste à intégrer une information de support sous forme d'une VA supplémentaire, et à recourir à une mesure de similarité statistique multivariée. Cette idée a été développée dans [11] pour une IM trivariée impliquant les images à recaler ainsi qu'un étiquetage de l'image cible. Elle se heurte toutefois à des difficultés théoriques liées à la généralisation multivariée de critères statistiques définis dans le cas bivarié.

Effectuant une synthèse des précédentes idées, nous proposons dans cet article d'injecter dans le critère de similarité un *a priori* spatial sous forme d'un modèle étiqueté de scène associé à l'image cible. Ce dernier induit un critère de similarité régionalisé fondé sur des densités de probabilité régionales, estimées via des noyaux de Parzen spécifiques sur chaque région du modèle. Nous formalisons ainsi les notions de *rapport de corrélation régionalisé* et de *f-information régionalisée*, dont nous calculons les flots de gradient sur des espaces de transformations non paramétriques et paramétriques. L'application au modèle de scène de la transformation optimale inverse induit alors une segmentation de l'image source. Nous illustrons les performances de cette approche pour la compensation de mouvements respiratoires et cardiaques complexes lors du transit d'un produit de contraste en IRM de perfusion.

2 Recalage non rigide statistique référencé région

Soient I_1 et I_2, des images définies sur un domaine $D \subset \mathbb{R}^n$ et à valeurs dans un intervalle Ω. Recaler I_2 sur I_1 consiste à identifier une transformation $\phi : D \to D$ au sein d'un espace $\mathcal{T}$ de fonctions régulières, telle que $I_2 \circ \phi$ soit similaire à I_1 au sens d'un critère prédéfini. Ceci équivaut à trouver un champ de déplacement $\mathbf{u} = \phi - \mathrm{Id} \in \mathcal{T}$ tel que $I_2^{\mathbf{u}}(\mathbf{x}) := I_2(\mathbf{x} + \mathbf{u}(\mathbf{x}))$ soit similaire à $I_1(\mathbf{x})$ en tout point $\mathbf{x} \in D$. Ce problème est mal-posé et doit être régularisé.

Une approche classique définit une solution comme un minimiseur $\mathbf{u}^*$ d'une fonctionnelle $\mathcal{J}(\mathbf{u}) := \mathcal{S}(I_1, I_2, \mathbf{u}) + \alpha\mathcal{R}(\mathbf{u})$ sur $\mathcal{T}$, combinant un terme de similarité $\mathcal{S}$ quantifiant l'écart entre I_1 et $I_2^{\mathbf{u}}$, et un stabilisateur $\mathcal{R}$ garantissant des solutions p.p. régulières, pondérés par un hyperparamètre $\alpha > 0$.

La minimisation variationnelle de $\mathcal{J}$ sur $\mathcal{T}$ s'effectue en intégrant par descente de gradient les équations d'Euler obtenues en annulant sa dérivée de Gâteaux. Deux cas sont à distinguer selon la nature de l'espace $\mathcal{T}$:

– *Transformations non-paramétriques*: le cas d'un espace de fonctions régulières arbitraires conduit à un problème de minimisation dense : $\mathbf{u}^* = \arg\min_{\mathbf{u}\in\mathcal{T}} \mathcal{J}(\mathbf{u})$, générant un flot sur D : $\frac{\partial \mathbf{u}}{\partial t}(\mathbf{x}) = -\partial_{\mathbf{u}}\mathcal{J}(\mathbf{u}(\mathbf{x}))$ $(\mathbf{u}(\cdot, 0) = \mathbf{0})$ où $\partial_{\mathbf{u}}\mathcal{J}$ est la 1ère variation de $\mathcal{J}$ sur $\mathcal{T}$.

– *Transformations paramétriques* : lorsque $\mathbf{u} := \mathbf{u}^\Theta$ appartient à un espace de fonctions régulières dépendant d'un paramètre $\Theta \in \mathbb{R}^p$, l'optimum $\mathbf{u}^*$ est recherché comme la solution $\mathbf{u}^{\Theta^*}$ d'un problème de minimisation épars : $\mathbf{\Theta}^* = \arg\min_{\Theta} \mathcal{J}(\mathbf{u}^{\mathbf{\Theta}})$, induisant un flot sur $\mathbb{R}^p$: $\Theta_t = -\partial_\Theta \mathcal{J}(\mathbf{u}^{\Theta(t)})$ $(\Theta(0) = \mathbf{0})$ où $\partial_\Theta\mathcal{J}$ est la 1ère variation de $\mathcal{J}$ sur $\mathcal{T}$. Nous nous restreindrons dans la suite à des modèles linéaires pour lesquels $\mathbf{u}^\Theta(\mathbf{x}) = \mathbf{B}(\mathbf{x})\Theta$, $\mathbf{B}(\mathbf{x})$ désignant une matrice encodant une base locale de $\mathcal{T}$ au point $\mathbf{x} \in D$.

2.1 Recalage non rigide statistique global

Nous considérons ici le cas de mesures de similarité statistiques pour lesquelles I_1 et $I_2^{\mathbf{u}}$ sont modélisées comme des VA sur l'espace d'état Ω, de densités respectives p^{I_1} et $p^{I_2^{\mathbf{u}}}$ et de densité jointe $p^{I_1, I_2^{\mathbf{u}}}$. Sans perte de généralité, nous envisagerons spécifiquement les cas suivants:

– $\mathcal{S}(I_1, I_2, \mathbf{u}) = \mathrm{CR}(I_2^{\mathbf{u}}|I_1)$ où CR est le rapport de corrélation [10] défini par :

$$\mathrm{CR}(I_2^{\mathbf{u}}|I_1) := 1 - \frac{E[Var(I_2^{\mathbf{u}}|I_1)]}{Var(I_2^{\mathbf{u}})} \quad . \tag{1}$$

– $\mathcal{S}(I_1, I_2, \mathbf{u}) = -I_f(I_1, I_2^{\mathbf{u}})$ où I_f est une f-information intégrale [13] telle que :

$$I_f(I_1, I_2^{\mathbf{u}}) := \int_{\Omega^2} p^{I_1}(i_1) p^{I_2^{\mathbf{u}}}(i_2) f\big(\rho^{\mathbf{u}}(i_1, i_2)\big)\, di_1 di_2 \tag{2}$$

f désignant une fonction continue convexe sur $\mathbb{R}^+$ et $\rho^{\mathbf{u}}(i_1, i_2) := \frac{p^{I_1, I_2^{\mathbf{u}}}(i_1, i_2)}{p^{I_1}(i_1)\, p^{I_2^{\mathbf{u}}}(i_2)}$. La mesure de Kullback $f_{\mathrm{KL}}(x) := x \log x$ conduit à l'IM, notée I. La fonction $f(x) = \frac{x^\alpha - \alpha x + \alpha - 1}{\alpha(\alpha-1)}$ $(\alpha \neq \{0, 1\})$ définit la I_α-information, notée I_α, qui vérifie: $\lim_{\alpha\to 1} I_\alpha = I$.

Le point central réside dans le calcul de la dérivée de Gâteaux du critère de similarité $\mathcal{S}(\mathbf{u})$. Dans ce but, nous utilisons des estimateurs de Parzen pour exprimer analytiquement les densités $p^{I_1, I_2^{\mathbf{u}}}$ et $p^{I_2^{\mathbf{u}}}$. La densité p^X d'une VA X sur D est ainsi approchée par :

$$p^X(X_0) = \frac{1}{\mu(D)} \int_D \mathcal{K}^d(X(\mathbf{x}) - X_0)\, d\mathbf{x}$$

où $\mathcal{K}^d$ est un noyau de même dimension d que X et $\mu(D) := \int_D d\mathbf{x}$ est le volume de D [14, 3]. Dans le cas de transformations non paramétriques, on montre alors que [3, 6] :

$$\partial_{\mathbf{u}} S = -V^{I_1, I_2^{\mathbf{u}}}(\mathbf{x})\, \nabla I_2^{\mathbf{u}}(\mathbf{x}) \quad (\mathbf{x} \in D) \quad , \tag{3}$$

où $V^{I_1, I_2^{\mathbf{u}}}$ désigne un terme statistique fonction des densités $p^{I_1, I_2^{\mathbf{u}}}$ et $p^{I_2^{\mathbf{u}}}$, décrivant la variation du couplage statistique entre images induite par la transformation $\mathbf{u}$. Il s'exprime sous la forme générale suivante :

$$V^{I_1, I_2^{\mathbf{u}}}(\mathbf{x}) := \frac{1}{\mu(D)} \left\{ \left[\mathcal{K}_2 \star \frac{\partial L_1^{\mathbf{u}}}{\partial i_2}\right] \big(I_1(\mathbf{x}), I_2^{\mathbf{u}}(\mathbf{x})\big) + \mathrm{E}_{I_1}\left(\left[\mathcal{K}_1 \star \frac{\partial L_2^{\mathbf{u}|i_1}}{\partial i_2}\right](I_2^{\mathbf{u}}(\mathbf{x}))\right)\right\} \tag{4}$$

où $\mathrm{E}_{I_1}(\cdot)$ est l'espérance par rapport à I_1, $\star$ désigne la convolution 2D (resp. 1D) sur l'espace d'intensité Ω^2 (resp. Ω), et $L_2^{\mathbf{u}|i_1}(\cdot) := L_2^{\mathbf{u}}(i_1, \cdot)$ pour $i_1 \in \Omega$ donné. La forme analytique des termes $L_1^{\mathbf{u}}$ et $L_2^{\mathbf{u}}$ dépend du critère $\mathcal{S}$ utilisé, et ai précisé dans les Tableaux 1 et 2, respectivement pour le rapport de corrélation et les f-informations intégrales.

Dans le cas de transformations paramétriques, le flot associé dans l'espace des paramètres prend la forme [6] :

$$\partial_\Theta \mathcal{S} = -\int_D V^{I_1, I_2^{\mathbf{u}^\Theta}}(\mathbf{x}) \left[\left(\nabla I_2^{\mathbf{u}^\Theta}\right)^t \mathbf{B}\right](\mathbf{x})\, d\mathbf{x} \quad . \tag{5}$$

TAB. 1: Structure analytique de la dérivée variationnelle du rapport de corrélation [3].

$L_1^{\mathbf{u}}$	$\frac{i_2}{Var(I_2^{\mathbf{u}})}\left[\, i_2 - 2E\big(I_2^{\mathbf{u}}\vert I_1 = i_1\big) + \big(\mathrm{CR}(I_2^{\mathbf{u}}\vert I_1) - 1\big)\big(i_2 - 2E\left(I_2^{\mathbf{u}}\right)\big)\right]$
$L_2^{\mathbf{u}}$	0

TAB. 2: Structure analytique de la dérivée variationnelle des f-informations intégrales [6].

$L_1^{\mathbf{u}}$	$f'\left(\rho^{\mathbf{u}}\right)$
$L_2^{\mathbf{u}}$	$f\left(\rho^{\mathbf{u}}\right) - \rho^{\mathbf{u}} f'\left(\rho^{\mathbf{u}}\right)$

2.2 Introduction d'un *a priori* régional

Supposons donnée une segmentation de l'image cible I_1 sous la forme d'une partition $\mathcal{D} := (D_l)_{0\leq l\leq L}$ du domaine D en L régions d'intérêt, D_0 correspondant par convention à l'arrière-plan. La projection triviale (injection canonique) de $\mathcal{D}$ dans l'image $I_2^{\mathbf{u}}$ induit une partition $\mathbf{u}^{-1}(\mathcal{D})$ de l'image source I_2 : celle-ci s'interprête comme la pseudo-segmentation de I_2 prédite par le modèle de scène $\mathcal{D}$ pour la transformation courante $\mathbf{u}$ (Figure 1). La restriction de I_1 (resp. $I_2^{\mathbf{u}}$) au domaine D_l, notée $I_{1|l}$ (resp. $I_{2|l}^{\mathbf{u}}$), définit la région image de support D_l. Mathématiquement, $I_{1|l}$ et $I_{2|l}^{\mathbf{u}}$ sont des VA sur Ω, de densités respectives $p_l^{I_1}$ et $p_l^{I_2^{\mathbf{u}}}$ et de densité jointe $p_l^{I_1,I_2^{\mathbf{u}}}$. Leur similarité peut donc être quantifiée au moyen d'une métrique statistique usuelle $\mathcal{S}$. En effectuant cette mesure paire à paire, nous définissons un critère de similarité statistique régionalisé référencé par le modèle de scène $\mathcal{D}$, noté $\mathcal{S}_{\mathcal{D}}$, en posant :

$$\mathcal{S}_{\mathcal{D}}(I_1, I_2, \mathbf{u}) := \sum_{l=0}^{L} S(I_{1|l}, I_{2|l}, \mathbf{u}) \quad . \tag{6}$$

Appliquée aux identités (1) et (2), l'équation (6) permet de définir respectivement le *coefficient de corrélation régionalisé* et la *f-informations régionalisée*. Comme dans le cas global, les densités de probabilités intervenant dans ces critères peuvent être estimées par une méthode de Parzen, en recourant cette fois-ci à des noyaux régionalisés $\mathcal{K}_l^d$ de support D_l :

$$p_l^X(X_0) = \frac{1}{\mu(D_l)} \int_{D_l} \mathcal{K}_l^d(X(\mathbf{x}) - X_0)\, d\mathbf{x} \quad .$$

La minimisation de $\mathcal{S}_{\mathcal{D}}$ sur un espace de transformations non paramétriques conduit alors au flot :

$$\partial_{\mathbf{u}} S_{\mathcal{D}} \;=\; -V_l^{I_1,I_2^{\mathbf{u}}}(\mathbf{x})\, \nabla I_2^{\mathbf{u}}(\mathbf{x}) \quad (\mathbf{x} \in D_l) \tag{7}$$

où $V_j^{I_1,I_2^{\mathbf{u}}}$ désigne un terme statistique fonction des densités $p_l^{I_1,I_2^{\mathbf{u}}}$ et $p_l^{I_2^{\mathbf{u}}}$, dont l'expression est identique au cas global (4). Ce terme décrit la variation du couplage statistique entre régions homologues induit par la transformation $\mathbf{u}$. Dans le cas de transformations paramétriques, le flot associé au critère $\mathcal{S}_{\mathcal{D}}$ dans l'espace des paramètres prend la forme :

$$\partial_\Theta \mathcal{S}_{\mathcal{D}} \;=\; -\sum_{l=0}^{L} \int_{D_l} V_j^{I_1,I_2^{\mathbf{u}^\Theta}}(\mathbf{x}) \left[\left(\nabla I_2^{\mathbf{u}^\Theta}\right)^t \mathbf{B}\right](\mathbf{x})\, d\mathbf{x} \;. \tag{8}$$

L'image $(\mathbf{u}^*)^{-1}(\mathcal{D})$ du modèle de scène $\mathcal{D}$ par la transformation optimale inverse $(\mathbf{u}^*)^{-1}$ induit finalement une segmentation de l'image source I_2. L'approche proposée permet ainsi de réaliser conjointement un recalage et une segmentation non rigides de l'image I_2 par propagation du modèle de scène $\mathcal{D}$. Notons qu'elle procède par adaptation itérative des données au modèle, ce dernier demeurant fixe. Elle s'oppose en cela au schéma classique de recalage orienté modèle où la mise en correspondance est réalisée par adaptation progressive du modèle aux données. De façon remarquable, cette stratégie permet de traiter la segmentation non rigide comme un problème d'optimisation à frontières fixes, la segmentation de I_2 étant obtenue par calcul direct une fois $\mathbf{u}^*$ estimée.

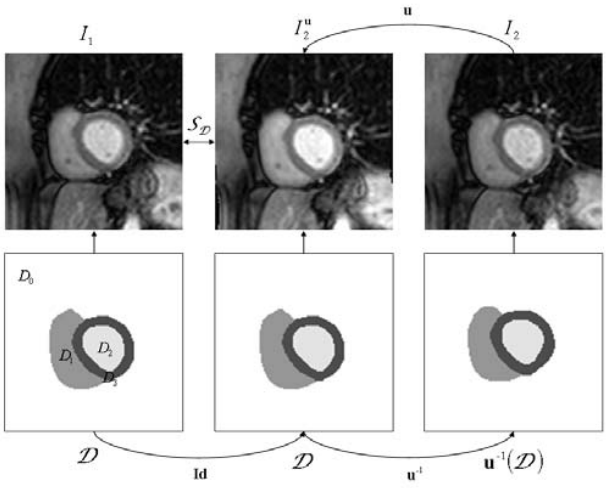


FIG. 1: *Principe du recalage non rigide statistique référencé région : l'image source I_2 est recalée sur l'image cible I_1 en maximisant la similarité statistique $\mathcal{S}_{\mathcal{D}}$ entre régions homologues de I_1 et $I_2^{\mathbf{u}}$, spécifiées par un modèle étiqueté de scène $\mathcal{D}$ associé à I_1 (ici, un modèle bi-ventriculaire du cœur en coupe petit-axe). La projection triviale $\mathbb{I}\mathrm{d}(\mathcal{D})$ de $\mathcal{D}$ sur $I_2^{\mathbf{u}}$ constitue une prédiction du modèle de scène dans l'image recalée par la transformation $\mathbf{u}$ estimée. $\mathbf{u}^{-1}(\mathcal{D})$ détermine une pseudo-segmentation de I_2.*

3 Résultats expérimentaux

Cette approche a été appliquée à la compensation de mouvements respiratoires et cardiaques dans des séquences d'images IRM à des fins d'estimation quantitative de la perfusion myocardique pour le diagnostic de pathologies cardiovasculaires. Dans ces séquences, le transit sanguin d'un produit de contraste (Gadolinium-DTPA) induit des variations de luminance très importantes dans les structures anatomiques et tissulaires irriguées, rendant le recalage complexe et intrinsèquement multimodal. Dans ce cadre, le modèle de scène $\mathcal{D}$ retenu est un modèle bi-ventriculaire de cœur segmenté en coupe petit-axe comprenant 4 régions : thorax (arrière-plan), cavité ventriculaire gauche, myocarde gauche et ventricule droit (Figure 1). A ce stade, ce modèle est obtenu par segmentation interactive d'une image de la séquence choisie comme référence. Deux mesures de similarité ont été évaluées : le rapport de corrélation régionalisé et la I_α-information régionalisée, et calculées via des noyaux de Parzen gaussiens dont les variances ont été automatiquement estimées sur chaque région du modèle par une méthode de cross-validation par maximum de vraisemblance.

Nous avons eu recours à un stabilisateur $\mathcal{R}$ de Nagel-Enkelmann [3, 6] et à une stratégie multirésolution par pyramides gaussiennes pour les images et les déplacements [6].

L'approche a tout d'abord été validée en recalant des images natives avec des versions de synthèse générées par transformation non linéaire de contraste puis déformation par un champ prédéfini. Des tests approfondis ont démontré que le recalage par rapport de corrélation et I_α-information régionalisée fournissent des résultats précis et de qualité comparables, tout en autorisant une accélération de convergence significative par rapport aux mesures globales analogues.

La Figure 2 présente un exemple de compensation de mouvement respiratoires et cardiaques sur quelques images au niveau de coupe basal, échantillonnées lors du passage du bolus de contraste, d'un examen IRM d'un patient post-opéré d'un infarctus du myocarde. Afin d'illustrer la précision du recalage, les contours du modèle bi-ventriculaire dans l'image de référence ont été superposés aux images natives avant et après recalage non rigide. On constate que l'approche proposée permet d'obtenir une segmentation précise de la cavité cardiaque, malgré des variations de contraste fortement non-linéaires et des déformations thoraciques importantes.

4 Conclusion et perspectives

Nous avons présenté un cadre variationnel générique pour le recalage non rigide statistique référencé région, fondé sur l'utilisation d'un modèle de scène étiqueté associé à l'image cible, couplé à une estimation non-paramétrique régionale des densités de probabilité des données. Ce cadre permet de disposer d'un moyen simple et flexible pour enrichir une mesure de similarité statistique par une information géométrique relative au support des objets imagés. Nous avons pu ainsi formaliser les concepts de rapport de corrélation et de f-information régionalisés, dont nous avons illustré la pertinence pour la compensation de mouvements respiratoires et cardiaques lors du transit d'un produit de contraste en IRM de perfusion. Les travaux en cours visent à étendre ce cadre en vue d'intégrer des contraintes de frontières. Les développements applicatifs concernent la génération automatique du modèle bi-ventriculaire.

Remerciements

Ces recherches ont été effectuées dans le cadre d'une convention CIFRE entre GET/INT et Philips Medical Systems, en partenariat médical étroit avec le Service de Radiologie du CHU Pitié-Salpêtrière (Pr. Philippe Cluzel).

Références

[1] J. Gee, J.B.A. Maintz, M. Vannier (Eds.), *Biomedical Image Registration*, Proceedings 2nd International Workshop on Biomedical Image Registration (WBIR'03), Philadelphia, PA - LNCS 2717, Springer-Verlag, Berlin, June 2003.

[2] P. Hellier, *Recalage non rigide en imagerie cérébrale: méthodes et validation* Thèse de Doctorat Université de Rennes 1, Décembre 2000.

[3] G. Hermosillo, *Variational methods for multimodal image matching*, Thèse de Doctorat Université de Nice Sophia-Antipolis, Mai 2002.

[4] H. Lester, S.R. Arridge, "A survey of hierarchical non-linear medical image registration", *Pattern Recognition*, 32(1), Janvier 1999, 129:149.

[5] F. Maes, A. Collignon, D. Vandermuelen, G. Marchal, P. Suetens, "Multimodality image registration by maximization of mutual information", *IEEE Transactions on Medical Imaging*, 16(2), Avril 1997, 187:198.

[6] C. Petitjean, *Recalage non rigide d'images par approches variationnelles statistiques. Application à l'analyse et à la modélisation de la fonction myocardique en IRM*, Thèse de Doctorat Université Paris V - René Descartes, Septembre 2003.

[7] J.P.W Pluim, *Mutual information based registration of medical images*, Ph.D. Dissertation Utrecht University, 2000.

[8] J.P.W. Pluim, J.B.A. Maintz, M.A. Viergever, "Image registration by maximization of combined mutual information and gradient information", *Computer Vision and Image Understanding*, 77(2), Février 2000, 211:232.

[9] J.P.W. Pluim, J.B.A. Maintz, M.A. Viergever, "Mutual-information-based registration of medical images: a survey" *IEEE Transactions on Medical Imaging*, 22(7), Juillet 2003, 986:1004.

[10] A. Roche, *Recalage d'images médicales par inférence statistique* Thèse de Doctorat Université de Nice Sophia-Antipolis, Février 2001.

[11] C. Studholme, D.L. Hill, D.J. Hawkes, "Incorporating connected region labelling into automated image registration using mutual information", *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA'96), San Francsico CA*, Juin 1996, 23:31.

[12] C. Studholme, D.L. Hill, D. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment", *Pattern Recognition*, 32(1), Janvier 1999, 71:86.

[13] I.Vajda, *Theory of Statistical Evidence and Information*, Kluwer Academic Publishers, Dordrecht, 1989.

[14] P. Viola, W.M. Wells III, "Alignment by maximization of mutual information", *IJCV*, 24(2), 1997, 137:154.

[15] B. Zitovà, J. Flusser, "Image registration methods: a survey", *Image and Vision Computing*, 21, 2003, 977:1000.

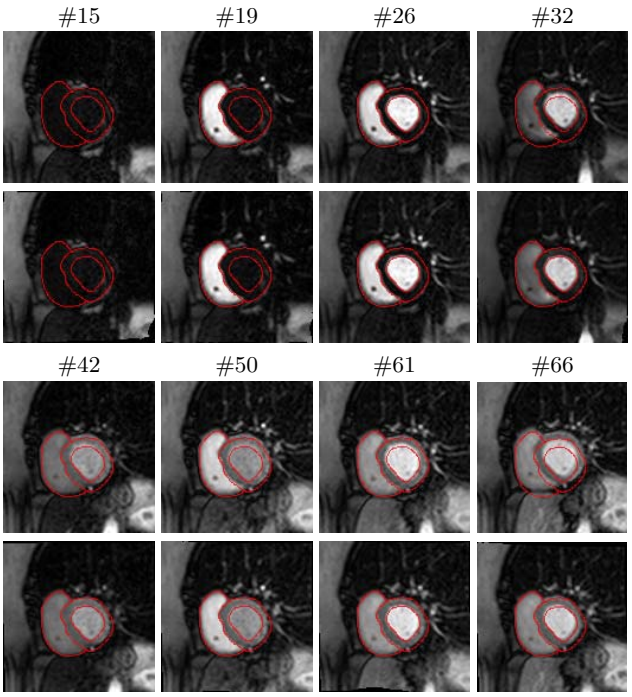


FIG. 2: *Compensation de mouvements respiratoires et cardiaques en IRM de perfusion par recalage non rigide par I_α-information régionalisée ($\alpha = 1.1$) utilisant un modèle bi-ventriculaire de cœur en coupe petit-axe présenté (*cf. *Figure 1). Rangées supérieures : séquence native. Rangées inférieures : séquence compensée. L'image cible de référence est l'image #26. Les frontières du modèle bi-venticulaire ont été superposées afin de mettre en évidence l'effet du recalage.*

Décomposition de déformation pour l'estimation d'un mouvement de caméra

Claire Jonchery[1], Françoise Dibos[1], Georges Koepfler[2]

[1]Laboratoire CEREMADE, Université Paris Dauphine
Place du Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France

[2]Laboratoire MAP5, Université Paris V
45, rue des Saints-Pères, 5270 Paris Cedex 06, France

jonchery@ceremade.dauphine.fr,dibos@ceremade.dauphine.fr
Georges.Koepfler@math-info.univ-paris5.fr

Résumé – Nous proposons dans ce papier une nouvelle approche globale pour estimer le mouvement d'une caméra filmant une scène fixe. Notre méthode considère deux images consécutives d'une séquence et utilise le contenu de ces images et le flot optique sur la séquence. L'originalité de notre approche repose sur une nouvelle modélisation du mouvement de caméra permettant de séparer la déformation entre deux images consécutives en deux composantes : une similitude et une déformation "purement" projective. La similitude est, dans un premier temps, estimée à partir des deux images par un algorithme d'estimation de mouvement paramétrique 2D. Dans un second temps, la déformation projective est évaluée à partir du flot optique et d'un critère de recalage entre les images.

Abstract – In this paper, we propose a new global approach for estimating motion parameters of a camera which films a static scene. Our method deals with adjacent frames of a sequence and utilizes both images content and optical flow. Thanks to an original model of camera motion, our approach separates the deformation between two adjacent frames in two components : a similarity and a "purely" projective application. The similarity is, in a first time, estimated from the two images with a parametric 2D motion model and then, the projective deformation is evaluated from the optical flow and from a registration criterion.

1 Introduction

L'estimation du mouvement d'une caméra est un problème difficile car le mouvement d'un pixel entre deux images dépend non seulement des six paramètres du mouvement de la caméra mais aussi de la profondeur du point de la scène projeté. La littérature présente un grand nombre de modèles différents et de techniques d'estimation variées, séparant ou non l'estimation du mouvement de celle de la structure de la scène. On distingue classiquement trois types d'approches. Les méthodes discrètes d'abord, utilisent des correspondances de points entre les images ; dans [2] par exemple, les appariements mènent à une estimation de la matrice essentielle permettant de déduire le mouvement de la caméra. Les méthodes différentielles ensuite, utilisent le flot optique préalablement calculé entre les images ; un grand nombre de ces méthodes est basé sur la contrainte épipolaire différentielle [6], sur la contrainte bilinéaire [3] ou sur le mouvement de parallaxe [7]. Une présentation et une comparaison de plusieurs techniques, dont la donnée de base est le flot optique, est présentée dans [8]. Les méthodes directes enfin, utilisent directement le contenu des images données. Elles reposent sur la contrainte d'illumination constante, à laquelle on ajoute une hypothèse sur le mouvement et sur les profondeurs de la scène filmée [4].

Notre méthode d'estimation se situe dans le contexte particulier, présenté dans la section 2, d'estimation du mouvement d'une caméra entre deux images consécutives dans une séquence. Dans ce cadre, nous approximons la déformation produite entre les deux images par une déformation plane, elle-même décomposée en une similitude et une application "purement" projective dans la section 3. Grâce à une approximation du flot optique associée à cette décomposition, nous élaborons un algorithme d'estimation du mouvement de caméra, proposé dans la section 4. La section 5 présente plusieurs résultats.

2 Contexte et modélisation

Considérons deux images f et g consécutives dans une séquence. Même dans le cas où la vitesse de la caméra est conséquente, le nombre élevé d'images par seconde limite nécessairement le mouvement de la caméra entre deux acquisitions consécutives. Les deux images f et g sont donc très proches l'une de l'autre et l'effet de parallaxe, n'affectant qu'un très petit nombre de pixels, peut être négligé. On peut aller plus loin ; si la scène filmée est suffisamment éloignée du centre optique, on peut la supposer plane et orthogonale à la direction de l'axe de la caméra avant le déplacement : la structure de la scène est négligée et l'image g est considérée comme une déformation plane de l'image f. Le problème d'estimation du mouvement de

caméra devient alors un problème d'estimation de transformations planes. Notons que plus la scène filmée est éloignée de la caméra, plus cette hypothèse est réaliste.

Nous utilisons le modèle de caméra sténopé classique. Soit (C, i, j, k) le repère orthonormal associé à la caméra, C représentant la position du centre optique et k la direction de l'axe optique. La projection m d'un point M, de coordonnées (X, Y, Z) dans le repère (C, i, j, k), est l'intersection du rayon optique (CM) avec le plan image $\mathcal{R} : \{Z = f_c\}$. Si c est l'intersection de l'axe optique avec $\mathcal{R}$, les coordonnées (x, y) de m dans le repère (c, i, j) du plan $\mathcal{R}$, sont données par

$$\begin{cases} x = f_c \dfrac{X}{Z} \\ \\ y = f_c \dfrac{Y}{Z}. \end{cases}$$

Comme la longueur focale f_c agit comme un facteur d'échelle sur l'image, on peut choisir en toute généralité de fixer f_c à 1.

Considérons maintenant un mouvement de caméra $D = (R, T)$ où R est une rotation d'axe contenant C et T une translation de vecteur t. Le repère associé à la caméra (C, i, j, k) est ainsi transformé par le déplacement D en $(C', R(i), R(j), R(k))$, avec $\overrightarrow{CC'} = t$. Dans ce cadre, en notant

$$R = \begin{pmatrix} a & b & c \\ a' & b' & c' \\ a'' & b'' & c'' \end{pmatrix} \text{ et } t = \alpha i + \beta j + \gamma k,$$

les images f et g sont liées par la relation

$$\begin{aligned} g(x, y) &= (\varphi f)(x, y) = f(\varphi(x, y)) \\ &= f\left(\frac{ax + by + c + \widetilde{\alpha}}{a''x + b''y + c'' + \widetilde{\gamma}}, \frac{a'x + b'y + c' + \widetilde{\beta}}{a''x + b''y + c'' + \widetilde{\gamma}} \right) \end{aligned}$$

où $\widetilde{t} = (\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\gamma})$ est le vecteur de translation quotienté par la profondeur Z_0 de la scène filmée. La matrice $3D$ associée à φ peut s'écrire

$$\mathcal{M}_\varphi = \begin{pmatrix} a & b & c + \widetilde{\alpha} \\ a' & b' & c' + \widetilde{\beta} \\ a'' & b'' & c'' + \widetilde{\gamma} \end{pmatrix} = R \begin{pmatrix} 1 & 0 & \langle \widetilde{t}, R(i) \rangle \\ 0 & 1 & \langle \widetilde{t}, R(j) \rangle \\ 0 & 0 & 1 + \langle \widetilde{t}, R(k) \rangle \end{pmatrix}$$

soit $\mathcal{M}_\varphi = RH$.

Réciproquement, on a

$$\begin{aligned} f(x, y) &= (\psi g)(x, y) = g(\psi(x, y)) \\ &= g\left(\frac{ax + a'y + a'' - \langle \widetilde{t}, R(i) \rangle}{cx + c'y + c'' - \langle \widetilde{t}, R(j) \rangle}, \frac{bx + b'y + b'' - \langle \widetilde{t}, R(j) \rangle}{cx + c'y + c'' - \langle \widetilde{t}, R(j) \rangle} \right) \end{aligned}$$

et la matrice $3D$ associée à ψ est

$$\mathcal{M}_\psi = \begin{pmatrix} a & a' & a'' - \langle \widetilde{t}, R(i) \rangle \\ b & b' & b'' - \langle \widetilde{t}, R(j) \rangle \\ c & c' & c'' - \langle \widetilde{t}, R(k) \rangle \end{pmatrix} = R^{-1} \begin{pmatrix} 1 & 0 & -\widetilde{\alpha} \\ 0 & 1 & -\widetilde{\beta} \\ 0 & 0 & 1 - \widetilde{\gamma} \end{pmatrix}.$$

Les expressions de φ et ψ montrent qu'il est vain de chercher à estimer, à partir des images f et g, à la fois la profondeur Z_0 de la scène et la translation de la caméra. En effet, si le vecteur (t, Z_0) est solution des équations précédentes, alors $\lambda(t, Z_0)$ l'est aussi. On ne pourra donc estimer que la direction de la translation.

Les transformations projectives φ et ψ, classiquement représentées dans le groupe projectif à huit paramètres, ne dépendent que des six paramètres du mouvement : trois respectivement pour la rotation et la translation. Il existe un groupe associé aux déplacements et adapté à la modélisation de telles transformations car il est isomorphe au groupe des déplacements rigides de l'espace $SE(3)$; c'est le groupe des recalages, défini dans [1], dans lequel nous nous plaçons. Il est essentiel pour nous de modéliser les déformations d'images dans un groupe car la structure de groupe permet d'inverser et de composer les déformations (et les mouvements de caméra associés).

3 Décomposition du mouvement d'une caméra

Tout d'abord, toute rotation de caméra R peut se décomposer en deux rotations R_1 et R_2 ; la première, R_1, d'axe Δ dans le plan (C, i, j) transforme la direction de l'axe optique k en $R(k)$ tandis que la seconde, R_2, est une rotation autour du nouvel axe $R(k)$. La rotation R_1 dépend de deux paramètres : θ pour la localisation de Δ dans (C, i, j) et α pour l'angle de rotation ; la rotation R_2 est définie par son angle β. En notant R_a^l la rotation d'axe l et d'angle a, on aboutit après calculs à

$$R = R_2 R_1 = R_\theta^k R_\alpha^i R_{-\theta}^k R_\beta^k = R_{\theta,\alpha} R_\beta^k,$$

avec $R_{\theta,\alpha} = R_\theta^k R_\alpha^i R_{-\theta}^k$.

Pour un mouvement de caméra complet, nous pouvons alors écrire

$$RH = R_{\theta,\alpha} R_\beta^k H.$$

Cette décomposition est intéressante par les transformations qu'elle induit sur l'image : la déformation « purement » projective $r_{\theta,\alpha}$ générée par $R_{\theta,\alpha}$ et la similitude s produite par $R_\beta^k H$. On a alors

$$g = \varphi f = s(r_{\theta,\alpha} f).$$

En conséquence, les six paramètres définissant un mouvement de caméra sont répartis comme suit : deux pour la rotation $R_{\theta,\alpha}$ et quatre pour la translation T et la rotation R_β^k. Nous définirons dorénavant un mouvement de caméra par les paramètres $(\theta, \alpha, \beta, A, B, \lambda)$ où $(-A, -B, \lambda - 1)$ sont les coordonnées de $\widetilde{t}$ dans la base $(R(i), R(j), R(k))$. Cette notation allège l'écriture de l'application projective ψ, l'inverse de φ dans le groupe des recalages

$$\psi(x, y) = \left(\frac{ax + a'y + a'' + A}{cx + c'y + c'' + 1 - \lambda}, \frac{bx + b'y + b'' + B}{cx + c'y + c'' + 1 - \lambda} \right).$$

Le tableau (1) donne les ordres de grandeur des valeurs des paramètres, obtenus expérimentalement en prenant des images variées, de dimensions quelques centaines de pixels, auxquelles nous avons appliqué des transformations projectives définies par les six paramètres. Les valeurs retenues permettent de générer des séquences visuellement réalistes, c'est-à-dire sans impression de saccades visuelles entre les images.

TAB. 1 – Ordres de grandeur des valeurs des paramètres entre deux acquisitions d'images consécutives

paramètre	ordre de grandeur
θ (radian)	$]-\pi,\pi]$
α (radian)	10^{-4}
$\|\beta\|$ (radian)	10^{-2}
$\|A\|$,$\|B\|$ (en pixel)	1
$\|1-\lambda\|$ (en pixel)	10^{-2}

En tenant compte des valeurs données dans le tableau, on obtient, en effectuant des développements de Taylor d'ordre 1 en α, d'ordre 2 en β et en $(1-\lambda)$ sur l'expression de ψ, une expression du flot optique à une précision de 10^{-4} sur les valeurs des six paramètres. Puis, en tenant compte de la dimension des images (quelques centaines de pixels), on aboutit à une approximation quadratique du flot optique en un point (x,y) de f, à une précision de 10^{-1} pixels

$$\left\{\begin{array}{lllll} x'-x\simeq & \underbrace{(\lambda-1)x+A}_{(a)} & \underbrace{+\beta y}_{(b)} & \underbrace{+\alpha x(y\cos\theta-x\sin\theta)}_{(c)} \\ y'-y\simeq & \underbrace{(\lambda-1)y+B}_{(a)} & \underbrace{-\beta x}_{(b)} & \underbrace{+\alpha y(y\cos\theta-x\sin\theta)}_{(c)}. \end{array}\right.$$

Ainsi, le flot optique est décomposé en trois termes indépendants : (a) est dû à la translation, (b) à la rotation R_β^k et (c) à la rotation $R_{\theta,\alpha}$. Ce dernier terme n'intervient que lorsque les termes quadratiques (x^2,y^2, xy) sont au moins d'ordre 10^3. Ceci signifie que pour des images de quelques centaines de pixels, la déformation « purement » projective $r_{\theta,\alpha}$ n'est visible qu'en périphérie de l'image. Le centre de l'image, qui correspond aux plus faibles valeurs de x et y (d'ordre 1 et 10), est principalement déformé par la similitude.

4 Algorithme

Pour f et g deux images successives d'une séquence et $(\theta,\alpha,\beta,A,B,\lambda)$ les six paramètres du mouvement entre les deux acquisitions, l'approximation précédente du flot optique donne, pour des valeurs de $|x|$ et $|y|$ restreintes aux ordres 1 et 10

$$\left\{\begin{array}{l} x'\simeq\lambda x+A+\beta y \\ y'\simeq\lambda y+B-\beta x. \end{array}\right.$$

En conséquence, le centre de l'image f est déformé principalement par la similitude. Cette remarque est à la base de notre méthode d'estimation de mouvement de caméra. Les étapes de l'algorithme proposé sont les suivantes :

(1) estimation des paramètres de la similitude $(\hat\beta,\hat A,\hat B,\hat\lambda)$ à partir des zones centrales des images f et g et en utilisant l'algorithme de mouvement paramétrique 2D d'Odobez et Bouthémy [5], disponible à l'adresse
http ://www.irisa.fr/Vista/Motion2D.
Nous utilisons un modèle de mouvement affine à quatre paramètres, permettant de modéliser les translations et rotations 2D et les homothéties. Ceci permet de calculer V_s, le flot optique généré par la similitude définie par les paramètres $(\hat\beta,\hat A,\hat B,\hat\lambda)$.

(2) calcul du flot optique V entre les images f et g par l'algorithme de Weickert et Schnörr [9],

(3) calcul du flot optique $V_p = V - V_s$.
Le flot V_s étant le flot dû à la similitude estimée à l'étape (1), notre approximation permet de considérer que le flot V_p est dû à la transformation « purement » projective (car $V \simeq V_s + V_p$).

(4) estimation du paramètre θ.
La matrice de la rotation $R_{\theta,\alpha}$ est approximée, à une précision de 10^{-4} sur les valeurs des coefficients par

$$R_{\theta,\alpha}\simeq\begin{pmatrix} 1 & 0 & \alpha\sin\theta \\ 0 & 1 & -\alpha\cos\theta \\ -\alpha\sin\theta & \alpha\cos\theta & 1 \end{pmatrix}.$$

Le flot optique généré par la rotation $R_{\theta,\alpha}$ seule, au point (x,y), est ainsi approximé par

$$\left\{\begin{array}{l} x'-x\simeq\dfrac{\alpha(-x^2\sin\theta+xy\cos\theta-\sin\theta)}{\alpha(x\sin\theta-y\cos\theta)+1} \\ \\ y'-y\simeq\dfrac{\alpha(-xy\sin\theta+y^2\cos\theta+\cos\theta)}{\alpha(x\sin\theta-y\cos\theta)+1} \end{array}\right. .$$

On observe alors que la direction $D_p(x,y)=\arctan(\frac{y'-y}{x'-x})$ du vecteur de flot est indépendante de α. On simule donc huit modèles de directions et on compare les directions D_{θ_k} dues aux rotations $R_{\theta_k,\alpha}$ ($\theta_k=-\pi+\frac{k\pi}{4}$, $1\le k\le 8$ et α arbitrairement choisi) à notre estimation D_p. La valeur sélectionnée minimise un critère L^2

$$\hat\theta=\text{argmin}_{\theta_k}\sum_{(x,y)\in f}(D_p(x,y)-D_{\theta_k}(x,y))^2.$$

(5) estimation de l'angle α.
On construit les images $s(r_{\hat\theta,\alpha_i}f)$, s étant la similitude générée par $(\hat\beta,\hat A,\hat B,\hat\lambda)$ et $r_{\hat\theta,\alpha_i}$ la déformation générée par $R_{\hat\theta,\alpha_i}$, avec α_i appartenant à un ensemble discret et $\hat\theta$ la valeur estimée à l'étape (4). La valeur α_i sélectionnée minimise la différence en norme L^2 $\|g-s(r_{\hat\theta,\alpha_i}f)\|^2$; elle correspond donc au meilleur recalage de f sur g connaissant une estimation des cinq autres paramètres du mouvement.

Dans la méthode présentée, on peut interpréter l'estimation de la déformation purement projective comme une étape de raffinement après l'estimation de la similitude, déformation principale du centre de l'image.

Remarquons que nous recherchons la similitude en utilisant les images et non le flot optique car le flot optique calculé sur une séquence est généralement assez bruité et comme cette première étape conditionne les étapes suivantes, il est préférable de limiter les erreurs d'estimation.

Il est également possible d'estimer directement les paramètres du mouvement de la caméra à partir des deux images en utilisant une autre version de l'algorithme d'estimation de mouvement paramétrique d'Odobez et Bouthémy. Le modèle 2D utilisé est alors un modèle quadratique à huit paramètres, que l'on sait regrouper pour obtenir

les six paramètres du mouvement de la caméra, grâce à la décomposition précédente du flot optique en les trois termes (a), (b) et (c).

5 Résultats et conclusion

Sur une séquence simulée à partir d'une image, nous obtenons de bons résultats ; le mouvement de la caméra est bien estimé. De plus, on a observé que l'étape d'estimation de l'angle θ est particulièrement robuste à un bruit impulsionnel ajouté sur les composantes du flot optique.

Nous avons également appliqué la méthode présentée à un film de notre bureau. En combinant les petits mouvements estimés entre les images successives, on accède au mouvement de la caméra entre des images éloignées dans le temps. Ceci nous permet de réaliser des panoramas, car la scène est suffisamment éloignée du centre optique. De tels mosaïquages sont donnés sur les figures (1) et (2). Remarquons que des discontinuités apparaissent lorsque les zones recalées ne correspondent pas à la profondeur moyenne de la scène. Notre prochain objectif est d'utiliser ces informations et le mouvement de la caméra pour déterminer une carte des disparités de la scène.

FIG. 1 – De gauche à droite, les images 70, 80 et 97 de la séquence du bureau et l'image reconstituée du point de vue de l'image 80.

FIG. 2 – De gauche à droite, les images 20, 35 et 50 de la séquence du bureau et l'image reconstituée du point de vue de l'image 35.

Références

[1] F. Dibos, G. Koepfler, P. Monasse, *Image Alignment*, Geometric Level Set Methods in Imaging, Vision and Graphics, Springer, 2003.

[2] O. Faugeras, *Three-dimensional computer vision : a geometric viewpoint*, MIT Press, 1993.

[3] D.J. Heeger, A.D. Jepson, *Subspace methods for recovering rigid motion i : algorithm and implementation*, IJCV, 7(2) :95-117, 1992.

[4] J.E. Ha, I.S. Kweon, *Robust direct motion estimation considering discontinuity*, Pattern Recognition Letters, 21(11) :999-1011, 2000.

[5] J.M. Odobez, P. Bouthémy, *Robust Multiresolution Estimation of Parametric Motion Models*, Jal. of Visual Communication and Image Representation, 6(4) :348-365, 1995.

[6] Y. Ma, J. Koseckà, S. Sastry, *Linear differential algorithm for motion recovery : A geometric approach*, IJCV, 36(1) :71-89, 2000.

[7] J.H. Rieger, D.T. Lawton, *Processing differential image motion*, Journal of the Optical Society of America, 2(2) :354-360, 1997.

[8] Y. Tian, C. Tomasi, D.J. Heeger, *Comparison of approaches to egomotion computation*, IEEE, Conference on Computer Vision and Pattern Recognition, 315-320, 1996.

[9] J. Weickert, C. Schnörr, *Variational Optic Flow Computation with a Spatio-Temporal Smoothness Constraint*, Technical Report, Computer Science Series, 2000.

Agrégation, sélection et utilisation de l'information de mouvement issue d'un flux MPEG

Renan COUDRAY, Bernard BESSERER,

Laboratoire Informatique Image Interaction, Université de La Rochelle
Av. Michel Crépeau, 17042 La Rochelle Cedex 1, France
renan.coudray@univ-lr.fr.fr, bernard.besserer@univ-lr.fr.fr,
tel:(33) 05 46 45 83 22, fax:(33) 05 46 45 82 42

Résumé – Le stockage d'une grande quantité de vidéos numériques, ainsi que leur diffusion, a été rendu possible grâce à l'utilisation de techniques de compression. Le standard MPEG, largement utilisé, atteint un taux de compression important en réduisant la redondance temporelle. Pour cela, des informations de compensation de mouvement sont calculées par l'encodeur et sont transmises dans le flux. Or, de nombreuses applications (segmentation, indexation, ...) requièrent une information relative au mouvement apparent d'une séquence d'image. Nous présentons dans cet article une façon de réorganiser les données présentes dans le flux MPEG afin de pouvoir réutiliser, avec une fiabilité accrue, l'information de mouvement. La charge calculatoire est moindre, car l'estimation de mouvement est effectuée à l'encodage. A titre d'exemple, nous montrons comment le flux MPEG et les informations qui y sont contenues peuvent être utilisées pour estimer le mouvement accompli par la camera (*global motion*).

Abstract –
Broadcast of digital video became possible using compression methods. MPEG standard is widely used because it allows large ratio compression by canceling redundant temporal information. MPEG compression is performed by a codec that generates information such as motion compensation field. Many applications (segmentation, indexing, ...), which process image sequences, rely on flow field to figure out apparent movement. In this article, we present a method which use the motion compensation field and select only valid information to improve its reliability. Applications gain better performance because motion estimation is carried out by MPEG codec at encoding step. For example, we demonstrate how MPEG flow can be efficiently used to extract global camera movements.

1 Introduction

De nombreuses applications impliquant la manipulation des objets vidéo comme l'indexation ou la segmentation ont besoin de calculer le mouvement présent dans une séquence d'images. Cette opération étant souvent coûteuse en temps de calcul, on se propose de réutiliser l'information de compensation de mouvement présente dans les formats de compression comme le MPEG, information estimée et enregistrée lors de l'encodage du flux vidéo. Le standard MPEG [1](Motion Picture Expert Group), le plus répandu actuellement, est notamment utilisé par la norme DVB (Digital Video Broadcasting, [2]), norme de diffusion des télévisions numériques, ouvrant un champ important d'applications potentielles.

Nous Présentons deux volets de notre recherche : La mise en forme et la sélection d'informations pertinentes de mouvement à partir du flux MPEG (en utilisant autant que possible les informations sous leur forme compressée) et un exemple d'utilisation de ces informations pour estimer le mouvement de camera (*global motion*).

2 Mise en forme des vecteurs mouvements du MPEG

Il est fondamental de rappeler que l'information de compensation de mouvement enregistrée dans un flux MPEG n'est pas forcement valide au sens du flot optique (déplacement des objets). Il s'agit de vecteurs pointant des blocs similaires, le but de cette information étant d'améliorer le taux de compression en réduisant les redondances temporelles. La compensation de l'image s'effectue par bloc (découpage de l'image en macro-blocs). Si ce bloc contient une structure, des contours, alors la probabilité d'estimer un mouvement (au sens du flot optique) augmente, du moins pour la composante orthogonale aux structures présentes dans le bloc. A l'inverse, une région uniforme peut être appariée avec n'importe quelle autre région uniforme ; les vecteurs de compensation affectés à ce bloc ne correspondent probablement pas au mouvement réel.

Nous choisissons de combiner les vecteurs mouvements présents dans les P et les B-frames (images compensées) afin d'obtenir un vecteur pour chaque macro-bloc d'une I-frame (notre information représente alors le mouvement d'une I-frame à une autre, soit les mouvements perçus le long d'un GOP (Group Of Picture)). Ces informations pourrons être rejetées sur la base de l'analyse des coefficients DCT (Discret Cosine Transform), selon la méthode décrite ci-après (2.2).

2.1 Combinaison des vecteurs

Le principe est de chaîner les vecteurs mouvements transmis avec les B et P-frames afin de relier les I-frames. Nous posons comme hypothèse d'être en présence de mouvements du type affine. Nous montrons dans [3] comment combiner et lier bout à bout les vecteurs mouvements, comme illustré sommairement dans **Fig. 1**.

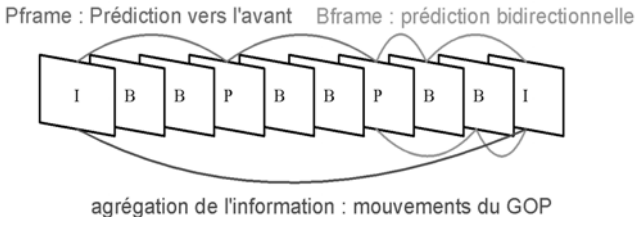


FIG. 1 – Calcul du mouvement du GOP

2.2 Rejet

L'opération de rejet est essentielle pour notre approche, or, la transformation DCT est calculée pour chaque bloc MPEG alors que la compensation de mouvement est calculée par macro-bloc (ensemble de 4 blocs). Afin de se prononcer sur la validité du vecteur mouvement associé au macro-bloc, il faut d'abord rassembler l'information des 4 blocs sous-jacents. Sans cette opération, les variations existantes d'un bloc à l'autre ne seraient pas prises en compte. Par exemple, si nos quatre blocs sont uniformes mais contrastés entre eux, nous ne pouvons pas détecter directement la présence de cette différence qui a pourtant joué un rôle lors de la compensation (opération d'appariement de bloc, ou *block matching*, effectuée par l'encodeur). Après combinaison de l'information, un indice de texture sera calculé, permettant d'estimer la fiabilité du vecteur mouvement issu de la compensation.

2.2.1 Sous échantillonnage

Nous nous appuyons, pour cette première étape, sur une méthode décrite dans [4] qui permet de sous-échantillonner les données DCT, c'est à dire de créer un seul bloc représentant l'information contenue à l'origine dans 4 blocs connexes.

Soit les constantes suivantes (dans le cas du MPEG $N = 8$) :

$$T^N(l,n) = \begin{cases} \frac{1}{\sqrt{N}} \; si \; l = 0 \\ \sqrt{\frac{2}{N}} \cos \frac{(2n+1)l\pi}{2N} \; sinon \end{cases} \quad (1)$$

$$[T_L|T_R] = T^N \quad T_s = T^{N/2} \quad (2)$$

$$C = \frac{1}{\sqrt{2}}\frac{1}{2}((T_L T_s^t) + (T_R T_s^t)) \quad (3)$$

$$D = \frac{1}{\sqrt{2}}\frac{1}{2}((T_L T_s^t) - (T_R T_s^t)) \quad (4)$$

T_N et T_s sont des matrices carrées de dimension N et $N/2$. T_L et T_R sont respectivement les parties gauche et droite T_N ; elles ont donc une taille de $(N, N/2)$.

En notant B_i les parties supérieures gauches de dimension $(N/2, N/2)$ des 4 blocs DCT d'un macro-bloc (numérotés de gauche à droite et de haut en bas) et B le résultat de la combinaison (de taille (N, N)), alors les calculs à effectuer sont les suivants :

– dans l'article [4] :

$$\begin{array}{c} X = C(B_1 + B_3) + D(B_1 - B_3) \\ Y = C(B_2 + B_4) + D(B_2 - B_4) \\ B = (X + Y)C^t + (X - Y)D^t \end{array} \quad (5)$$

– ou de manière équivalente :

$$\begin{array}{c} X = (B_1 + B_2)C^t + (B_1 - B_2)D^t \\ Y = (B_3 + B_4)C^t + (B_3 - B_4)D^t \\ B = C(X + Y) + D(X - Y) \end{array} \quad (6)$$

Il reste à calculer une valeur proportionnelle à la quantité de "formes" présentes dans B, pour estimer la confiance que l'on peut attribuer au vecteur mouvement attaché à notre macro-bloc.

2.2.2 Texture DCT

Ce que nous dénommons texture ou forme correspond à la variance des niveaux de gris présents dans le macro-bloc. Les valeurs des pixels d'un bloc sont décrites grâce aux coefficients DCT (décomposition fréquentielle) et ceux-ci ne sont pas invariants au changement de position de ces formes. Les travaux décrits dans [6] ont pour but d'indexer des blocs DCT et présentent des indices de texture statistiquement stables aux décalages. Nous proposons de modifier deux de ces indices en retirant la composante continue pour obtenir deux indices t_x, t_y qui représentent respectivement la présence de formes horizontales ou verticales :

$$t_x(B) = \sum_{i=1}^{N} B(0,i)^2 \quad (7)$$

$$t_y(B) = \sum_{i=1}^{N} B(i,0)^2 \quad (8)$$

2.2.3 Combinaison des techniques

L'enchaînement du sous échantillonage et de l'extraction des indices de forme en restant dans le domaine compressé entraîne énormément de calculs. Nous avons étudié la possibilité d'obtenir $t_x(B)$ et $t_y(B)$ directement à partir des valeurs des B_i sans passer par X et Y. La simplification des équations **Eq. (9)** et **Eq. (10)** fut notre point de départ. Une démonstration exhaustive, par l'utilisation successives de règles trigonométriques, conduisant à ces formules serait trop imposante dans le cadre de cet article.

$$t_x(X) = \sum_{n=1}^{N/2} X(0,n)^2 = ... = \frac{t_x(B_1 + B_3)}{4} \quad (9)$$

$$\begin{array}{l} t_y(X) = \sum_{l=1}^{N} X(l,0)^2 = ... \\ \quad = \frac{t_y(B_1) + t_y(B_3)}{2} + \frac{(B_1(0,0) - B_3(0,0))^2}{4} \end{array} \quad (10)$$

On trouve des équations équivalentes pour Y (remplacer B_1 par B_2, et B_3 par B_4). Pour fusionner X et Y les équations sont inversées entre t_x et t_y :

$$t_x(B) = \frac{t_x(X) + t_x(Y)}{2} + \frac{(X(0,0) - Y(0,0))^2}{4} \quad (11)$$

$$t_y(B) = \frac{t_y(X + Y)}{4} \quad (12)$$

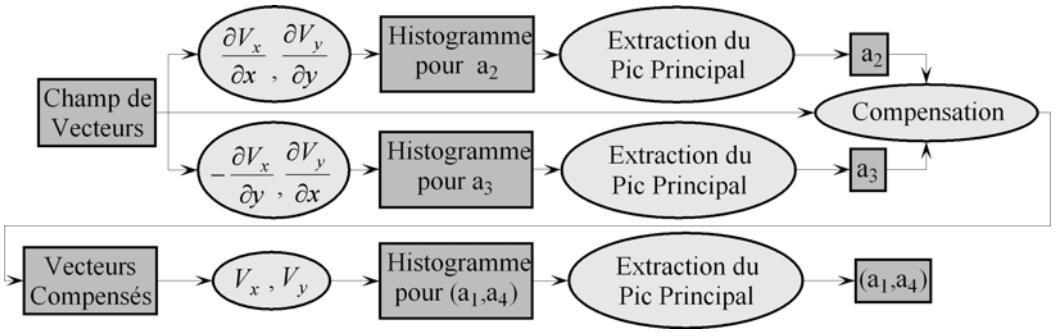


FIG. 2 – Estimation du mouvement global : Synoptique de la méthode basée sur le modèle affine simplifié

Sachant que la composante continue d'un bloc reconstruit est la moyenne des composantes continues des blocs fusionnés et en utilisant les équations **Eq. (9)** et **Eq. (10)**, on obtient t_x en fonction des B_i :

$$t_x(B) = \frac{\frac{t_x(B_1+B_3)}{4} + \frac{t_x(B_2+B_4)}{4}}{2} + \frac{\left(\frac{B_1(0,0)+B_3(0,0)}{2} - \frac{B_2(0,0)+B_4(0,0)}{2}\right)^2}{4} \quad (13)$$

Pour t_y, on ne peut pas connaître la valeur de $t_y(X+Y)$ sans passer par le calcul de X et Y, que l'on veut éviter. Il faut donc se placer dans le cas équivalent (**Eq. (6)**), c'est à dire regrouper les blocs horizontalement en premier pour obtenir au final :

$$t_y(B) = \frac{\frac{t_y(B_1+B_2)}{4} + \frac{t_y(B_3+B_4)}{4}}{2} + \frac{\left(\frac{B_1(0,0)+B_2(0,0)}{2} - \frac{B_3(0,0)+B_4(0,0)}{2}\right)^2}{4} \quad (14)$$

Dans le cas de vidéos entrelacées, les lignes paires et impaires sont souvent encodées séparément en faisant des regroupements (les lignes paires dans deux blocs DCT et les lignes impaires dans les deux autres). Dans ce cas, nous avons 2 vecteurs mouvements différents (les lignes paires et impaires sont compensées de manières indépendantes), mais nous affectons le même indice de texture à ces deux vecteurs. L'influence de la démarcation artificielle introduite pas le regroupement séparé des lignes paires et impaires peut être annulée en retirant l'énergie due aux composantes continues pour l'indice de texture verticale (suppression du terme de droite de la somme de l'équation **Eq. (14)**).

Ces calculs produisent 2 indices de textures qui sont associés aux composante x, y de nos vecteurs mouvements. Ces indices sont transformés en une valeur de fiabilité comprise entre 0 et 1 (utilisation de *least-power influence function*, [11]) permettant de pondérer la qualité de l'information de mouvement selon l'application envisagée. Si la fiabilité d'un vecteur est demandée (et non la fiabilité d'une de ses composantes), le minimum de la fiabilité de ses deux composantes est alors utilisé.

3 Estimation du mouvement global

3.1 Méthode

Les données disponibles pour l'estimation du mouvement global correspondent à un champ de vecteurs épars, et, malgré notre technique de rejet, des vecteurs aberrants subsistent. De nombreuses méthodes d'estimation du mouvement utilisent une technique du type "moindres carrés" [10, 9]. Ces méthodes sont coûteuses en temps de calcul, surtout dans leurs versions récursives permettant le rejet des mouvements des objets. Dans [3] nous proposons une approche d'accumulation dans un espace paramétrique (accumulation des informations proportionnellement à l'indice de fiabilité précédemment présenté), permettant d'écarter les vecteurs correspondants aux mouvements des objets présents dans la vidéo (exemple des footballeurs dans **Fig. 3**) ainsi que les vecteurs erronés restants. La **Fig. 2** représente le processus d'estimation du mouvement de la camera, mouvement respectant le modèle affine simplifié (**Eq. (15)**, 4 paramètres). Nous avons étendu la méthode pour le modèle affine (**Eq. (16)**, 6 paramètres).

$$V = \begin{pmatrix} a_1 + a_2 x - a_3 y \\ a_4 + a_3 x + a_2 y \end{pmatrix} \quad (15)$$

$$V = \begin{pmatrix} a_1 + a_2 x + a_3 y \\ a_4 + a_5 x + a_6 y \end{pmatrix} \quad (16)$$

3.2 Résultats

De nombreuses applications ayant besoin d'une information sur le mouvement présent dans une vidéo peuvent s'appuyer sur ces travaux, en s'accommodant toutefois des restrictions du MPEG et de la méthode : champ de vecteurs épars, un seul champ de vecteurs par GOP seulement et une qualité d'estimation différente selon l'encodeur MPEG utilisé. Nos essais démontrent que les vidéos diffusées sur les chaînes numériques ou provenant de cartes d'acquisition temps réel possèdent une qualité d'estimation suffisante permettant d'imaginer de nombreux cas concrets d'application. Dans [8] des *storyboards* sont crées de manière automatique grâce au calcul du mouvement de la camera. Dans [5] des séquences de sport sont classées par genre grâce au mouvement présent dans les vidéos. Ces deux exemples d'application utilisent les séquences décompressées mais notre technique d'estimation du mouvement conviendrait bien à ceux-ci.

FIG. 3 – Résultat d'une application créant automatiquement une mosaïque à partir d'une séquence (retransmission d'un match de football) au format MPEG. Les I-frames successives sont ajustées entre elles grâce à l'estimation du mouvement de la camera (panoramiques et zooms dans cette sequence).

L'estimation du mouvement caméra ou *global motion* représente l'application générique et simple de notre approche, pouvant être confrontée avec d'autres techniques. Dans [3], nous comparons nos résultats de l'estimation avec une méthode du domaine non compressé, mettant en évidence des résultats très similaires, avec des coûts de calculs réduits (3 à 4 fois le temps réel, sur un pentium 4 2.4 GHz). La **Fig. 3** montre le résultat d'un *mosaicing*, l'arrangement des images étant basé sur l'estimation du mouvement caméra délivrée par cette méthode. Seule la composante continue de chaque bloc est décodée, ce qui explique la qualité de l'image. Ce résultat montre que le mouvement effectué par la camera est correctement estimé par notre méthode. Il reste des imperfections, pour obtenir une mosaïque de qualité supérieure il faudra faire appel aux données des images pour affiner le résultat.

La réutilisation des informations de compensation de mouvement d'un flux MPEG est aussi étudiée pour générer une signature vidéo, destinée à une reconnaissance des contenus diffusés. Des tests sont en cours, notamment la mise en place d'un protocole de sélection d'indices extraits du champ de vecteurs mouvement, protocole adapté à notre problème d'identification de séquences d'images.

4 Conclusion

Nous avons présenté une méthode qui utilise exclusivement les données compressées d'un flux MPEG et qui permet d'obtenir (en temps réel) une information de mouvement exploitable. Malgré les restrictions du MPEG, nos méthodes d'agrégation des vecteurs et leurs rejets éventuels sur la base d'indices de texture, fiabilisent l'utilisation ultérieure de ce champ de vecteurs mouvements. Nos méthodes actuelles délivrent un champ de vecteurs par GOP ; si une précision temporelle plus importante est demandée, il est probable qu'un traitement des données exclusivement dans le domaine compressé ne suffise plus. Malgré ces restrictions, le potentiel recherche est riche et les applications possibles sont nombreuses.

Références

[1] *ISO/IEC 13818-1 and ISO/IEC 13818-2*. 2000.

[2] H. Benoit. *Digital Television MPEG-1, MPEG-2 and Principles of the DVB System, second edition*. Focal Press, 2002.

[3] R. Coudray and B. Besserer. Global motion estimation for MPEG-encoded streams. *Proc. of IEEE Int. Conf. on Image Processing*, 2004.

[4] R. Dugad and N. Ahuja. A fast scheme for image size change in the compressed domain. *IEEE Trans. Circuits Syst. Video Technol.*, 11 :461–474, April 2001.

[5] R. Fablet, P. Bouthemy, and P. Perez. Non parametric motion characterization using temporal gibbs models for content-based video indexing and retrieval. *IEEE Trans. Image Processing*, accepted for publication, 2003.

[6] R. E. Frye and R. S. Ledley. Texture discrimination using discrete cosine transformation shift-insensitive (DCTSIS) descriptors. *Pattern Recognition*, 33(10) :1585–1598, 2000.

[7] J. Heuer and A. Kaup. Global motion estimation in image sequences using robust motion vector field segmentation. *ACM Multimedia*, pages 261–264, November 1999.

[8] S. Porter, M. Mirmehdi, and B. Thomas. Video indexing using motion estimation. *The British Machine Vision Conference*, 2003.

[9] A Smolic and J-R Ohm. Robust global motion estimation using a simplified m-estimator approach. *ICIP*, 2000.

[10] R. Wang and T. Huang. Fast camera motion analysis in MPEG domain. *IEEE Trans. Image Processing*, 3 :691–694, 1999.

[11] Z. Zhang. Parameter estimation techniques : A tutorial with application to conic fitting. *Image and Vision Computing Journal*, 15(1) :56–76, 1997.

Reconnaissance de sauts d'athlète dans les vidéos : fusion du mouvement de caméra et de la trajectoire de points caractéristiques.

Emmanuel RAMASSO, Denis PELLERIN, Michèle ROMBAUT

Laboratoire des Images et des Signaux
LIS - INPG, 46 Avenue Félix Viallet, 38031 Grenoble, France
{Prénom.Nom}@lis.inpg.fr

Résumé – La reconnaissance de l'activité humaine est un champ de recherches très actif dans la communauté du traitement d'image. Dans cet article, nous nous intéressons à la reconnaissance de l'action vue en tant que composante de l'activité. Pour cela, nous utilisons une méthode de fusion de données basée sur la théorie des fonctions de croyances et plus précisément le Modèle de Croyance Transférable. Les données utilisées sont, d'une part, l'estimation du mouvement de la caméra et, d'autre part, la position de points caractéristiques de la silhouette humaine. La méthode est testée pour reconnaître les actions *course*, *saut* et *chute* d'un athlète dans des vidéos de meeting d'athlétisme contenant diverses séquences de saut.

Abstract – Human activity recognition is an active field of research in image understanding and computer vision. We pay attention to action recognition, as an atomic part of an activity. The method is based on data fusion using belief theory and more precisely the Transferable Belief Model. Data concern the camera motion and the trajectory of feature points. The developed method is tested on videos of athletics meetings in order to recognize *running*, *jumping* and *falling* actions in different video sequences of jump.

1 Introduction

L'analyse du mouvement humain dans les vidéos [1] est un challenge important dans le domaine de la vision par ordinateur. Les applications sont nombreuses notamment pour la surveillance de lieux publics ou privés, la recherche et l'archivage de vidéos, le diagnostic médical ou encore l'analyse sportive.

La finalité d'une analyse de mouvement humain est de reconnaître une *activité* que nous percevons comme une *séquence d'actions*. Dans cet article, nous nous focalisons sur la reconnaissance d'actions en vue de celle de l'activité.

« Reconnaître » nécessite de comparer une observation à des modèles et de choisir le plus approprié. De nombreuses méthodes proposées dans la littérature sont basées sur les probabilités [2] avec particulièrement les chaînes de Markov cachées (HMM) et les réseaux bayésiens dynamiques (DBN), ces deux méthodes étant décrites et utilisées dans [3].

La reconnaissance d'actions nécessite d'extraire, à partir des vidéos, des attributs pertinents. Ces attributs sont généralement dépendants de l'application et proviennent dans la majorité des cas d'algorithmes de suivi de points caractéristiques [4] de la silhouette. Quelle que soit leur provenance, les attributs sont généralement entachés d'*imprécisions* dues aux capteurs (e.g. la caméra) ou aux algorithmes de traitement des données. Classiquement, la théorie des probabilités est utilisée pour prendre en compte ces *imperfections*. Nous avons choisi le Modèle de Croyance Transférable (TBM) proposé par Smets et Kennes [5] car ce modèle, en plus de prendre en compte l'imprécision des données, explicite le *doute* ainsi que le *conflit* entre des sources d'informations. Le TBM semble bien adapté à la reconnaissance des actions : d'une part, le doute permet d'exprimer les situations d'*incertitude* entre des hypothèses ce qui s'avère fort intéressant pour décrire l'*aspect graduel des transitions* entre les actions et, d'autre part, le conflit *quantifie* la nécessité d'améliorer la modélisation des informations fusionnées et peut aussi remettre en cause la *qualité* des données.

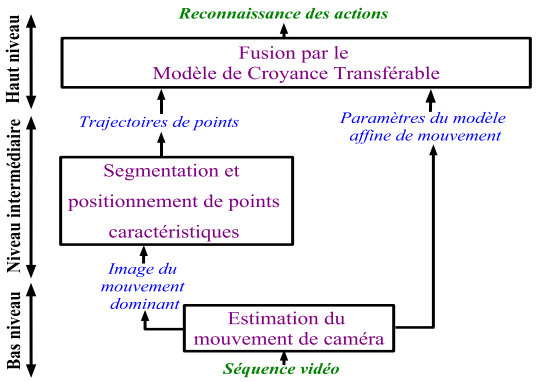


FIG. 1: Architecture du système de reconnaissance d'actions.

Une méthode originale basée sur le TBM est proposée pour le reconnaissance des actions d'athlètes dans des vidéos acquises avec caméra mobile. Une architecture en trois niveaux est présentée dans la section 2. La section 3 est consacrée à la description des deux premiers niveaux qui fournissent les paramètres utilisés par le module haut niveau. Ce dernier, décrit dans la section 4, réalise la fusion et la reconnaissance. Le système est testé sur des vidéos d'athlétisme où l'objectif est de retrouver les actions *course*, *saut* et *chute* dans quatre types de sauts différents. Les résultats présentés dans la section 5 montrent la validité de la méthode.

2 Architecture du système

Le synoptique du système est représenté figure 1. Le fonctionnement est basé sur deux hypothèses liées à l'application :

TAB. 1: Paramètres des niveaux bas et intermédiaire.

Mouvement de caméra		Les points	
a_0	horizontal	(x_c, y_c)	centre de gravité
a_1	vertical	(x_h, y_h)	tête
a_2, a_5	divergence	(x_l, y_l)	pieds

(i) la caméra suit le mouvement de l'athlète, (ii) les mouvements dans l'image de la tête, du centre de gravité et d'un pied de l'athlète renseignent sur l'action en cours. A partir de ces deux hypothèses, nous proposons une structure à trois niveaux, chacun d'eux correspondant à un niveau d'abstraction des données traitées. Tout d'abord, partant de la vidéo originale (fig. 2(a)), un module *bas niveau* estime le mouvement de caméra et fournit, d'une part, six paramètres correspondant au modèle affine et, d'autre part, une image dont l'intensité d'un pixel reflète son appartenance au mouvement dominant. Ensuite, au *niveau intermédiaire*, le positionnement des points caractéristiques sur la silhouette préalablement segmentée est réalisé. Enfin, au *haut niveau*, le mouvement de caméra et la position des points sont combinés dans le cadre du TBM pour obtenir la croyance sur la réalité des actions.

3 Paramètres caractérisant les actions

3.1 Le mouvement de caméra

Nous supposons que la caméra suit grossièrement l'athlète dont les mouvements sont analysés. Nous avons utilisé la méthode robuste d'estimation de modèles paramétriques de mouvement présentée dans [6]. Cette méthode itérative et multi-résolution utilise deux images successives. Le modèle affine (eq. 1) est généralement suffisant pour un grand nombre d'application :

$$\begin{cases} v_x = a_0 + a_2 \cdot x + a_3 \cdot y \\ v_y = a_1 + a_4 \cdot x + a_5 \cdot y \end{cases} \quad (1)$$

Les paramètres a_i utilisés par le module de reconnaissance sont regroupés dans la table 1. Ces paramètres correspondent au mouvement de caméra avec : les translations horizontale et verticale (a_0 et a_1), le zoom (a_5 et a_2) et la rotation (a_4 et a_3 non utilisés). Un lissage de ces paramètres est réalisé par un filtrage Gaussien. Ces paramètres permettent de connaître l'appartenance de chaque pixel au mouvement dominant qui correspond généralement au mouvement du fond (fig. 2(b)).

3.2 Positionnement de points caractéristiques

La méthode décrite dans [7] est exploitée afin de positionner trois points caractéristiques sur une silhouette : deux sont situés aux extrémités (assimilées à la tête et à l'extrémité d'une jambe), et le troisième correspond au centre de gravité (tronc). La trajectoire de ces points est supposée suffisamment pertinente pour la reconnaissance d'actions globales telles qu'une *course*, un *saut* ou une *chute*. Préalablement au positionnement, il est nécessaire de séparer la silhouette du fond.

Segmentation de la silhouette : La segmentation est basée sur l'image du mouvement dominant, obtenue par l'analyse du mouvement de caméra. Un filtrage médian et des opérations morphologiques permettent d'isoler la silhouette (fig. 2(c)).

TAB. 2: Méta-paramètres.

paramètres bruts	méta-paramètres
a_0, a_2	mouvement horizontal
a_1	mouvement vertical
$(x_c, y_c), (x_h, y_h), (x_l, y_l)$	« swing »
$(x_c, y_c), (x_h, y_h), (x_l, y_l)$	alternance des jambes
y_c	variation verticale

Positionnement des points caractéristiques : Le centre de gravité de la silhouette segmentée est calculé et ce premier point caractéristique est ensuite utilisé pour déterminer l'orientation de l'axe principal de la silhouette. Les deux points extrêmes de cette dernière sont alors déduits et correspondent aux deux autres points caractéristiques (fig. 2(c)). Une procédure de re-classification des pixels de la silhouette est réalisée à chaque image afin de rendre plus robuste le positionnement : le principe est de seuiller la plus petite des distances entre les pixels de la silhouette et les points caractéristiques obtenus à l'image précédente puis d'affecter les pixels trop éloignés au fond.

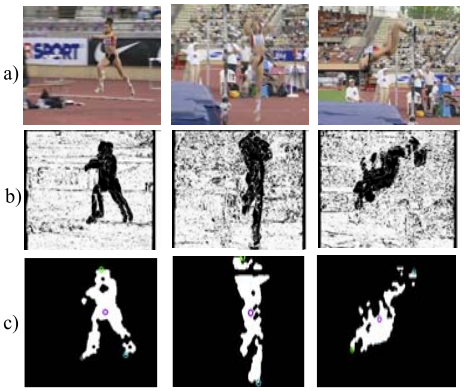


FIG. 2: Exemple de résultat pour un saut en hauteur : (a) images originales pour les actions *course*, *saut* et *chute*, (b) images du mouvement dominant et (c) segmentation et positionnement des points caractéristiques.

3.3 Méta-paramètres

Afin de faciliter la description des actions, des paramètres plus évolués (tab. 2) sont calculés à partir de ceux obtenus par les algorithmes (tab. 1). Par exemple, l'angle que fait l'axe de la silhouette de l'athlète avec l'horizon (« swing ») est calculé à partir des coordonnées des points caractéristiques et permet d'avoir une information sur la manière dont est positionné l'athlète dans l'espace (fig. 3).

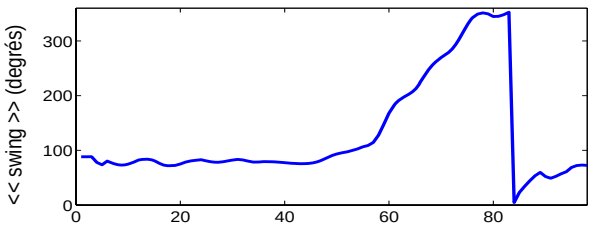


FIG. 3: Angle entre l'axe de la silhouette et l'horizon (« swing ») pour un saut en hauteur.

4 Fusion et reconnaissance

L'évolution dans le temps des différents paramètres obtenus aux deux précédents niveaux doit permettre de reconnaître l'action en cours. Nous proposons une structure de fusion de données basée sur le Modèle de Croyance Transférable (TBM). A chaque valeur des paramètres est associée une croyance en la réalité des actions possibles. Ces croyances sont fusionnées dans le cadre du TBM afin d'obtenir une croyance plus complète tenant compte de l'ensemble des paramètres.

4.1 Le Modèle de Croyance Transférable

Le TBM a été proposé par Smets et Kennes dans [5] et fait suite aux travaux de Dempster et Shafer sur la théorie de l'évidence. Le TBM offre des outils pour la combinaison de sources d'évidence, chacune d'elles exprimant une croyance sur des hypothèses concernant le monde dans lequel elle se trouve.

L'espace de discernement regroupe les différentes hypothèses possibles. Dans le cadre de cet article, l'espace de discernement est $\Omega_A = \{V_A, F_A\}$ et représente l'ensemble des hypothèses concernant une action A, où V_A correspond à « *A est vraie* » et F_A signifie « *A est fausse* ». Les hypothèses sont exclusives et l'espace de discernement est exhaustif.

La masse d'évidence $m^{\Omega_A}_{S_i}$ représente la quantité de croyance affectée à chaque partie (formée par les hypothèses et unions d'hypothèses) de Ω_A par une source S_i. Pour chaque image d'une vidéo, une telle affectation est réalisée à partir de la valeur de chaque paramètre. Cette distribution de masses est réalisée par une application définie $\forall X \in 2^{\Omega_A}$ à valeur dans [0,1] où $2^{\Omega_A} = \{V_A, F_A, V_A \cup F_A\}$. Par construction, $m^{\Omega_A}_{S_i}(\emptyset) = 0$. L'ensemble $V_A \cup F_A$ représente le doute sur la véracité de l'action A sans privilégier une des deux hypothèses.

L'allocation de la masse d'évidence est réalisée pour chaque paramètre mesuré, et relativement aux actions, par une méthode inspirée des sous ensembles flous et revient à une conversion numérique-symbolique. Une illustration est donnée à la figure 4, où a_0 représente le mouvement de translation horizontale de la caméra et où les paramètres de règles ont été définis pour une action *course*. Si $a_0 = 7$ (pixels/image) alors on a $m^{\Omega_A}_{a_0}(V_A) = 1$ (croyance catégorique) et cela signifie que la proposition « A = *course* » est vraie. Si $a_0 = 2,3$ alors $m^{\Omega_A}_{a_0}(F_A) = 0,67$ et $m^{\Omega_A}_{a_0}(V_A \cup F_A) = 0,33$ ce qui signifie que l'action *course* est plutôt fausse mais qu'il persiste un doute. La difficulté est de régler les paramètres de conversion qui sont pour le moment fixés par heuristiques. Cette difficulté est récurrente quel que soit le formalisme adopté (probabilité, évidence, possibilité) si l'on ne dispose pas d'une base d'apprentissage suffisamment riche comme c'est le cas ici.

La combinaison des distributions de masses permet de déterminer la véracité des actions prenant en compte l'ensemble des paramètres. Plusieurs règles de combinaisons ont été proposées, par exemple, la combinaison conjonctive, notée ⓝ, définie par :

$$(m^{\Omega_A}_{S_1} ⓝ m^{\Omega_A}_{S_2})(X) = \sum_{\substack{X,Y,Z \subseteq \Omega_A \\ Y \cap Z = X}} m^{\Omega_A}_{S_1}(Y).m^{\Omega_A}_{S_2}(Z) \qquad (2)$$

avec S_1 et S_2 deux sources distinctes et toutes deux définies sur le même espace de discernement Ω_A. La règle ⓝ étant commutative et associative, il est possible de combiner n sources distinctes S_i en appliquant cette règle en cascade et on obtient $m^{\Omega_A}_{S_{1,2\ldots n}} = m^{\Omega_A}_{S_1} ⓝ m^{\Omega_A}_{S_2} \ldots ⓝ m^{\Omega_A}_{S_n}$. La règle ⓝ est utilisée lorsque plusieurs sources doivent concorder concernant la véracité de A. Lorsque des sources sont en conflit, de la masse apparaît sur l'ensemble vide ($m^{\Omega_A}_{S_{1,2\ldots n}}(\emptyset) > 0$) et ce conflit doit être analysé et résolu car il est absorbant par la règle ⓝ.

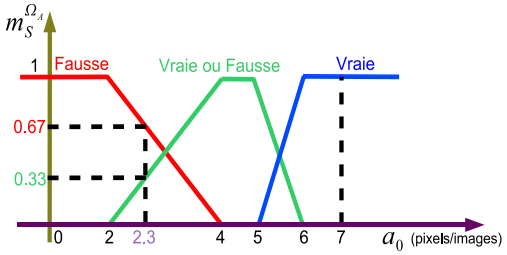


FIG. 4: Un exemple d'allocation de masse pour le paramètre de mouvement de translation horizontale a_0.

4.2 Reconnaissance d'action

La reconnaissance d'une action A est réalisée en trois étapes. Tout d'abord, les distributions de masses, définies sur 2^{Ω_A}, sont déterminées (fig. 4) pour chaque paramètre, les seuils des conversions numériques-symboliques étant fixés pour chaque action et pour chaque type de saut. Ensuite, la combinaison des distributions est réalisée (eq. 2). Enfin, des contraintes temporelles sont appliquées sur la durée minimale des actions : si une action A devient vraie après combinaison, c'est à dire que $m^{\Omega_A}_{S_{1,2\ldots n}}(V_A) > 0$, alors un changement potentiel dans le comportement est détecté. Si cette durée est inférieure à un seuil Δ_A alors la masse est transférée sur le doute, $V_A \cup F_A$, assimilant $m^{\Omega_A}_{S_{1,2\ldots n}}(V_A)$ à du bruit. Ce processus agit comme un filtre passe bas.

Les actions sont indépendantes c'est à dire que la croyance sur la véracité d'une action n'influe pas sur la croyance concernant la véracité des autres actions. De plus, les actions peuvent être vraies en même temps puisqu'elles ont été définies les unes indépendamment des autres (non exclusives). Ce choix a été fait pour pouvoir poursuivre le travail présenté dans cet article vers la reconnaissance d'activités. En effet, une activité étant vue comme une séquence d'actions, une transition entre deux actions successives est une zone (ensemble d'images) où les deux actions peuvent être vraies en même temps.

5 Expérimentations

Conditions expérimentales : La méthode est testée pour la reconnaissance des actions *course*, *saut* et *chute* dans quatre types de saut différents et sur 33 vidéos (tab. 3). Les vidéos, au format 290 × 292 pixels, présentent des angles de vue différents et *a priori* inconnus. Généralement, un seul athlète bouge dans la vidéo mais il arrive que d'autres personnes soient en mouvement en même temps puisque ces vidéos sont des extraits de meetings d'athlétisme. Le réglage des paramètres des conversions numériques-symboliques est réalisé une fois pour chaque action dans chaque type de saut. Enfin, le système de reconnaissance fonctionne, pour le moment, image par image.

Décision et évaluation : Les résultats en sortie du module de reconnaissance sont sous la forme de masses d'évidence.

TAB. 3: La base de vidéos : le nombre d'images concernant les actions *course*, *saut* et *chute* (col. 3-5) et le nombre total de vidéos pour chaque saut (N_V).

saut/action	N_V	course	saut	chute	total
hauteur	9	604	351	205	1160
longueur	8	632	220	213	1065
perche	8	598	417	243	1258
triple saut	8	576	505	377	1458
total	33	2410	1493	1038	4941

TAB. 4: Rappels et précisions (en %) de la reconnaissance des actions *course*, *saut* et *chute* basée sur leur crédibilité.

saut/action	course		saut		chute	
	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{P}$
hauteur	$96,4$	$79,6$	$75,4$	$73,3$	$74,1$	$91,6$
longueur	$90,0$	$80,6$	$58,2$	$52,9$	$64,8$	$70,8$
perche	$81,9$	$75,6$	$74,5$	$70,9$	$72,8$	$75,6$
triple saut	$85,9$	$70,6$	$55,6$	$63,3$	$62,9$	$52,1$
total	$88,6$	$76,7$	$59,6$	$66,1$	$67,8$	$64,0$

Aucune prise de décision n'est volontairement réalisée car ces données alimenteront un module de reconnaissance d'activités. Cependant, l'évaluation de la reconnaissance d'actions nécessite de savoir si une action est vraie ou fausse. Pour cela, nous considérons que si la crédibilité d'une action est non nulle ($m^{\Omega_A}_{S_{1,2\ldots n}}(V_A) > 0$) alors l'action est vraie (fig. 5). C'est une décision stricte par rapport à celle qui considérerait la plausibilité ou la probabilité pignistique. On utilise les indices de rappel et de précision ($\mathcal{R}$ et $\mathcal{P}$) pour évaluer les résultats avec $\mathcal{R}= \frac{C \cap R}{C}$ et $\mathcal{P}= \frac{C \cap R}{R}$ où C est l'ensemble de référence, R est l'ensemble des images retrouvées par le module de reconnaissance (relativement au critère basé sur la crédibilité) et $C \cap R$ est le nombre d'images correctement retrouvées.

Description et analyse des résultats : La table 4 regroupe les valeurs de $\mathcal{R}$ et $\mathcal{P}$. La dernière ligne est la moyenne des résultats sur toutes les vidéos. L'action *course* est convenablement reconnue avec $\mathcal{R} \geq 81.9\%$ et $\mathcal{P} \geq 70,6\%$. Quant aux actions *saut* et *chute*, on peut distinguer deux ensembles de résultats selon le type de saut : un premier pour le saut en hauteur et à la perche avec $\mathcal{R} \geq 72.8\%$ et $\mathcal{P} \geq 70.9\%$, et un second pour le saut en longueur et le triple saut avec $\mathcal{R} \geq 55.6\%$ et $\mathcal{P} \geq 52,1\%$. Les résultats sont dégradés par d'autres personnes ou objets en mouvement (e.g. la perche perturbe le positionnement des points). De plus, les actions ont été reconnues de manière statique mais une description tenant compte des dépendances temporelles entre les actions serait plus pertinente.

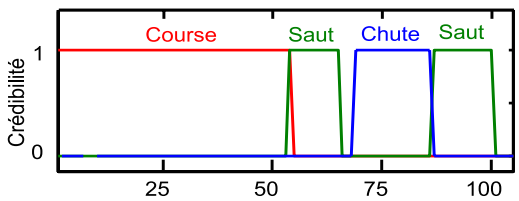


FIG. 5: Evolution de la crédibilité d'actions dans une séquence de saut en hauteur, le deuxième saut est dû à l'athlète qui exprime sa joie lorsqu'il réussit le saut.

6 Conclusion

Une méthode originale de reconnaissance de sauts d'athlète basée sur le Modèle de Croyance Transférable (TBM) a été proposée. Les données combinées proviennent d'un estimateur de mouvement de caméra ainsi que d'un module de positionnement de points caractéristiques de la silhouette. Les valeurs des paramètres ont été traduites en masses d'évidence obtenues grâce à un procédé inspiré des règles floues et fusionnées dans le cadre du TBM. Les tests ont porté sur 33 vidéos où l'objectif était de reconnaître les actions *course*, *saut* et *chute* dans quatre types de saut différents (longueur, hauteur, perche et triple saut). Les résultats ont permis de mettre en évidence l'intérêt de la méthode proposée qui se distingue de celles majoritairement employées dans la littérature, d'ailleurs généralement basées sur les probabilités. Un point fort de la méthode est la mise en évidence du conflit qui rend immédiatement compte d'un réglage inadapté des règles d'allocations de masses et permet de rectifier en conséquence. De plus, le doute entre les hypothèses permettra d'expliciter les transitions entre actions lors de la reconnaissance des activités. Enfin, la prise en compte d'informations complémentaires dans le processus de reconnaissance, comme des règles expertes ou d'autres *a priori*, est rapidement réalisable dans le cadre du TBM. Un point critique de la méthode concerne le réglage des paramètres permettant les allocations des masses : ne disposant pas d'une base d'apprentissage suffisante, nous visons à les adapter en ligne.

Remerciements

Cette recherche est en partie soutenue par le réseau d'excellence SIMILAR. Les auteurs remercient l'équipe Vista de l'Irisa/Inria Rennes pour le logiciel Motion2D sur lequel s'appuient les développements de cette étude.

Références

[1] L. Wang, W. Hu, and T. Tan. Recent developments in human motion analysis. *PR*, 36(3):585–601, 2003.

[2] S. Hongeng, R. Nevatia, and F. Bremond. Video-based event recognition and probabilistic recognition methods. *CVIU*, 96:129–162, 2004.

[3] T. Xiang and S. Gong. Discovering bayesian causality among visual events in a complex outdoor scene. In *Proc. of the IEEE on Advanced Video and Signal based Surveillance*, pages 177–182, 2003.

[4] J. Wang and S. Singh. Video analysis of human dynamics-a survey. *Real-Time Imaging*, 9(5):321–346, 2003.

[5] P. Smets and R. Kennes. The Transferable Belief Model. *Artificial Intelligence*, 66(2):191–234, 1994.

[6] J.M. Odobez and P. Bouthemy. Robust multiresolution estimation of parametric motion models. *J. of Vis. Comm. and Image R.*, 6(4):348–365, 1995.

[7] C. Panagiotakis and G. Tziritas. Recognition and tracking of the members of a moving human body. In *Articulated motion and deformable objects*, pages 86–98, 2004.

Détection MV-TFAC dynamique dans un fouillis de mer hétérogène

Nicolas BON[1], Ali KHENCHAF[1], Jean-Michel QUELLEC[2], Myriam CHABAH[2], René GARELLO[3]

[1]ENSIETA, Laboratoire E3I2-EA 3876
2 rue François Verny, 29806 Brest cedex 9, France

[2]Thales Airborne Systems
BP 58324, 29283 Brest cedex, France

[3]GET-ENST Bretagne
Technopôle Brest Iroise - CS 83818, 29238 Brest cedex, France

bonni@ensieta.fr, ali.khenchaf@ensieta.fr, jean-michel.quellec@fr.thalesgroup.com,
myriam.chabah@fr.thalesgroup.com, rene.garello@enst-bretagne.fr

Résumé – Nous proposons un détecteur TFAC adaptatif qui tient compte de la présence de zones hétérogènes de fouillis dans la fenêtre d´analyse. Il permet d´améliorer les performances des détecteurs classiques comme le CA-TFAC en maintenant un seuil quasiment constant. Nous présentons son application aux fouillis de mer ainsi qu´une extension qui vient modéliser la texture du vecteur Gaussien composé du fouillis comme une somme de sinusoïdes dont les paramètres sont estimés par maximum de vraisemblance par la méthode EM.

Abstract – We propose a ML-CFAR detector which models different intensity areas in the reference window. Thus, it enables to improve classical CFAR detectors in clutter-edge environments and we present its application to spatially correlated sea clutter. An extension which models the texture component of the compound Gaussian clutter vector as the sum of sinusoïds is also proposed to improve performances.

1 Introduction

Un radar doit par définition être capable de détecter les cibles présentes dans un domaine de détection. En l'absence de fouillis ou dans un fouillis uniforme, cette tâche est facilement remplie en fixant un seuil de détection global. Pour maintenir des performances constantes dans un environnement non stationnaire, le radar doit estimer localement les paramètres du fouillis pour fixer un seuil local. C'est le principe des détecteurs à taux de fausse alarme constant (TFAC) qui se basent sur l'observation des données d'une fenêtre d'analyse pour déterminer ce seuil local de détection comme le montre la figure 1.

L'exemple d'un tel détecteur est le CA-TFAC (*Cell Averaging*) [1] qui estime le niveau de fouillis moyen en moyennant les cellules de références. C'est le détecteur optimal quand les cellules de références sont indépendantes et identiquement distribuées selon une loi exponentielle. En présence de zones de fouillis d'intensités différentes dans la fenêtre d´analyse (par exemple zones de pluies, cibles multiples, transitions terre-mer, corrélation spatiale du fouillis) ce détecteur TFAC et ses variantes classiquement rencontrées, *Greatest Of*-TFAC, *Smallest Of*-TFAC et *Ordered Statistics*-TFAC, voient leurs performances fortement diminuées [2], le problème majeur restant la détermination du seuil de détection pour maintenir une fausse alarme constante.

Nous présentons dans la section 2 notre cadre de travail et particulièrement le modèle de fouillis comme vecteur aléatoire Gaussien modulé par une texture. La corrélation spatiale de cette texture engendre des zones d´intensités différentes dont la présence diminue les performances des détecteurs TFAC classiques. Pour en tenir compte, le détecteur TFAC que nous proposons dans cet article modélise la texture comme une fonction paramétrée dont l´amplitude varie dans la fenêtre d´analyse. Les paramètres de la texture sont estimés au sens du maximum de vraisemblance. Dans la section 3, la texture de la fenêtre est modélisée comme deux zones d´intensités constantes distinctes et le logarithme de la texture est modélisé par la somme de p composantes sinusoïdales dans la section 4. Des résultats seront présentés pour chacune des approches.

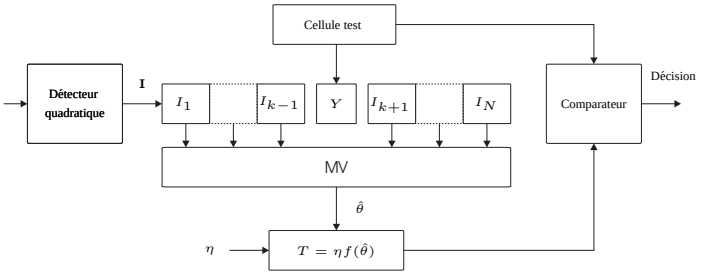


FIG. 1 – Schéma d'un détecteur MV-TFAC.

2 Fouillis de mer Gaussien composé

Le modèle statistique couramment retenu dans la littérature pour un radar à haute résolution distance et aux angles d'incidence rasants est le modèle Gaussien composé [3]. Un tel processus s'exprime comme le produit de deux composantes : $z(n) = \sqrt{\tau(n)}s(n)$ où $s(n) = s_I(n) + s_Q(n)$, couramment appelé *speckle*, c'est un processus Gaussien complexe de moyenne nulle et de puissance unité $E\{|s(n)^2|\} = 1$. Le facteur $\tau(n)$ est un processus aléatoire réel non-négatif couramment appelé *texture*. La surface de la mer peut être vue comme la super-

position de vagues de capillarité et de gravité. Le *speckle* est la conséquence des premières et présente des temps de corrélation d'environ 10 ms à fréquence fixe et est décorrélé par l'agilité de fréquence. La texture présente des longueurs de corrélation beaucoup plus longues [4]. Le fouillis est ainsi couramment modélisé par un processus Gaussien invariant sphériquement (*SIRP*) qui suppose que la texture est invariante pendant le temps d'intégration cohérente : $z(n) = \sqrt{\tau}s(n)$ [3].

Nous nous plaçons ici dans le cas d'un détecteur non-cohérent. L'intensité du fouillis est alors modélisée par $I(n) = \tau(n)|s(n)|^2$ et nous cherchons à améliorer les performances des détecteurs TFAC classiques lorsque la texture est corrélée spatialement. Nous proposons ainsi de modéliser la texture comme une fonction paramétrée représentative des zones d´intensités différentes. Dans la prochaine section la texture est représentée par deux zones d´intensités constantes disjointes et le logarithme de la texture est modélisé comme une somme de sinusoïdes dans la section 4.

3 MV-TFAC dans un fouillis hétérogène

3.1 Deux zones d'intensités différentes

Nous modélisons ici la texture comme deux zones d´intensités constantes disjointes dans la fenêtre d´analyse. La figure 2 illustre cette approche en la modélisant par cinq configurations distinctes suivant la position et la taille des deux zones par rapport à la cellule test. Ces cinq configurations différentes H_i,

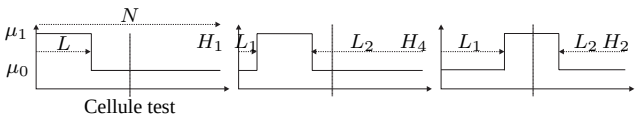


FIG. 2 – configurations H_1, H_2 et H_4. Les configurations H_3 et H_5 sont respectivement les symétriques de H_1 et H_4 par rapport à la cellule test.

$i = 1, \ldots, 5$, sont paramétrées par L ou (L_1, L_2) et les intensités μ_0 et μ_1 (fig.2). On pourra également tenir compte d'une seule zone homogène H_0 (cas du CA-TFAC). L´objectif est ainsi de déterminer la configuration la plus vraisemblable et d'en estimer les paramètres aux vues des observations $I(j), j \in \mathcal{I} = \{1, \ldots, k-1, k+1, \ldots, N\}$ telle que :

$$(\hat{\mu}_0, \hat{\mu}_1, \hat{H}, \hat{\mathbf{L}}(\hat{H})) = \underset{(\mu_0,\mu_1,H,\mathbf{L}(H))}{\arg\max} \prod_{i\in\mathcal{I}} p(I(i)|\mu_0, \mu_1, H, \mathbf{L}(H)) \tag{1}$$

où $\mathbf{L}(H)$ est le vecteur (L_1, L_2) pour H_2, H_4, H_5 et (L) pour H_1, H_3. Nous définissons les ensembles $\mathcal{I}_0$ et $\mathcal{I}_1$ correspondant respectivement aux zones d´intensités moyennes μ_0 et μ_1 telles que $\mathcal{I} = \mathcal{I}_0 \cup \mathcal{I}_1$ et $\mathcal{I}_0 \cap \mathcal{I}_1 = \{\oslash\}$ et en supposant les $I(j), j \in \mathcal{I}$ indépendantes :

$$(\hat{\mu}_0, \hat{\mu}_1, \hat{H}, \hat{\mathbf{L}}(\hat{H})) = \underset{(\mu_0,\mu_1,H,\mathbf{L}(H))}{\arg\max} \prod_{k=1,2} \prod_{i\in\mathcal{I}_k} p(I(i)|\mu_k, H) \tag{2}$$

Pour plus de stabilité numérique, nous calculons ensuite la log-vraisemblance. Le fouillis étant supposé suivre une loi exponentielle et en notant $N_k = card(\mathcal{I}_k), k = 0, 1$, elle s´exprime par :

$$\ln p(\mathbf{I}|\mu_0, \mu_1, H, \mathbf{L}(H)) = -\sum_{k=1,2} \left(N_k \ln(\mu_k) + \sum_{i\in I_k} \frac{I(i)}{\mu_k} \right) \tag{3}$$

Pour une configuration donnée, les moyennes μ_0 et μ_1 sont estimées par annulation de la dérivée de la log-vraisemblance et sont bien sûr égales à $\hat{\mu}_k = \sum_{i\in\mathcal{I}_k} I(i)/N_k$ pour $k \in \{1, 2\}$. Notons que nous pourrons définir une valeur minimale des longueurs L, L_1 et L_2 pour limiter l'erreur d'estimation de la moyenne locale. Une fois que nous avons estimé les intensités moyennes $\hat{\mu}_0$ et $\hat{\mu}_1$ ainsi que la configuration $(\hat{H}, \hat{\mathbf{L}}(\hat{H})$ la plus vraisemblable, nous pouvons appliquer le seuil et décider s´il y a détection en fonction de l´appartenance de la cellule test à telle ou telle zone de fouillis de mer.

3.2 Performances du *MV-TFAC*

Pour présenter les performances de notre méthode, nous l'appliquons à la détection d'une cible suivant le modèle de Swerling 1 dans la configuration H_1 pour $L = 8$, $N = 32$, $\mu_0 = 1$ et $\mu_1 = 15$. Nous effectuons des simulations de Monte-Carlo pour tracer les courbes de détection en fonction du RSB. Le RSB est calculé en fixant l'intensité de la cible par rapport à l'intensité de la zone de fouillis où elle se trouve. En notant Y la cellule test et $\hat{\mu}$ l'estimée de la moyenne locale, le test effectué est $Y/\hat{\mu} \gtrless \eta$. La figure 3 présente les courbes de détection pour $P_{fa} = 10^{-4}$, le seuil de détection étant fixé tel que $\eta = -\ln P_{fa} = 9.21$ (seuil du CA-TFAC sur un fouillis homogène). Nous avons indiqué à titre indicatif les P_{fa} calculées par les simulations de Monte Carlo pour chaque méthodes. Dans cette configuration, nous remarquons une baisse

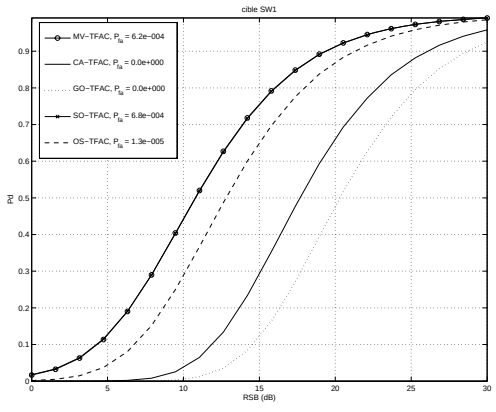


FIG. 3 – Probabilité de détection d´une cible SW1 dans la configuration H_1 avec $\mu_0 = 1$, $\mu_1 = 15$, $L = 8$, $N = 32$, estimées par 10^6 simulations de Monte-Carlo. La 20è statistique est arbitrairement choisie pour OS-TFAC. Le seuil est fixé à $\eta = -\ln P_{fa}$ et pour chaque détecteur TFAC, la P_{fa} calculée est donnée.

significative de performances des méthodes TFAC classiques.

Afin de mettre en évidence l'adaptivité de notre méthode, nous présentons dans le tableau 1 le seuil η à fixer pour maintenir une probabilité de fausse alarme constante de 10^{-4} pour chacune des configurations $H_i, i \in \{1, \ldots, 5\}$ et pour chaque détecteur. Dans toutes les configurations, la zone d'intensité la plus haute est de longueur $N/4$ où N est la taille de la

fenêtre. Notons que si l'intensité moyenne est connue, le seuil

	H_1	H_3	H_2	H_5
MV-TFAC	9.7997	9.5504	11.4746	9.7446
CA-TFAC	2.1852	12.6631	36.5659	2.1840
GO-TFAC	1.2230	9.7237	30.0916	1.2909
SO-TFAC	9.8281	19.5373	51.4121	9.8696
OS-TFAC	6.3565	14.9670	95.3914	6.3562

TAB. 1 – Seuils (η) de chaque détecteur TFAC pour maintenir une $P_{fa} = 10^{-4}$. $N = 32$.

désiré est $\eta = -\ln P_{fa} = 9.21$. Quelques soient les configurations, le seuil du MV-TFAC est constamment proche de cette valeur, c'est une propriété importante qui traduit l'adaptivité du détecteur. Lombardo [11] retrouve les mêmes propriétés en identifiant la fenêtre d´analyse à un mélange de deux lois exponentielles et en identifiant les deux zones homogènes par continuité spatiale. Nous avons appliqué notre méthode sur un

L_c/N	25	50	75	100	125	150	200	300
MV	0.96	0.61	0.54	0.48	0.47	0.43	0.40	0.33
CA	1.64	0.79	0.59	0.46	0.41	0.36	0.31	0.24

TAB. 2 – EQM sur $\hat{\mu}$ pour les détecteurs MV-TFAC et CA-TFAC en fonction du rapport (%) de la longueur de corrélation du fouillis L_c sur la taille N de la fenêtre d'analyse. 10^6 simulations de Monte Carlo, $N = 40$.

fouillis de mer de loi K corrélée spatialement selon [5]. Le tableau 2 montre l´importance d´adapter la taille de la fenêtre d´analyse à la longueur de corrélation du fouillis de mer. Au-delà d´une longueur de corrélation du fouillis égale environ à la taille de la fenêtre d´analyse, la moyenne locale est mieux estimée par le CA-TFAC. Ainsi, en adaptant la taille de la fenêtre d´analyse, notre méthode permet d´améliorer les performances des détecteurs classiques en présence de fouillis corrélé spatialement.

4 Texture comme somme de sinusoïdes

Afin d'améliorer le caractère adaptatif en présence de fouillis de mer corrélé, nous adaptons la même approche que Noga [6] qui travaille sur le logarithme des données tel que :

$$y(n) = \ln I(n) = \ln \tau(n) + \ln |s(n)|^2 \tag{4}$$

et on approche $\ln |s(n)|^2$ par une Gaussienne de paramètres μ, σ^2 tels que :

$$\mu = \psi^{(0)}(1) + \ln(2) \quad \text{et} \quad \sigma^2 = \psi^{(1)}(1) \tag{5}$$

avec $\psi^{(k)}(x) = d^k(\ln \Gamma(x))/dx^k$ fonction polygamma. Noga modélise $\ln \tau(n)$ par un modèle auto-régressif symétrique sur ses voisins (modèle hétéroscédastique). Gini [7] adopte une approche similaire en modélisant la texture elle-même par une somme de sinusoïdes en se basant sur les propriétés de cyclostationarité du fouillis de mer. Ceci motive notre choix de modéliser le logarithme de la texture par :

$$\ln \tau(n) = a_0 + \sum_{k=1}^{p} a_k \cos(2\pi f_k n + \theta_k) \tag{6}$$

où les amplitudes $a_i, i = 0 \dots p$ et les fréquences des sinusoïdes $f_i, i = 1 \dots p$ sont supposées déterministes et les phases sont des v.a. uniformément réparties sur $[-\pi, \pi[$. Avec ce modèle, $b_0(n) = y(n) - \sum a_k \cos(2\pi f_k + \theta_k)$ suit donc une loi Gaussienne $\mathcal{N}(\psi^{(0)}(1) + \ln 2 + a_0, \psi^{(1)}(1))$.

En notant $z(n) = y(n) - \hat{a}_0 - \mu_{b_0}, n = 1 \dots N$, où on remplace a_0 par sa MV-estimée $\hat{a}_0 = \sum_{n=1}^{N} y(n)/N$. Nous nous ramenons à un problème d'estimation spectrale avec estimation des amplitudes et des phases de sinusoïdes noyées dans un bruit blanc Gaussien de moyenne nulle. Sous forme matricielle :

$$\mathbf{z} = \mathbf{E}\mathbf{a} + \mathbf{b}, \left\{ \begin{array}{l} \mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_p] \\ \mathbf{a} = [a_1, \dots, a_p] \end{array} \right. \text{et } \mathbf{b} = \mathcal{N}(\mathbf{0}, \mathbf{R}_b) \tag{7}$$

où $\mathbf{e}_k = [1, \cos(2\pi f_k + \theta_k), \dots, \cos(2\pi f_k (N-1) + \theta_k)]^T$. Les fréquences pouvant être rapprochées, le périodogramme ne peut être utilisé car sa résolution fréquentielle est de l'ordre de $1/N$. Nous choisissons de venir estimer les paramètres au sens du maximum de vraisemblance tels que :

$$(\hat{\mathbf{a}}, \hat{\mathbf{f}}, \hat{\theta}) = \arg\max_{\mathbf{a},\mathbf{f},\theta} \frac{1}{\det(\mathbf{R}_z)} \exp(-\mathbf{z}^T \mathbf{R}_z^{-1} \mathbf{z}/2) \tag{8}$$

avec $\mathbf{R}_z = \mathbf{E}\mathbf{A}\mathbf{E}^T + \mathbf{R}_b$ dépend des vecteurs des paramètres $(\mathbf{a}, \mathbf{f}, \theta)$ à estimer et $\mathbf{R}_b = \psi^1(1)\mathbf{I}_N$ matrice des covariances du bruit.

4.1 Estimation par la méthode EM

Étant donnée la complexité du problème de part le nombre de sources et de paramètres à estimer, la maximisation directe de la vraisemblance est impossible (ou très incertaine par les méthodes d´optimisation classiques de type Newton-Raphson). Dans cette configuration, l'algorithme itératif EM [8] (*Expectation Maximization*) présente des propriétés attractives. Il permet tout d'abord de trouver l'estimée MV par des calculs plus simples sur les données. Ensuite il est stable numériquement et chaque itération augmente la vraisemblance. Chung[9] et Lavielle[10] l'ont précédemment utilisé pour estimer les directions d'arrivées de sources discrètes.

Cette méthode repose sur le concept de données *complètes* et *incomplètes* (voir [8] pour plus de commentaires et illustrations). Nous considérons ici que les données complètes sont constituées par l´ensemble des composantes $\mathbf{z}_k, k = 1 \dots p$ telles que :

$$\mathbf{z}_k = a_k \mathbf{e}_k(f_k) + \beta_k \mathbf{b} \tag{9}$$

où $\mathbf{e}_k = [1, \cos(2\pi f_k + \theta_k), \dots, \cos(2\pi f_k (N-1) + \theta_k)]^T$ et $\sum_{k=1}^{p} \beta_k = 1$. On a donc $\mathbf{z} = \sum_{k=1}^{p} \mathbf{z}_k$.

$$\mathbf{R}_{z_k} = a_k^2 \mathbf{d}_k(f_k) \mathbf{d}_k(f_k)^T + \beta_k \mathbf{R}_b \tag{10}$$

avec $\mathbf{d}_k = [1, \cos(2\pi f_k), \dots, \cos(2\pi f_k (N-1))]^T$. Ce choix est motivé par cette interprétation : si les p sources pouvaient être observées indépendamment, l'estimation des paramètres serait faite directement par p maximisations en parallèle.

Par manque de place, nous ne détaillons par les calculs et données les valeurs des fréquences et amplitudes estimées à l´itération $i+1$:

$$\left\{ \begin{array}{l} f_k^{(i+1)} = \arg\max_{f_k} \frac{\mathbf{d}_k(f_k)^T \mathbf{P}_k^{(i+1)} \mathbf{d}_k(f_k)}{||\mathbf{d}_k(f_k)||^2} \\ a_k^{2(i+1)} = \frac{1}{||\mathbf{d}_k(f_k^{(i+1)})||^2} \left(\frac{\mathbf{d}_k(f_k^{(i+1)})^T \mathbf{P}_k^{(i+1)} \mathbf{d}_k(f_k^{(i+1)})}{||\mathbf{d}_k(f_k^{(i+1)})||^2} - \beta \right) \end{array} \right. \tag{11}$$

avec :

$$\mathbf{P}_k^{(i)} = \mathbf{Q}_k^{(i)} - \mathbf{Q}_k^{(i)}\mathbf{R}_k^{-1(i)}\left(\mathbf{I}_N + \mathbf{z}\mathbf{z}^H\mathbf{R}_k^{-1H(i)}\right)\mathbf{Q}_k^{(i)} \quad (12)$$

où $\mathbf{R}_k^{(i)} = \sum_{m=1}^{p} a_m^{2(i)}\mathbf{d}_m^{(i)}\mathbf{d}_m^{H(i)} + \mathbf{R}_b$ et $\mathbf{Q}_k^{(i)} = a_k^{2(i)}\mathbf{d}_k^{(i)}\mathbf{d}_k^{H(i)} + \beta_k\mathbf{R}_b$. A partir de ces résultats, l´algorithme EM est implémenté de la manière suivante :

répéter les étapes 1-3 pour $k = 1 \dots p$:

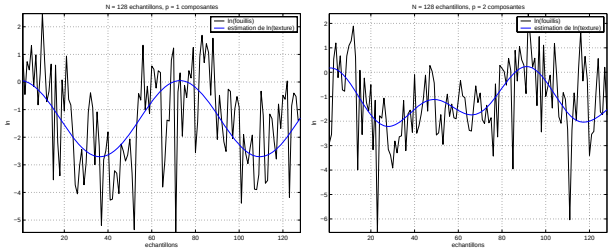
1. choisir $\{a_k^{(0)}, f_k^{(0)}\}$, $i = 0$
2. $i = i + 1$,
   - estimer $\mathbf{P}_k^{(i+1)}$ par (12).
   - estimer la fréquence par (11),
   - injecter la fréquence estimée puis calculer l´estimée de l´amplitude par (11).
3. continuer jusqu´à ce que les estimées convergent.

Nous avons ainsi fourni un estimateur MV des amplitudes et fréquences. Il nous reste à estimer les phases des sinusoïdes. Cette tâche est accomplie par les moindres carrés en venant estimer les amplitudes complexes $\alpha_i = a_i \exp(j\theta_i)/2$ des $2p$ sinusoïdes complexes dont les vecteurs forment les colonnes de la matrice $\mathbf{B} = [\mathbf{e}(-f_p), \dots, \mathbf{e}(-f_1), \mathbf{e}(f_1), \dots, \mathbf{e}(f_p),]$, $\mathbf{e}(f) = [1, \exp(2\pi j f), \dots, \exp(2\pi j f(N-1))]$ avec $\hat{\alpha} = [\alpha_p^*, \dots, \alpha_1^*, \alpha_1, \dots, \alpha_p]^T$ et $\alpha_i = a_i \exp(j\theta_i)/2$:

$$\hat{\alpha} = \left(\mathbf{B}^H\mathbf{B}\right)^{-1}\mathbf{B}^H\mathbf{z} \quad (13)$$

4.2 Résultats

Nous présentons sur la figure 4 l'application de notre méthode à l'estimation du logarithme de la texture d'une loi K corrélée simulée selon [5]. Le nombre de composantes p a été fixé manuellement. Gini [7] propose d´automatiser ce choix en observant la décroissance des valeurs propres de la matrice des covariances du signal. La texture semble correctement estimée



(a) p=1 (b) p=2

FIG. 4 – test de la méthode EM sur l´estimation du log de la texture d´un fouillis de mer de loi K corrélé synthétisé selon [5] dans deux cas différents où les valeurs optimales du nombre de composantes p sont respectivement $p = 1$ et $p = 2$.

et cette méthode pourrait permettre de rendre indépendant le détecteur de la longueur de corrélation du fouillis et de donner une meilleure estimée de la moyenne du fouillis dans la cellule test.

5 Conclusion

Nous avons présenté dans cet article un détecteur MV-TFAC adaptatif qui modélise deux zones d´intensités différentes dans la fenêtre d´analyse et estime leurs paramètres en maximisant la vraisemblance. Cette méthode permet d´améliorer les performances des détecteurs TFAC classiques en présence de transitions fouillis de mer/fouillis terrestre dans le cas de détection côtière par exemple. L´adaptivité de ce détecteur est caractérisée par un seuil de détection maintenu quasiment constant quelques soient les positions des zones d´intensités différentes. Nous avons aussi montré qu´en adaptant la taille de la fenêtre, les performances peuvent être améliorées en présence de fouillis de mer suivant une loi K composée corrélée spatialement. Afin de rendre notre détecteur plus adaptatif dans ces conditions, nous en avons proposé une extension qui vient modéliser le logarithme de la texture par une somme de p composantes sinusoïdales dont les paramètres sont estimés par la méthode EM. Prendre le logarithme du fouillis permet de transformer le modèle multiplicatif (Gaussien composé) en un modèle additif, la densité de probabilité du logarithme du fouillis étant approchée par une loi Gaussienne. Une interprétation qualitative des résultats préliminaires de cette nouvelle méthode nous encourage à poursuivre nos recherches dans cette voie. Nous envisageons ainsi d´évaluer les performances en terme de P_{fa} et P_d de cette méthode sur un fouillis de mer corrélé en distance.

Références

[1] H.M. Finn, R.S. Johnson, *Adaptive detection mode with threshold control as a function of spatially sampled clutter level estimates*, RCA review, 29, 414-464, 1968.

[2] P.P. Ghandi, S.A. Kassam *Analysis of CFAR processors in nonhomogeneous background*, IEEE trans. on Aerospace and Sig. Proc., Vol. 24, No. 4, pp 427-444, 1988.

[3] Conte E., Longo M. *Characterisation of radar clutter as a spherically invariant random process*, IEE Proc., vol.134, part F, No. 2, 1987, pp. 191-197.

[4] K.D. Ward, C.J. Baker, S. Watts, *Maritime surveillance radar. Part 1 : Radar scattering from the ocean surface*, IEE Proc., Vol. 137, No. 2, pp 51-62, 1990.

[5] R. J. A. Tough and K. D. Ward, *The correlation properties of gamma and other non-Gaussian processes generated by memoryless nonlinear transformation*, J. Physics D : Applied Physics, vol. 32, pp. 3075–3084, Dec. 1999.

[6] J. L. Noga, *Bayesian State-Space Modelling of Spatio-Temporal NonGaussian Radar Returns*, Ph.D thesis, Cambridge University, 1998.

[7] F. Gini, G.B. Giannakis, M. Greco, G. Tong Zhou, *Time-averaged subspace methods for radar clutter texture retrieval*, IEEE trans. on sig. proc., vol.49, no. 9, 2001.

[8] A. Dempster, N. Laird, D. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. Royal Statis. Soc. Ser. B, vol 39, pp.1-38, 1977.

[9] P.J. Chung, J.F.Böhme, *DOA estimation using fast EM and SAGE algorithms*, Signal Processing, vol. 82, 1753-1762, 2002

[10] M. Lavielle, E. Moulines, J.F. Cardoso, *A Maximum Likelihood Solution to Doa Estimation for Discrete Sources*, IEEE 7th SP Workshop on Stat. Sig. and Array Proc., 1994.

[11] P. Lombardo *Adaptive CFAR Detection for Clutter-Edge Heterogeneity Using Bayesian Inference*. IEEE Trans. on AES, 39(4) :1462–1470, october 2003.

Imagerie d'objets mobiles à l'aide d'un radar bande étroite multistatique

G. GINOLHAC, F. DAOUT et F. SCHMITT

GEA, Univ. Paris X, PST Ville d'Avray, 1 ch. Desvallières 92410 Ville d'Avray

guillaume.ginolhac,franck.daout,francoise.schmitt@u-paris10.fr

Résumé – *Cet article traite de l'imagerie de cibles mobiles à l'aide d'un radar multistatique (dans notre cas plusieurs émetteurs et un seul récepteur). Tout d'abord, nous développons un algorithme original multistatique basé sur les méthodes spatio-temporelles de Radar à Synthèse d'Ouverture (RSO). Comme le signal émis est bande étroite et que sa fréquence centrale est faible, les résolutions finales de l'image dépendent principalement de deux paramètres : le nombre d'émetteurs et la longueur de l'antenne synthétique. La fonction d'ambiguïté du système est calculé numériquement pour étudier l'influence de ces deux paramètres. Ensuite, l'algorithme développé est testé sur des cibles réalistes. Les images sont intéressantes et permettent des premiers résultats de classification. A l'aide des modèles numériques des cibles, nous montrons aussi l'importance du placement des émetteurs ainsi que la nécessité d'un second récepteur.*

Abstract – *This article deals with the imagery of mobile targets by using a multistatic radar (several transmitters and one receiver). First, we develop an original multistatic algorithm based on spatio-temporal Synthetic Aperture Radar (SAR) methods. As the transmitted signal is narrowband and the central frequency is weak, the final resolutions of the image depend on two parameters: the number of transmitters and the synthetic array length. The ambiguity function of the multistatic system is used to study the influence of these both parameters. We check this algorithm on realistic targets. Obtained images are interesting and allow first classification results. Thanks to numerical models of targets, we also show the importance of the transmitters locations as well as the need of a second receiver.*

1. Introduction

Le Radar à Synthèse d'Ouverture (RSO) [1] consiste à utiliser le mouvement du radar dans le but de simuler une antenne de grande taille et donc obtenir une très bonne résolution spatiale. Si les signaux émis par le radar sont larges bandes, alors il est possible d'avoir une image tout à fait correcte de la zone couverte par le RSO. Le problème dual au RSO consiste à utiliser le mouvement de la cible que l'on désire imager et au contraire d'avoir un radar fixe. Ce type de système est appelé Radar à Synthèse d'Ouverture Inverse (RSOI) [1] [4]. On considère dans cet article ce type de système mais avec quelques caractéristiques particulières : la fréquence centrale du signal émis est assez petite (de 100 MHz à 800 MHz suivant l'application), la bande est étroite (moins de 100 kHz), le radar est multistatique. Dans ce papier, il est constitué de plusieurs émetteurs et d'un seul récepteur. Ce type de système peut se retrouver dans les radars passifs multistatiques fonctionnant avec des émetteurs FM ou TV [4]. Dans ce cas, on utilise les signaux d'opportunité pour imager une cible qui se déplacerait dans la zone couverte par le radar.

Néanmoins, l'imagerie est très délicate à cause des caractéristiques décrites ci-dessus. Tout d'abord, le caractère multistatique du radar nécessite de faire évoluer les algorithmes d'imagerie [1] [3] existants. Afin d'atteindre des résolutions correctes, il faut agir en particulier sur deux paramètres : augmenter le nombre d'émetteurs pour pallier à l'aspect bande étroite des signaux et augmenter la longueur d'antenne synthétique (temps d'illumination de la cible) pour contrer la faiblesse de la fréquence centrale. Nous calculons la **fonction d'ambiguïté** du système pour étudier l'influence de ces deux paramètres sur les résolutions ainsi que sur l'importance des lobes secondaires (*Principal Lobe Secondary Ratio : PLSR*). Nous montrons ensuite l'intérêt de ce type de système pour classifier des cibles réalistes [2]. Ainsi, des cibles blanches en fréquence mais anisotrope spatialement sont utilisées. La première hypothèse est justifiée par la largeur de bande du signal considérée. En revanche, l'anisotropie spatiale permet de décrire un comportement réaliste de cible vis à vis des différentes positions des couples émetteurs – récepteurs. Par ce biais, l'importance des positions des émetteurs sera évaluée.

Cet article est organisé de la façon suivante : nous présentons dans la section 2 le contexte et l'algorithme multistatique d'imagerie. Dans la section 3, nous donnons les performances de cet algorithme. Finalement, nous l'appliquons dans la section 4 sur deux cibles géométriquement proches pour le tester dans des conditions non favorables.

2. Algorithme d'imagerie

Le schéma du système radar étudié est montré sur la figure 1. Un repère fixe *(0,x,y,z)* est utilisé pour fournir les localisations du récepteur $R\ (x_r, y_r, -h)$ et des M émetteurs $E_i\ (x_{ei}, y_{ei}, -h)_{1\le i\le M}$. La cible se déplace le long de l'axe *y* et est vue par le récepteur entre y_{min} et y_{max}. Un repère mobile *(x',y',z')* est utilisé pour décrire la cible à imager. Son origine

est localisée sur l'axe *y* et il suit le mouvement de la cible. Comme pour le RSO, nous faisons l'hypothèse « *stop and go* » (la cible est supposée fixe pendant un certain temps puis se déplace directement à la position suivante) qui permet de décrire le mouvement de la cible. Le repère mobile est ainsi localisé à certains points tous les Δy (échantillonnage spatial).

La cible est supposée avoir une vitesse uniforme linéaire de 150 m/s. Le signal émis, noté *e*, est une Fréquence Modulée Linéairement (FML) de fréquence centrale $f_0=300$ MHz, de bande spectrale $B=75$ kHz et de durée $T=0.2$ ms. Ce type de signal a des propriétés spectrales proches des signaux FM ou TV [4] [5] car la bande est très petite et la fréquence centrale est très inférieure à celle des applications de RSOI traditionnelles [1]. L'échantillonnage temporel f_e est égal à 170 kHz, alors que l'échantillonnage spatial Δy est égal à 0.25 m. Pour tester l'algorithme d'imagerie, nous utilisons $M=103$ émetteurs localisés dans un carré de 100 km×100 km. Leur disposition est réaliste car elle est identique à celle des émetteurs FM près de Guéret.

L'algorithme multistatique est ici développé à partir des méthodes spatio-temporelles car les méthodes fréquentielles sont très difficiles à mettre en oeuvre dans des configurations multistatiques générales. Tout d'abord, nous donnons la relation du retard que met l'onde entre l'émetteur *i*, un point situé dans le repère mobile (sa position est repérée par *n*) à la position *(x',y',0)* et finalement le récepteur :

$$\delta_i(n,x',y') = \frac{1}{c}\left(\sqrt{(x_{ei}-x')^2+(y_{\min}+n.\Delta y+y'-y_{ei})^2+h^2} + \sqrt{(x_r-x')^2+(y_{\min}+n.\Delta y+y'-y_r)^2+h^2}\right)$$

Le modèle du signal reçu est donné par :

$$m(t,n,x',y') = \sum_{i=1}^{M} e(t-\delta_i(n,x',y'))$$

où $t=k/f_e$ est le retard pour chaque position du repère mobile qui est aussi la position de la cible à $y_{\min}+n.\Delta y$. Alors, l'algorithme consiste à effectuer la corrélation entre le signal reçu *s* et le modèle *m* afin d'obtenir l'image *I(x',y')* de la cible :

$$I(x',y') = \sum_{n=1}^{N}\int_{t=0}^{T_{acq}} s(t,n)m^*(t,n,x',y')dt$$

où T_{acq} est la durée d'acquisition du signal reçu pour chaque position de la cible et *N* est le nombre total de positions du repère mobile ($y_{\max}-y_{\min}=N.\Delta y$). Ce programme est notée Corrélation Spatio-Temporel Multistatique (CSTM). Une version plus rapide de cet algorithme est possible en calculant le filtre adapté pour chaque position de la cible avant le processeur RSOI. Ceci est inspiré par l'algorithme *Back Projection* [1] développé pour les RSO monostatique. Cet algorithme est noté *Back Projection* Multistatique (BPM).

3. Performances du BPM

Dans cette partie, nous exprimons la fonction d'ambiguïté du système multistatique décrit dans ce papier. Celle ci représente le produit scalaire entre les signaux émis par deux points brillants localisés en différents lieux de l'espace. L'un est supposé centré à l'origine du repère mobile, alors que le second est localisé à la position *(x',y')* :

$$A(x',y')=\sum_{n=1}^{N}\int_{t=0}^{Tacq}\sum_{i=1}^{M} e(t-\delta(n,0,0))e^*(t-\delta(n,x',y'))dt$$

Il est très difficile de calculer analytiquement cette expression et donc nous proposons de l'évaluer numériquement. En particulier, nous étudions les évolutions des résolutions à –3 dB et de la PLSR en fonction du nombre d'émetteurs et de la longueur de l'antenne synthétique ($y_{\max}-y_{\min}$). La figure 2 montre l'évolution de la résolution pour les deux axes *x'* et *y'* en fonction du nombre d'émetteurs. La longueur de l'antenne synthétique est ici égale à 3 km. Nous notons qu'une résolution correcte (autour de 1 m qui est proche de $\lambda = c/f_0$) est atteinte dans les deux axes avec seulement 10 émetteurs (ils sont choisis aléatoirement parmi les 103 émetteurs présentés dans la section 2). La figure 3 montre, pour les mêmes paramètres de simulation, l'évolution de la PLSR pour les deux axes en fonction du nombre d'émetteurs. Un résultat correct (autour de –10 dB) est atteint avec **50 émetteurs**. Nous faisons la même étude en fonction de la longueur de l'antenne synthétique (le nombre d'émetteurs est fixé dans ce cas à 21). Nous donnons les principaux résultats : des résolutions correctes et une PLSR assez faibles sont possibles avec une longueur d'antenne synthétique de **70 km**.

Cette section a permis de montrer que l'image d'une cible est possible même si le radar est bande étroite et une faible fréquence centrale. Mais, l'imagerie nécessite d'avoir à disposition plusieurs émetteurs et surtout une importante longueur d'antenne synthétique. Ce dernier point est un défaut de la méthode proposée car celle-ci nécessite que la cible soit stationnaire durant toute la longueur de l'antenne synthétique. Il paraît donc ambitieux qu'une cible reste aussi longtemps stationnaire et cela nécessite une étude plus précise.

4. Simulation sur cibles réalistes

La procédure de simulation pour calculer et insérer les signatures des cibles est présentée dans [2]. Nous considérons 2 cibles (voir figure 4) : un cylindre avec un cône et le même cylindre avec deux ailes et une dérive. Les différences ne sont pas importantes, et de plus, leur taille est faible (autour de 3λ) et donc la classification entre ces deux cibles risquent d'être difficiles. Tout d'abord, nous rappelons les résultats obtenus dans [2] avec un radar bistatique et une longueur d'antenne synthétique de 70 km. Seule la longueur du cylindre était mesurable sur les deux images et de plus la classification entre les deux cibles était impossible. Nous effectuons la même simulation mais en considérant 103 émetteurs. Les figures 5 et 6 montrent les images, obtenues par le BPM, des deux cibles pour la polarisation principale. Nous pouvons noter que la longueur et la largeur du cylindre sont mesurables. De plus, l'image de l'avion contient des zones d'énergies plus importantes et certaines faibles différences sont visibles (comme montré par le cercle noir). La figure 7 montre les mêmes résultats pour une polarisation croisée. Le niveau de l'image du cylindre est très faible alors qu'il reste important pour l'avion. La classification en polarisation croisée entre les deux cibles est donc simple grâce a l'algorithme développé. Les différences proviennent principalement de la présence des ailes.

Nous désirons maintenant expliquer les résultats obtenus. Les phases des Section Electromagnétique Radar (SER) des deux cibles sont montrées sur les figures 8, 9, 10 et 11 pour les polarisations principales et croisées. Les angles incidents et reçus pour la simulation considérée sont ajoutés aux

figures. Tout d'abord, nous notons que les phases des SER en polarisation croisée sont très différentes entre le cylindre et l'avion. Il est donc normal que la classification dans ce mode soit assez facile. Pour la polarisation principale, les différences sont moins importantes et de plus certaines zones comportant des différences ne sont pas prises en compte par le radar multistatique (par exemple dans le cercle noir). Le résultat de classification présenté n'est donc pas optimal. En fait, nous notons l'importance d'avoir des émetteurs localisés dans toute la zone pour couvrir tous les angles (θ, θ_s). Il semble aussi qu'un second récepteur serait intéressant pour avoir plus de couples (θ, θ_s).

5. Conclusion

Nous avons développé un algorithme original d'imagerie d'objets mobiles pour un radar bande étroite multistatique. La fonction d'ambiguïté du système est numériquement calculée pour étudier les performances du radar multistatique. Nous montrons l'importance du nombre d'émetteurs et de la longueur de l'antenne synthétique pour obtenir une image correcte d'une cible. Finalement, nous appliquons cet algorithme sur deux cibles légèrement différentes. Nous montrons que les images obtenues permettent plusieurs mesures ainsi que des premiers résultats de classification. De plus, nous montrons l'importance du placement des émetteurs et l'intérêt de l'apport d'un second récepteur pour une meilleure classification.

Références

[1] M. Soumekh. "*Synthetic Aperture Radar Signal Processing*", Wiley – Interscience Publication, 1999.

[2] F. Daout, G. Ginolhac, F. Schmitt and M. Lesturgie, "Passive Bistatic Radar Imagery of Targets Modelled by NEC", *RADAR 2004, Toulouse, FRANCE*, 2004.

[3] M. Soumekh. "Bistatic Synthetic Aperture Radar Inversion with application in Dynamic Object Imaging", *IEEE Trans. on Signal Proc.*, vol. 39, n°9, pp 2044-2055, 1991.

[4] Y. Wu and D.C. Munson. "*Multistatic Synthetic Aperture Imaging of Aircraft using Reflected Television Signals*", in Algorithms for Synthetic Aperture Radar Imaging VIII, Proc. SPIE 4382, Ed. E.G. Zelnio, Orlando, FL, April 2001.

[5] P. Forster and M. Lesturgie, "Passive Radar for Target Tracking and Classification", *RADAR 2004, Toulouse, FRANCE*, 2004.

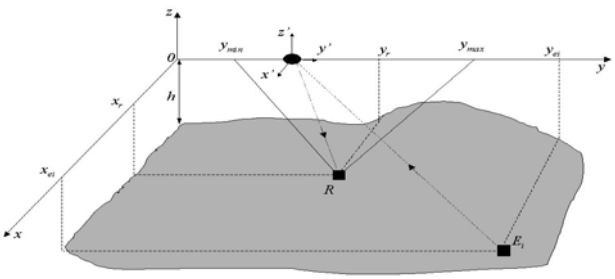


FIG. 1 : Schéma du radar multistatique.

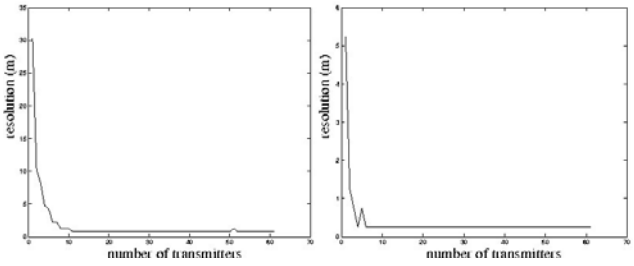


FIG. 2 : Evolution des résolutions sur les axes *x'* (gauche) et *y'* (droite) en fonction du nombre d'émetteurs.

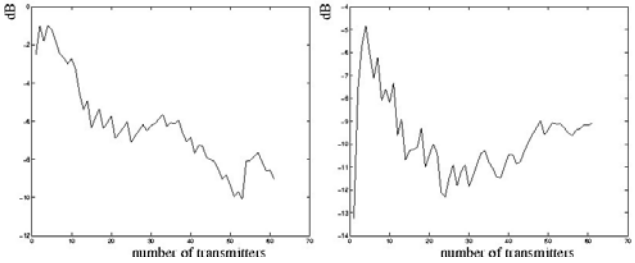


FIG. 3 : Evolution de la PLSR sur les axes *x'* (gauche) et *y'* (droite) en fonction du nombre d'émetteurs.

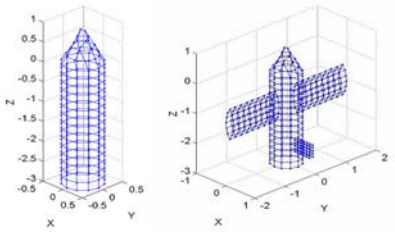


FIG. 4 : Modèles de cibles : cylindre (gauche) et avion (droite).

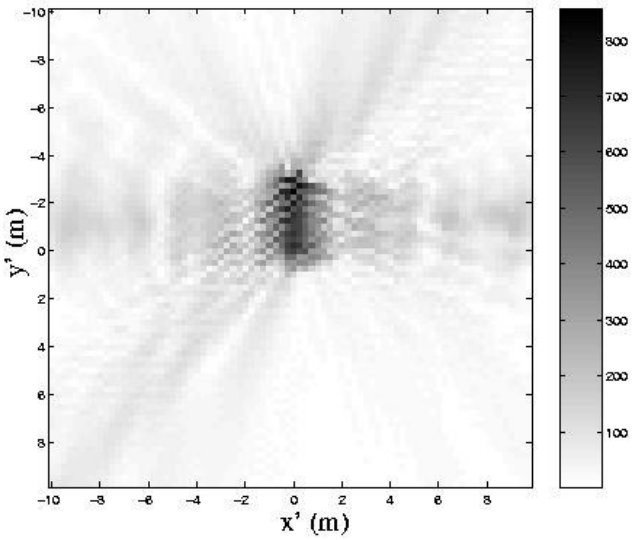


FIG. 5 : Image multistatique du cylindre en polarisation principale.

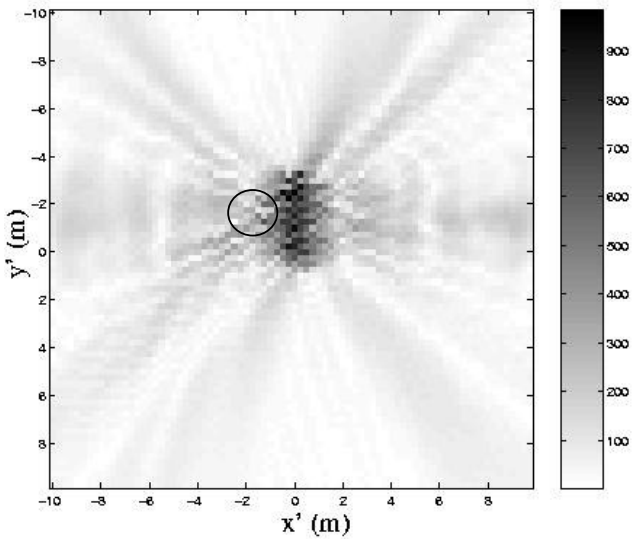


FIG. 6 : Image multistatique de l'avion en polarisation principale.

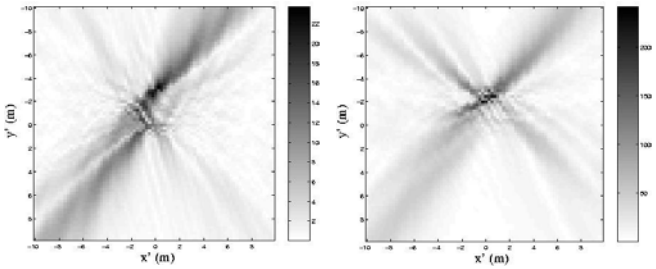


FIG. 7 : Images multistatiques du cylindre (gauche) et de l'avion (droite) en polarisation croisée.

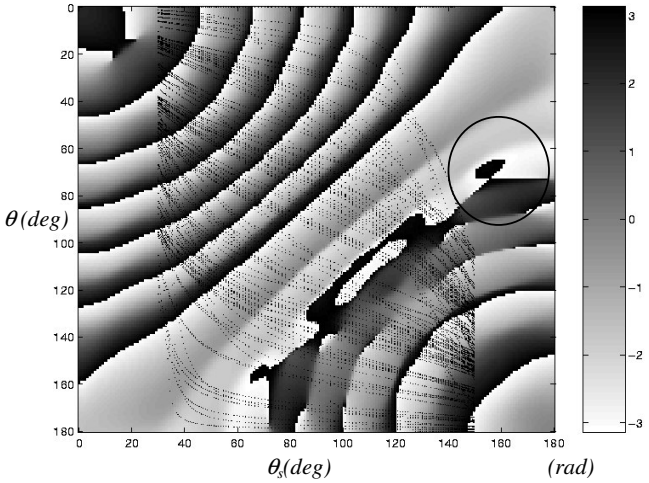


FIG. 8 : Phase de la SER du cylindre en polarisation principale.

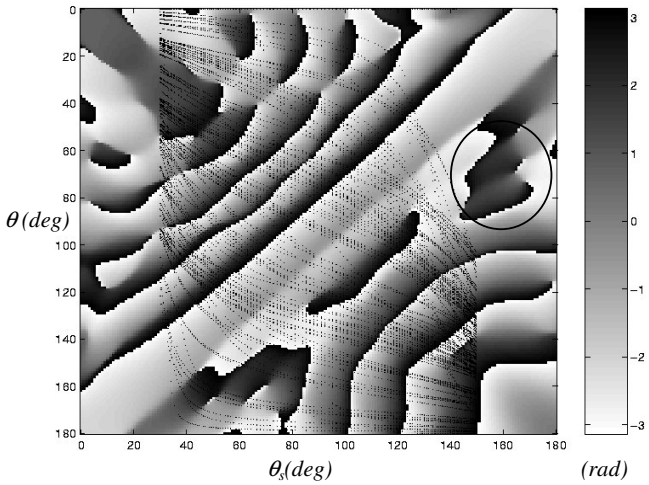


FIG. 9 : Phase de la SER de l'avion en polarisation principale.

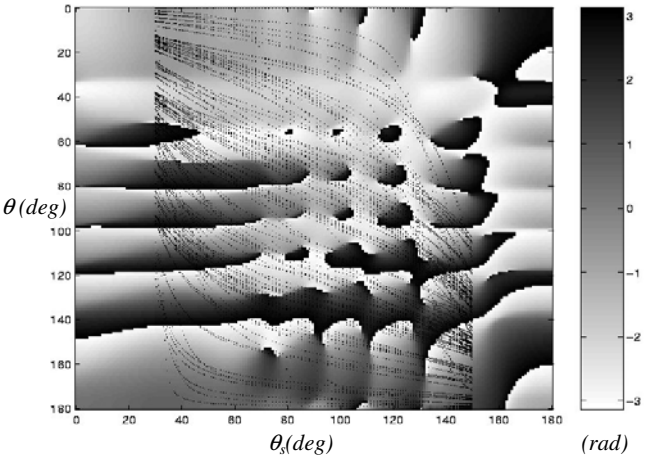


FIG. 10 : Phase de la SER du cylindre en polarisation croisée.

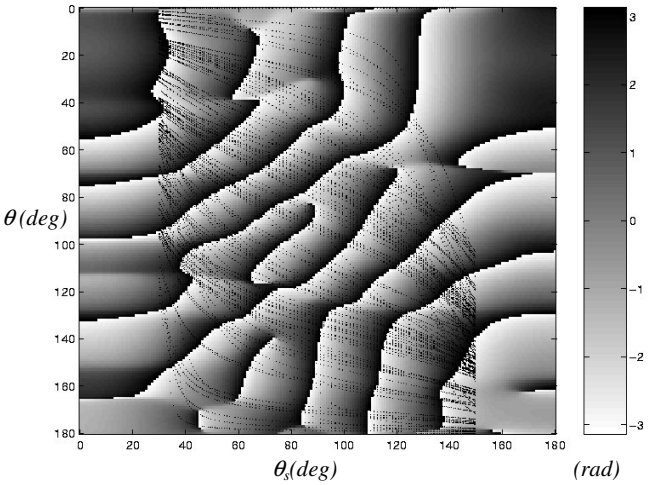


FIG. 11 : Phase de la SER de l'avion en polarisation croisée.

Un algorithme de Gradient Conjugué modifié pour le traitement Spatio-Temporel adaptatif en radar aéroporté

Hocine BELKACEMI[1], Hichem SEMIRA[2], Sylvie MARCOS[1]

[1]Laboratoire des Signaux et Systèmes (L2S) UMR-CNRS 8506
Supélec, 3 rue Joliot-Curie 91192 Gif-sur-Yvette France

[2]Département d'électronique
Université d'Annaba 23000 Algérie

belkacemi@lss.supelec.fr, se_hichem@yahoo.fr, marcos@lss.supelec.fr.

Résumé – L'algorithme du Gradient Conjugué (GC) a montré une efficacité équivalente à celle du filtrage de Wiener multi-étages (MSWF) [1, 2] pour compenser les interférences dans les systèmes de télécommunications. Dans cet article, nous proposons l'application du GC au problème du traitement spatio-temporel adaptatif en radar aéroporté. On présente aussi une version modifiée de l'algorithme du GC pour le filtrage du fouillis radar à faible support de données en utilisant le lissage *forward/backward* (f/b) pour l'estimation de la matrice de covariance. La version ainsi modifiée permet à la fois la compression en rang et la compression en support de données.

Abstract – The Conjugate Gradient (CG) algorithm has been shown to be as efficient as the Multistage Wiener Filter (MSWF) [1, 2] to compensate interferences in telecommunication systems. In this paper, we propose the application of the CG to interference suppression in Space-Time Adaptive Processing (STAP) radar. We also present a modified version of the algorithm for low sample support by using the forward/backward (f/b) averaging in the estimation of the covariance matrix. The new algorithm then allows both rank and sample support compression.

1 Introduction

La détection de cibles mobiles est l'un des objectifs essentiels de la télédétection notamment pour les applications radar. Les retours de la cible sont souvent mélangés avec de forts retours de fouillis de sol. Les performances en détection de cibles des radars MTI classiques se trouvent alors fortement dégradées par l'étalement Doppler des retours de fouillis, induits par le mouvement de la plateforme radar. Les cibles évoluant à faible vitesse peuvent, par conséquent, rester indétectables [3]. Le traitement adaptatif spatio-temporel (STAP) qui est une technique conçue pour résoudre ces problèmes inhérents aux techniques conventionnelles a récemment attiré l'attention de nombreux chercheurs dans la communauté radar. A la différence d'un formateur de voies classique (*beamformer*), le traitement STAP inclut une dimension temporelle et constitue un algorithme de filtrage bidimensionnel en angle et en Doppler.

Considérons un réseau spatio-temporel avec N capteurs uniformément espacés et M éléments de retard pour chaque capteur. Le vecteur de pondération optimal $\mathbf{w}_{opt}$ de dimension NM qui maximise le rapport signal à interférence ($SINR$) à la sortie du réseau est donné par [4] $\mathbf{w}_{opt} = \kappa \mathbf{R}_i^{-1}\mathbf{s}$, où κ est une constante, $\mathbf{R}_i$ est la matrice de covariance des interférences (foullis, brouilleurs et bruit thermique) de dimension $NM \times NM$ et $\mathbf{s}$ est le vecteur directionnel de la cible. Comme la matrice $\mathbf{R}_i$ est inconnue dans la pratique, Reed et al. [4] ont proposé la méthode dite *Sample Matrix Inversion* (SMI) basée sur l'estimée de la matrice de covariance des interférences $\hat{\mathbf{R}}_i$. En général, les algorithmes sont choisis selon deux critères : (i) convergence ; (ii) coût de calcul. La méthode SMI présente une faible vitesse de convergence ($2NM$ échantillons pour un SINR de 3dB au dessous de l'optimal) [4] et un coût de calcul élevé $O((NM)^3)$. Pour accélérer la convergence, les auteurs dans [5, 6] ont proposé des méthodes de sous-espaces basées sur la décomposition en valeurs propres (DVP) de $\mathbf{R}_i$. L'étude asymptotique de ces méthodes montre qu'on peut atteindre les mêmes performances que la méthode SMI avec seulement $2r$ échantillons où r et le rang de la matrice $\hat{\mathbf{R}}_i$, au lieu de $2NM$ avec un coût de calcul $O((NM)^3)$. Dans [7], Goldstein et *al.* ont proposé le filtrage de Wiener Multi-étages (MSWF). MSWF présente une compression du rang considérable comparé aux méthodes basées sur la DVP pour une vitesse de convergence comparable [8]. Récemment, l'équivalence entre le MSWF et le Gradient Conjugué a été établie [1] en se basant sur les sous-espaces de Krylov [9]. Dans cet article, nous proposons d'appliquer l'algorithme du gradient conjugué au traitement adaptatif spatio-temporel pour les radars aéroportés. Nous introduisons, en plus, une version modifiée du GC en utilisant le lissage *forward/backward* (f/b) [10] pour accélérer la convergence de l'algorithme. Le modèle des signaux est donné dans la section 2. Dans la section 3, on présente le principe du MSWF ainsi que quelques exemples de simulations. La section 4 introduit l'algorithme GC et la version modifiée. Enfin les performances de cette version sont testées dans la section 5, et la section 6 résume les principaux résultats.

2 Modéle des signaux

Le retour radar spatio-temporel $\mathbf{x}$, à une distance donnée est constitué de l'écho de la cible ainsi que d'interférences (fouillis, brouilleurs et bruit thermique) [3]

$$\mathbf{x} = \alpha_t \mathbf{v}_t + \mathbf{n} \tag{1}$$

où
- $\mathbf{x} = [x_1, ..., x_{MN}]^T$ est le vecteur des sorties du réseau spatio-temporel
- α_t et $\mathbf{v}_t \equiv \mathbf{v}(\varpi_t, \nu_t)$ sont le facteur d'atténuation et le vecteur directionnel de la cible respectivement, associés aux paramètres spatial ϖ_t et Doppler ν_t

$$\mathbf{v}(\varpi_t, \nu_t) = \mathbf{b}(\varpi_t) \otimes \mathbf{a}(\nu_t) \tag{2}$$

où
  - $\mathbf{a}(\nu_t) = [1 \quad e^{j2\pi\nu_t} \cdots \quad e^{j2\pi(M-1)\nu_t}]^T$ est le vecteur directionnel temporel $\nu_k = \frac{f_t}{f_r}$, f_t est la fréquence Doppler de la cible et f_r est la fréquence de transmission du radar.
  - $\mathbf{b}(\varpi_t) = [1 \quad e^{j2\pi\varpi_t} \cdots \quad e^{j2\pi(N-1)\varpi_t}]^T$ est le vecteur directionnel spatial ($\varpi_t = \frac{d}{\lambda} sin(\theta_t)$, d est la distance de séparation entre les éléments de l'antenne, λ est la longueur d'onde et θ_t est l'azimuth de la cible).
- $\mathbf{n} = [n_1, ..., n_{MN}]^T$

Le vecteur des interférences $\mathbf{n}$ est composé du fouillis $\mathbf{n}_c$ du bruit thermique $\mathbf{n}_b$ supposé blanc et de brouilleurs $\mathbf{n}_j$.

$$\mathbf{n} = \mathbf{n}_c + \mathbf{n}_b + \mathbf{n}_j \tag{3}$$

En supposant que ces composantes sont non corrélées, on a

$$\mathbf{R}_i = E\{\mathbf{n}\mathbf{n}^H\} = \mathbf{R_c} + \sigma^2 \mathbf{I}_{NM} + \sum_{i=1}^{N_j} \mathbf{R}_j(i) \tag{4}$$

où N_j est le nombre de brouilleurs et $\mathbf{R}_j(i)$ est la matrice de covariance du *i-ème* brouilleur, $\mathbf{R_c}$ la matrice de covariance du fouillis, σ^2 est la variance du bruit et $\mathbf{I}_{NM}$ la matrice identité de dimension $NM \times NM$.

3 Filtre à Multi-étages de Wiener

La matrice de covariance $\mathbf{R}_i$ est généralement de rang faible (voir figure 1(a)). Les algorithmes de sous-espaces exploitent cette propriété pour supprimer les interférences. L'idée de base est de décomposer l'espace total des observations en deux sous-espaces : le sous-espace interférence qui correspond aux vecteurs propres dominants et le sous-espace complémentaire appelé sous-espace bruit. Dans des environnements radar réels, la dimension du sous-espace interférence est affectée par plusieurs paramètres. Les figures 1(a) et 1(b) montrent, respectivement, l'effet de la présence de brouilleurs et le mouvement interne du fouillis radar (*internal clutter motion* ou ICM) sur la dimension du sous-espace interférence. Il est connu [5] que la sous-estimation du rang introduit des grandes pertes en SINR rendant inefficace les algorithmes basés sur la DVP. Pour pallier à ce problème, Goldstein et *al.* [11] ont proposé l'algorithme de Wiener à multi-étages (MSWF). Cet algorithme est basé sur la structure classique du filtre de Wiener.

La figure 2 illustre une structure standard du MSWF à 3 étages. En général, le MSWF revient à transformer la matrice de covariance $\mathbf{R}_{\mathbf{x}_0}$ (où $\mathbf{x}_0$ est l'observation à la sortie de la matrice du blocage $\mathbf{B}_0$) à rang complet en une matrice à rang réduit $\mathbf{R}_d$ où $\mathbf{d} = \mathbf{L}^H \mathbf{x}_0$ [12] (voir figure 3), la matrice résultante $\mathbf{R}_d$ est alors donnée par

$$\mathbf{R}_d = \mathbf{L}\mathbf{R}_{\mathbf{x}_0}\mathbf{L}^H \tag{5}$$

où $\mathbf{L}$ est la matrice de transformation

$$\mathbf{L}^H = \left[\begin{array}{cccc} \mathbf{h}_1 & \mathbf{B}_1^H\mathbf{h}_2 & \mathbf{B}_1^H\mathbf{B}_2^H\mathbf{h}_3 & \cdots \quad \prod_{i=1}^{r} \mathbf{B}_i^H\mathbf{h}_r \end{array} \right] \tag{6}$$

où les $\mathbf{h}_i = \frac{\mathbf{r}_{\mathbf{x}_{i-1}d_{i-1}}}{\|\mathbf{r}_{\mathbf{x}_{i-1}d_{i-1}}\|}$ sont définis comme les inter-corrélations normalisées entre le signal de référence d_{i-1} et le vecteur observé $\mathbf{x}_{i-1}$ de chaque étage avec le précédent et $\mathbf{B}_i = \mathbf{I} - \mathbf{h}_i\mathbf{h}_i^H$ sont les matrices de blocage. Le rang du filtre MSWF correspond au nombre d'étages retenus après une troncature à un rang r qui peut être choisi de manière adaptative [13].

Le vecteur de pondération $\mathbf{w}$ pour un filtre à rang r est donné par [11]

$$\mathbf{w} = \mathbf{s} - \mathbf{B}_0^H \mathbf{L} \mathbf{w}_{mswf} \tag{7}$$

$$\mathbf{w}_{mswf} = \mathbf{R}_d^{-1} \mathbf{r}_{\mathbf{d}d_0} \tag{8}$$

où $\mathbf{r}_{\mathbf{d}d_0} = E[\mathbf{d}d_0^H]$ avec $\mathbf{d} = [\ d_1 \quad d_2 \quad \cdots \quad d_r \ ]$

4 Algorithme du Gradient Conjugué

L'algorithme de gradient conjugué (GC) consiste à ramener la résolution du système linéaire de Wiener-Hopf à celui de la minimisation de la fonction de coût [9]

$$\phi(\mathbf{w}_{GC}) = \mathbf{w}_{GC}^H \mathbf{R}_{x_0} \mathbf{w}_{GC} - 2\Re(\mathbf{r}_{x_0 d_0} \mathbf{w}_{GC}) \tag{9}$$

La table 1 résume une version itérative du gradient conjugué où les $\mathbf{p}_i$ sont les vecteurs de recherche du GC qui sont $\mathbf{R}_{\mathbf{x}_0}$-conjugués

$$\mathbf{p}_i \mathbf{R}_{\mathbf{x}_0} \mathbf{p}_j = 0 \quad \forall i \neq j \tag{10}$$

$\mathbf{w}_i$ est le vecteur de pondération à l'itération i. Le GC conjugué est équivalent aussi à appliquer une transformation de réduction du rang $\mathbf{T}_{CG}$ à la matrice de covariance $\mathbf{R}_{\mathbf{x}_0}$, cette transformation est décrite par les vecteurs de recherche du GC $\mathbf{p}_i$

$$\mathbf{T}_{CG} = [\ \mathbf{p}_1, \mathbf{p}_2 \ \ \ldots, \mathbf{p}_D] \tag{11}$$

Une propriété du MSWF identifiée dans [1, 2] est que la matrice de transformation donnée par l'équation (6) et la matrice $\mathbf{T}_{CG}$ engendrent le même sous espace de Krylov $\mathcal{K}^D(\mathbf{r}_{x_0 d_0}, \mathbf{R}_{\mathbf{x}_0})$[1] entraînant l'équivalence des solutions du MSWF et du GC.

L'avantage du GC par rapport au MSWF est qu'on peut avoir le vecteur de pondération à chaque itération (étage) ce qui permet d'implanter un test d'arrêt (version adaptative). Pour accélérer la convergence du GC en support de données nous proposons d'estimer la matrice de covariance par un lissage (f/b) [10] comme résumé dans la table 2 où $\mathbf{R}_f(n)$ et $\mathbf{R}_b(n)$ sont les matrices de covariance *forward et backward* à l'itération n. Les avantages de notre approche sont que d'une part, on aura une compression du rang par le GC et, d'autre part, une compression en support de données.

5 Simulations et Résultats

Comme mesure de performances, on prend la perte en SINR ($SINR_{Loss}$) [3] à la sortie du réseau

$$SINR_{Loss} = \frac{|\mathbf{w}^H \mathbf{s}|}{\mathbf{w}^H \mathbf{R}_i \mathbf{w}} \tag{12}$$

Les simulations ont été réalisées sur 100 réalisations de Monte Carlo, l'antenne est linéaire à N=14 éléments et M=16 impulsions. Le fouillis radar est supposé gaussien

[1] $\mathcal{K}^D(\mathbf{r}_{x_0 d_0}, \mathbf{R}_{\mathbf{x}_0}) \equiv span\{\mathbf{r}_{x_0 d_0}, \mathbf{R}_{\mathbf{x}_0}\mathbf{r}_{x_0 d_0}, \ldots, \mathbf{R}_{\mathbf{x}_0}^{D-1}\mathbf{r}_{x_0 d_0}\}$

$\mathbf{w}_0(n) = 0,\ \mathbf{p}_1(n) = \mathbf{g}_0(n) = \mathbf{r}_{\mathbf{x}_0\mathbf{d}_0}$
$\rho_0(n) = \mathbf{g}_0(n)^H \mathbf{g}_0(n)$
$\mathbf{R}(n) = \mathbf{R}(n-1) + \mu \mathbf{x}(n)\mathbf{x}^H(n)$

for $i = 1$ to D do
$\mathbf{z}_i(n) = \mathbf{R}(n)\ \mathbf{p}_i(n)$
$\alpha_i(n) = \frac{\rho_i(n)}{\mathbf{p}_i(n)^H \mathbf{z}_i(n)}$
$\mathbf{w}_i(n) = \mathbf{w}_{i-1}(n) + \alpha_i(n)\mathbf{p}_i(n)$
$\mathbf{g}_i(n) = \mathbf{g}_{i-1}(n) - \alpha_i(n)\mathbf{z}_i(n)$
$\rho_i(n) = \mathbf{g}_i(n)^H \mathbf{g}_i(n)$
$\delta_i(n) = \frac{\rho_i(n)}{\rho_{i-1}(n)}$
$\mathbf{p}_{i+1}(n) = \mathbf{g}_i(n) + \delta_i(n)\mathbf{p}_i(n)$
End for

TAB. 1 – Algorithme du Gradient Conjugué itératif

$\mathbf{R}_f(n) = \mathbf{R}_f(n-1) + \mu \mathbf{x}_f(n)\mathbf{x}_f^H(n)$
$\mathbf{R}_b(n) = \mathbf{R}_b(n-1) + \mu \mathbf{x}_b(n)\mathbf{x}_b^H(n)$
$\mathbf{R}_{fb}(n) = 0.5(\mathbf{R}_b(n) + \mathbf{R}_f(n))$

TAB. 2 – Estimation de $\mathbf{R}_i$ par lissage *Forward/Backward*

avec un rapport fouillis sur bruit CNR=40 dB en présence de 4 brouillleurs avec les paramètres suivants : rapport brouilleur à bruit JNR=[40 30 40 30] dB et angles azimuth [45° -30° -45° 60°]. Le signal désiré est à 0° en azimuth avec un SNR=0 dB.

Dans les figures 4 et 5, on compare les performances du MSWF pour une matrice de covariance théorique connue avec les algorithmes basés sur la DVP (composantes principales (PC) et Cross spectral Metric (CSM)). Dans les deux cas, présence de brouilleurs ou ICM, les figures montrent bien la compression en rang du MSWF.

La figure 6 montre une comparaison des trois algorithmes MSWF, GC itératif comme décrit dans la table 1 et PC en fonction du rang. On constate que le gain apporté par le GC en SINR est important par rapport au PC et que le GC maintient ce gain lorsque le rang est surestimé. La figure 7 montre bien que les performances du GC avec lissage f/b dépassent celles du GC en terme de support de données.

6 Conclusion

Nous avons montré dans cet article que les algorithmes à rang réduit convergent plus vite que les algorithmes basés sur la décomposition en valeur propres de la matrice de covariance (PC, CSM). Nous avons appliqué aussi le gradient conjugué dans le contexte du STAP, les résultats de simulations ont montré l'efficacité de cet algorithme et sa robustesse vis à vis du rang et nous avons montré aussi que l'on peut améliorer la convergence du gradient conjugué en terme du nombre de données en utilisant le lissage f/b.

Références

[1] G. Dietel, M. Zoltowski, and M. Joham, "Reduced-rank equalization for EDGE via conjgate gardient implementation of multi-stage nested Wiener filter," *In Proc. of 54th IEEE Vehicular Technology Conference.*

[2] S. Burykh and K. Abed-Meraim, "Reduced-rank adaptive filtering using Krylov subspace," *EURASIP JASP*, vol. 2002, no. 12, pp. 1387–1400, Dec. 2002.

[3] J. Ward, "Space-time adaptive processing for airborne radar," Lincolin Laboratory," Technical report 1015, December 1994.

[4] L. E. Brennan and I. S. Reed, "Theory of adaptive radar," *IEEE Transaction on Aerospace and Electronic Systems*, vol. AES-9, pp. 237–252, Mar. 1973.

[5] A. M. Haimovitch, "Eigenanalysis-based space-time adaptive radar : Performance analysis," *IEEE Transaction on Aerospace and Electronic Systems*, vol. 33, pp. 1170–1179, October 1997.

[6] J. Goldstein and I. Reed, "Reduced-rank adaptive filtering," *IEEE Transactions on Signal Processing*, vol. 45, no. 2, pp. 492–496, Feb. 1997.

[7] J. Goldstein, I. Reed, and P. Zulch, "Multistage partially adaptive STAP CFAR detection algorithm," *IEEE Transaction on Aerospace and Electronic Systems*, vol. 35, no. 2, pp. 645–662, April 1999.

[8] J. Hiemstra, "Robust implementation of the multistage Wiener filter," Ph.D. dissertation, Virginia Polytechnic Institute and State University, April 2003.

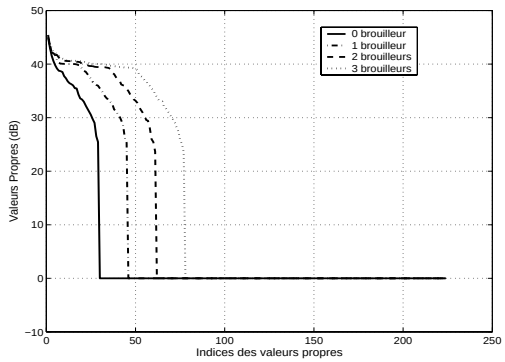
[9] G. Golub and C. V. Loan, *Matrix computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, 1996.

[10] M. Zatman and D. Marshall, "Forward-backward averaging in the presence of array manifold errors," *IEEE Transactions on Antennas and Propagation*, vol. 46, no. 11, pp. 1700–1704, Nov 1998.

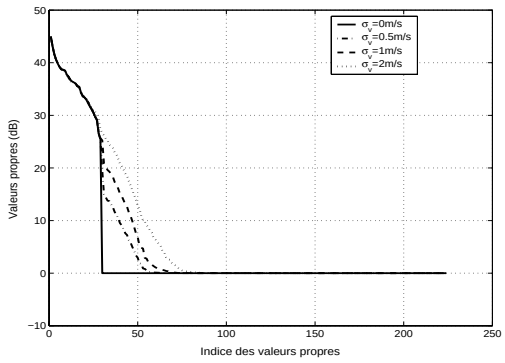
[11] J. Goldstein, I. Reed, and L. Sharf, "A multistage representation of Wiener filter based on orthogonal projections," *IEEE Transaction on Aerospace and Electronic Systems*, vol. 44, no. 7, pp. 2943–2959, Nov. 1998.

[12] J. Guerci, J. Goldstein, and I. Reed, "Optimal and adaptive reduced-rank STAP," *IEEE Transaction on Aerospace and Electronic Systems*, vol. 36, no. 2, pp. 647–661, April 2000.

[13] M. Honig and J. Goldstein, "Adaptive reduced-rank interference suppression based on the multistage Wiener filter," *IEEE Transactions on Communications*, vol. 50, no. 6, pp. 986–994, June 2002.



(a)



(b)

FIG. 1 – Paramètres affectant la dimension du sous-espace interférence (a) Effet des brouilleurs (b) Effet du mouvement interne du fouillis (ICM)

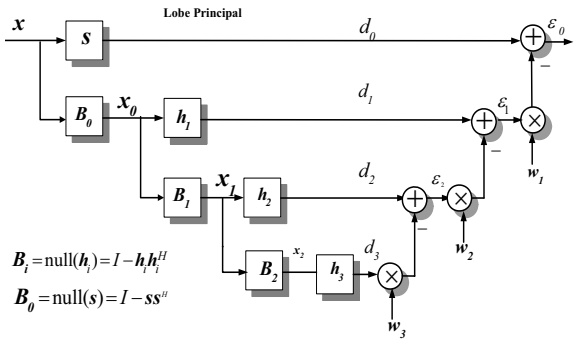


FIG. 2 – Illustration d'un Filtre de Wiener à 3 étages

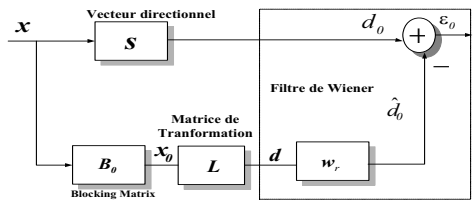


FIG. 3 – Illustration du Filtrage à rang réduit

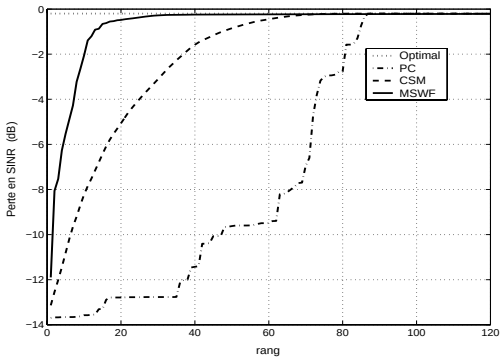


FIG. 4 – Perte en SINR en fonction du rang pour une matrice de covariance connue avec les paramètres suivants (a) En présence de 4 brouilleurs à [-60° -30° 45° 60°] avec JNR [40 30 40 30] dB respectivement

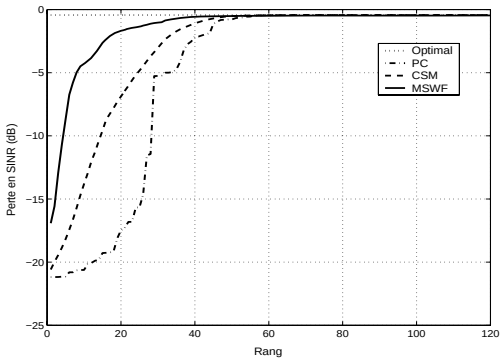


FIG. 5 – Perte en SINR en fonction du rang pour une matrice de covariance connue en présence du mouvement interne du fouillis (ICM) avec une vitesse du vent $\sigma_v = 1m/s$.

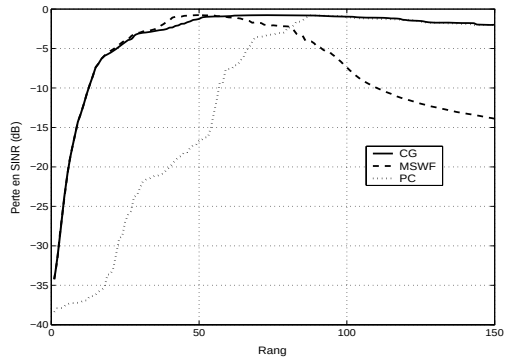


FIG. 6 – Perte en SINR en fonction du rang pour PC, MSWF et GC itératif pour 200 échantillons

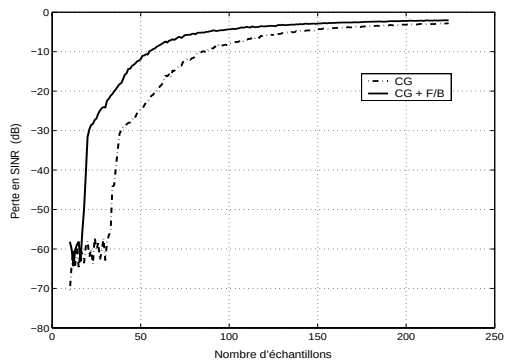


FIG. 7 – Perte en SINR en fonction du nombre d'échantillons pour GC et GC avec lissage (f/b), r=80

Antibrouillage RADAR en contexte de rotation d'antenne

Marc OUDIN[1], Jean Pierre DELMAS[1], Claude ADNET[2], Cécile GERMOND[2], Frédéric BARBARESCO[2]

[1]GET/INT, Département CITI, UMR-CNRS 5157, 9 rue Charles Fourier, 91011 Evry Cedex, France

[2]Thales/TAD/JRS/RBPEF, 7/9 rue des Mathurins, 92221 Bagneux Cedex, France

marc.oudin(jean-pierre.delmas)@int-evry.fr,
claude.adnet(cecile.germond,frederic.barbaresco)@fr.thalesgroup.com

Résumé – Nous nous intéressons dans cet article à l'algorithme Opposition dans les Lobes Secondaires (OLS) utilisé en traitement d'antenne RADAR dans un but d'antibrouillage en contexte de rotation d'antenne. Nous montrons que l'utilisation de la méthode Extended Sample Matrix Inversion (ESMI) [2] conduit alors à des performances supérieures à la méthode standard.

Abstract – In this paper, we focus on the Sidelobe Canceller (SLC) algorithm applied for RADAR with rotating antenna. We show that the ESMI approach leads to better performance than the standard one.

1 Introduction

Le fonctionnement d'un grand nombre d'applications : télécom - munications, RADAR, sonar, se trouve perturbé par la présence d'interférences. [1] Des techniques de traitement d'antenne adaptatives consistant à utiliser un réseau de capteurs et à pondérer les signaux en sortie de ces derniers par des coefficients dépen - dant des données sont souvent utilisées. Elles permettent de réduire l'influence des nuisances sur le signal reçu.

Cependant pour un RADAR en mode veille scrutant l'espace, ce problème peut se trouver compliqué par une rotation du réseau de capteurs qui rend l'environnement non stationnaire [1].

Dans [2], Hayward propose la méthode ESMI basée sur un développement limité du filtre spatial optimal. Il s'agit d'une extension de la méthode SMI présentée dans [3]. Les performances asymptotiques en terme de SINR sont ensuite étudiées dans [4]. Dans cet article, nous appliquons cette méthode à l'approche OLS (c.f. Section 3), en nous basant sur un modèle physique de rotation d'antenne. Par rapport au modèle proposé dans [2], notre modèle tient compte de la forme des gains d'antenne. Nous étudions les performances de cette méthode en terme de réduction de la puissance résiduelle de nuisance.

Dans la section 2 de cet article, nous proposons une modélisa - tion mathématique simplifiée du problème en se plaçant dans le cas où le réseau de capteurs est composé de deux voies. La première voie a un diagramme de rayonnement directif alors que l'autre a un diagramme large. Puis, dans la section 3, nous étudions le principe de l'algorithme étudié avant d'en faire une étude de performances théoriques dans la section 4. Enfin, nous illustrons les résultats par des simulations dans la partie 5.

2 Modélisation du problème

2.1 Signal utile, interférences et bruit thermique

Considérons une antenne composée de deux voies. Ces voies peuvent par exemple correspondre à un regroupement analogique de capteurs élémentaires. La première voie (voie principale) possède un diagramme de rayonnement directif alors que la seconde (voie auxiliaire) a un diagramme de rayonnement large.

L'environnement est constitué d'un signal utile et d'un signal de nuisance provenant tous deux de réflecteurs fixes. Le signal utile $s(t)$ est complexe déterministe de puissance σ_s^2. Par contre, le signal perturbateur $b(t)$ est modélisé par un processus stochastique du second ordre complexe centré de puissance σ_b^2. Les signaux utiles et perturbateurs sont supposés à bande étroite.

Les deux voies génèrent un bruit thermique interne ($n_1(t)$ et $n_2(t)$) venant se rajouter au signal reçu. Ces bruits thermiques sont modélisés par un processus stochastique du second ordre gaussien complexe centré de puissance σ_n^2.

2.2 Modélisation de la rotation d'antenne

On se place dans un problème plan dans lequel le réseau d'antennes est animé d'un mouvement de rotation uniforme à la vitesse ω rad/s. Les gains des capteurs sont supposés connus et constants dans le temps. Ils sont fonction uniquement de l'angle sous lequel est vu l'élément réflecteur.

On note $G_p(\theta)$ et $G_a(\theta)$ les gains complexes respectifs des voies principale et auxiliaire pour un signal reçu sous l'angle θ. Afin de simplifier le problème, on suppose que l'antenne est focalisée initialisement dans la direction 0 de la cible. On note θ_b l'angle initial du signal perturbateur. A l'instant t, les angles sous lesquels sont vus la cible ($\theta_c(t)$) et la nuisance ($\theta_b(t)$) sont donc :

$$\theta_c(t) = \omega t \quad \text{et} \quad \theta_b(t) = \omega t + \theta_b$$

Finalement, les signaux résultants en sortie des voies principale et auxiliaire peuvent s'écrire respectivement :

$$V_p(t) = G_p(\omega t)s(t) + G_p(\omega t + \theta_b)b(t) + n_1(t) \qquad (1)$$

$$V_a(t) = G_a(\omega t)s(t) + G_a(\omega t + \theta_b)b(t) + n_2(t) \qquad (2)$$

1. Ce travail est réalisé dans le cadre d'une thèse DGA/Thales Air Defence

Il est important de noter que le seul paramètre inconnu de ces équations est l'angle initial de la nuisance θ_b. On dispose dans la suite des données correspondant à tous les échantillons d'une période d'écoute RADAR : $V_p(t)$ et $V_a(t)$ pour $t_1 \leq t \leq t_N$.

3 Description de l'algorithme OLS

3.1 Principe de l'algorithme

L'algorithme OLS soustrait au signal sur la voie principale une pondération du signal sur la voie auxiliaire choisie de telle sorte à minimiser l'influence de la perturbation sur le signal résultant. Il peut fonctionner dans la mesure où la contribution du signal utile soustraite à la voie principale est négligeable devant le signal utile sur cette même voie. On se ramène alors à un problème de référence bruit seul. Les conditions d'application de l'algorithme sont donc la voie principale focalisée dans la direction utile, le signal perturbateur absent du lobe principal et le gain sur la voie auxiliaire supérieur au gain de la voie principale au niveau de ses lobes secondaires.

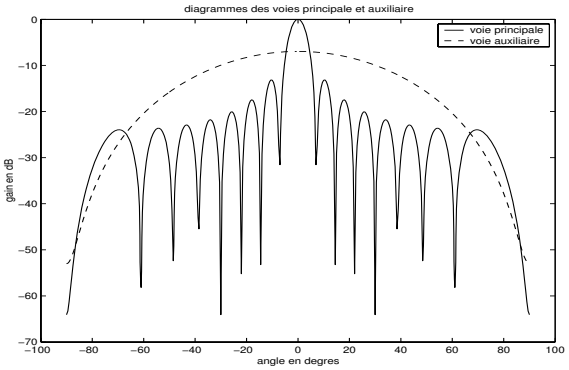


Fig.1 : Diagrammes des voies principale et auxiliaire

Le critère utilisé dans cette méthode est l'erreur quadratique moyenne entre le signal sur la voie principale et la pondération du signal sur la voie auxiliaire.

L'algorithme OLS standard consiste à rechercher le coefficient α qui minimise l'expression : $E|V_p(t) - \alpha^* V_a(t)|^2$.

Le coefficient solution de ce problème vérifie alors l'équation :

$$\alpha(t) = E(|V_a(t)|^2)^{-1} E(V_a(t) V_p^*(t))$$

L'approche OLS optimale serait donc de calculer un coefficient à chaque instant. Cependant, cela n'est pas réalisable en pratique d'une part parce que les espérances sont estimées par une moyenne empirique sur plusieurs instants et d'autre part parce qu'une telle approche serait trop lourde à mettre en oeuvre.

3.2 Approche avec coefficients variables dans le temps

Cette approche a pour objectif de décomposer le coefficient recherché sur une base polynômiale de façon à tenir compte des variations temporelles du coefficient optimal [2]. L'estimation à effectuer revient alors à celle des coefficients dans la base choisie.

Supposons que l'on fasse une décomposition du coefficient sur la base de polynômes canonique contenant $K+1$ éléments :

$$\alpha(t) = \sum_{k=0}^{K} \alpha_k (t - t_0)^k$$

Les coefficients à estimer sont maintenant les coefficients $(\alpha_k)_{k=0..K}$. t_0 est un instant de la durée d'écoute choisi de façon à optimiser les performances en terme de puissance résiduelle. Dans toute la suite de l'article, on fait l'hypothèse selon laquelle la variation de gain est suffisamment faible pour que l'on puisse se limiter à $K = 1$,

3.3 Implémentation de l'algorithme

L'algorithme OLS peut être mis en oeuvre de différentes manières. En effet, l'utilisateur dispose de toutes les données correspondant aux cases distance de la récurrence RADAR et doit choisir les échantillons à utiliser dans le calcul du coefficient. Ce dernier peut de plus être calculé une ou plusieurs fois. Comme le choix des échantillons n'est pas le propos de cet article, nous supposerons dans la suite que le coefficient n'est calculé qu'une seule fois, à partir de tous les échantillons de la récurrence.

4 Etude de performances

4.1 Critère de performances retenu

On cherche à comparer le gain en performances résultant de l'utilisation de l'algorithme OLS de coefficient variable dans le temps. Pour cela, nous retenons comme critère de performances la puissance résiduelle de nuisance instantanée. Ainsi, l'algorithme est d'autant plus efficace que la puissance résiduel - le de nuisance est faible.

D'après (1) et (2), la puissance résiduelle de nuisance à l'instant t s'écrit :

$$P_{res}(t) = |G_p(\omega t + \theta_b) - \alpha^*(t) G_a(\omega t + \theta_b)|^2 \, \sigma_b^2 \quad (3)$$

4.2 Expression du critère de performances

Nous calculons dans cette partie l'expression de la puissance résiduelle de nuisance pour l'algorithme OLS standard puis pour sa version avec coefficients variables dans le temps. Pour aboutir à ces deux expressions, nous effectuons un développe - ment limité déterministe du gain sur la voie principale et une approximation sur des fonctions du signal perturbateur qui interviennent dans le calcul. Dans les deux cas, nous calculons la puissance en fonction de l'instant t_l avec $1 \leq l \leq N$.

4.2.1 Version standard

Le coefficient OLS est calculé par moyenne empirique selon la méthode SMI [3] :

$$\alpha = (\sum_{k=1}^{N} |V_a(t_k)|^2)^{-1}) (\sum_{k=1}^{N} V_a(t_k) V_p^*(t_k)) \quad (4)$$

Afin de simplifier cette étude, nous effectuons les hypothèses préalables suivantes :

– le gain de la voie auxiliaire est constant ($G_a(\theta) = G_a$)

– le signal utile est négligé devant le signal perturbateur : $\sigma_s^2 \ll \sigma_b^2$
– le bruit thermique est négligé devant le signal perturbateur : $\sigma_n^2 \ll \sigma_b^2$

Et en utilisant (1), (2) avec ces hypothèses dans (4) :

$$\alpha = (\sum_{k=1}^{N} |G_a|^2 |b(t_k)|^2)^{-1} (\sum_{k=1}^{N} G_a G_p^*(\omega t_k + \theta_b) |b(t_k)|^2) \tag{5}$$

Introduisons les notations suivantes :

$$S = \sum_{k=1}^{N} |b(t_k)|^2 \quad \text{et} \quad T = \sum_{k=1}^{N} |b(t_k)|^2 G_p(\omega t_k + \theta_b) \tag{6}$$

La puissance résiduelle de nuisance peut alors s'exprimer de la manière suivante en introduisant (5) et (6) dans (3) :

$$P_{res}^{(1)}(t_l) = \left| G_p(\omega t_l + \theta_b) - \frac{T}{S} \right|^2 \sigma_b^2 \tag{7}$$

Notons que même si le coefficient OLS est constant sur l'ensemble des cases distances, la puissance résiduelle ne l'est pas car elle tient compte du gain de nuisance à l'instant étudié. Ainsi, la méthode d'antibrouillage peut être plus ou moins efficace selon la case distance traitée.

4.2.2 Version à coefficients variables dans le temps

Cherchons maintenant à exprimer la puissance résiduelle de nuisance dans le cas de l'algorithme OLS à coefficient variable dans le temps. Le coefficient est maintenant estimé par l'expression :

$$\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = (\sum_{k=1}^{N} \underline{V}_a(t_k) \underline{V}_a^H(t_k))^{-1} (\sum_{k=1}^{N} \underline{V}_a(t_k) V_p^*(t_k))$$

$$\alpha(t_l) = \alpha_0 + \alpha_1 (t_l - t_0)$$

où $\underline{V}_a(t_k) = \begin{pmatrix} V_a(t_k) \\ (t_k - t_0) V_a(t_k) \end{pmatrix}$.

Afin de ne pas trop alourdir les expressions à venir, nous introduisons les notations suivantes :

$$U = \sum_{k=1}^{N} (t_k - t_0) |b(t_k)|^2 \quad \text{et} \quad V = \sum_{k=1}^{N} (t_k - t_0)^2 |b(t_k)|^2 \tag{8}$$

$$W = \sum_{k=1}^{N} (t_k - t_0) |b(t_k)|^2 G_p(\omega t_k + \theta_b) \tag{9}$$

$$\text{et} \quad X = \sum_{k=1}^{N} (t_k - t_0)^3 |b(t_k)|^2$$

Avec ces notations et les hypothèses simplificatrices de la section 4.2.1, l'expression du coefficient OLS devient :

$$\begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix} = (|G_a|^2 \begin{pmatrix} S & U \\ U & V \end{pmatrix})^{-1} (G_a \begin{pmatrix} T^* \\ W^* \end{pmatrix}) \tag{10}$$

On en déduit la puissance résiduelle de nuisance après OLS avec coefficient variable dans le temps en introduisant (8) et (9) et en remplaçant (10) dans (3) :

$$P_{res}^{(2)}(t_l) = |G_p(\omega t_l + \theta_b) - \frac{((VT - UW) + (SW - UT)(t_l - t_0))}{SV - U^2} \Big|^2 \sigma_b^2 \tag{11}$$

De même que pour la méthode OLS classique, la puissance résiduelle de nuisance est fonction de l'instant auquel on la calcule. Cependant, on remarque dans cette expression que le terme soustrait au gain de nuisance sur la voie principale est également fonction du temps.

4.2.3 Simplification des expressions de puissance résiduelle de nuisance

Afin de comparer les expressions du critère (7) et (11), nous procédons maintenant à un développement limité (DL) du gain du signal perturbateur sur la voie principale. Ce DL est réalisé autour de l'instant traité t_l et se justifie dans la mesure où la variation du gain sur la voie principale est faible pendant la durée d'observation.

Comme le montre la figure 2 suivante où est représentée la variance du gain principal pendant la récurrence, en fonction de la valeur de l'angle initial du signal perturbateur θ_b, l'hypothèse est valable lorsque le signal perturbateur n'est pas vu initialement dans un creux de diagramme. Les valeurs numériques choisies sont telles que $\omega T_{rec} = 0.036 deg.$

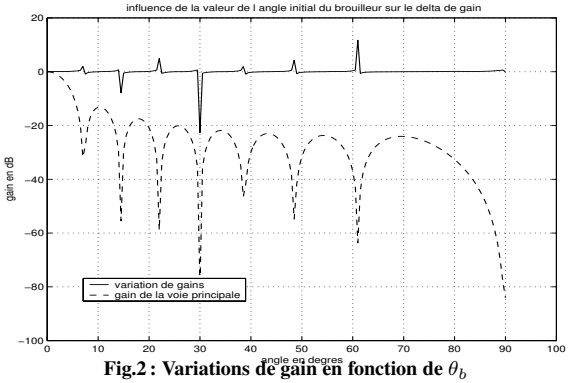


Fig.2 : Variations de gain en fonction de θ_b

Ce DL s'effectue au premier ordre pour $P_{res}^{(1)}$, mais au deuxiè - me ordre pour $P_{res}^{(2)}$, le DL à l'ordre 1 de cette expression étant nul.

Pour continuer à simplifier les équations (7) et (11), nous effectuons maintenant une approximation asympotique des fonctions de nuisance correspondant aux équations (8) et (9).

La convergence en moyenne quadratique des expressions suivantes permet de justifier ces approximations (avec $\rho = \frac{t_0}{NT_e}$):

$$\frac{S}{N} \approx \sigma_b^2 \quad \text{et} \quad \frac{U}{N^2 T_e} \approx (\frac{1}{2} - \rho) \sigma_b^2 \tag{12}$$

$$\frac{V}{N^3 T_e^2} \approx (\frac{1}{3} - \rho + \rho^2) \sigma_b^2 \tag{13}$$

$$\frac{X}{N^4 T_e^3} \approx (\frac{1}{4} - \rho + 3/2\rho^2 - \rho^3) \sigma_b^2 \tag{14}$$

En utilisant (12) dans (7), on obtient un équivalent sur la puissance résiduelle de nuisance avec l'algorithme OLS standard :

$$P_{res}^{(1)}(t_l) \approx \left| \frac{\omega T_e}{2}(N-2l)\dot{G}_p(\omega t_l + \theta_b) \right|^2 \sigma_b^2 \qquad (15)$$

Puis, en utilisant (12), (13) et (14) dans (11) et après quelques simplifications :

$$P_{res}^{(2)}(t_l) \approx \left| \frac{\omega^2 T_e^2}{2}(\frac{N^2}{6} - lN + l^2)\ddot{G}_p(\omega t_l + \theta_b) \right|^2 \sigma_b^2 \quad (16)$$

Nous remarquons que cette dernière expression ne dépend pas de t_0, ce qui veut dire que le choix de cette valeur n'a pas d'influence sur les performances de l'algorithme.

5 Simulations

Les simulations sont réalisées à partir d'une voie principale formée grâce à une antenne ALU de 16 capteurs et d'une voie auxiliaire à partir d'une antenne ALU de 2 capteurs. Ces deux réseaux respectent la condition de Nyquist et ont le même centre de phase. On choisit comme durée d'observation une période de récurrence RADAR. Les puissances sont exprimées en dB.

TAB. 1: Paramètres de la simulation

N	T_{rec}(ms)	ω(rd/s)	σ_b^2	$G_p(0)$	$G_a(0)$	N_{mc}
100	0.1	2π	50	0	-7	100

5.1 Comparaison des performances des deux versions de l'algorithme OLS

Les figures suivantes comparent les puissances résultantes des deux versions de l'algorithme OLS obtenues par Monte Carlo sur N_{mc} réalisations et par les formules théoriques (15) et (16). On observe une nette diminution de la puissance résiduelle moyenne pour la version avec coefficients variables dans le temps par rapport à la version standard.

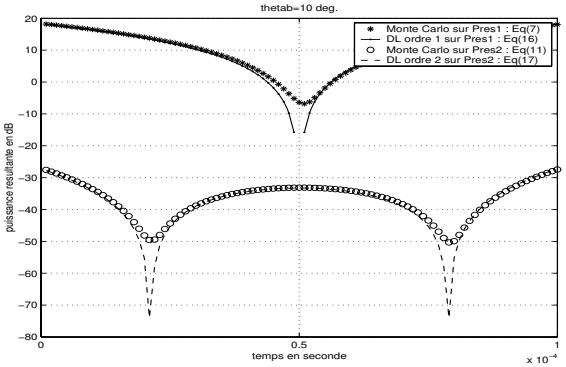


Fig.3 : $\theta_b = 10deg$

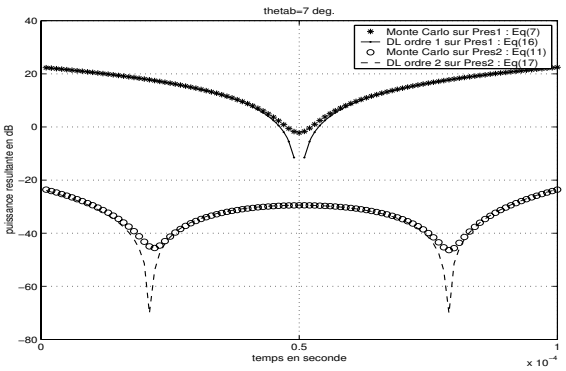


Fig.4 : $\theta_b = 7deg$

On note que même en zone de gain de la voie principale fortement variable (c.f. figure 2 avec $\theta_b = 7deg$), les courbes théoriques et de Monte Carlo sont très proches. On vérifie que les courbes de $P_{res}^{(1)}$ montrent un creux pour $l = \frac{N}{2}$, alors que celles de $P_{res}^{(2)}$ montrent un creux pour les valeurs de l racines de $(\frac{N^2}{6} - lN + l^2)$.

5.2 Robustesse du modèle par rapport aux hypothèses

On étudie maintenant la robustesse des résultats théoriques aux erreurs de modèle.

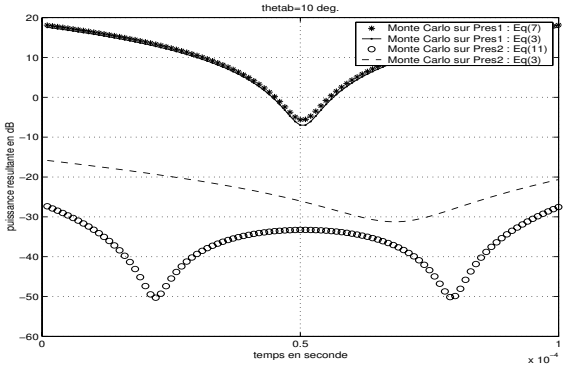


Fig.5 : $\theta_b = 10deg$, $\sigma_n^2 = -20dB$, $\sigma_s^2 = -10dB$

On remarque que l'expression théorique de $P_{res}^{(2)}$ est moins robuste que celle de $P_{res}^{(1)}$, mais que les performances de la version à coefficients variables restent sensiblement meilleures que celles de la version standard.

6 Conclusion

Nous avons vu dans cet article que l'algorithme OLS avec coefficients variables dans le temps appliqué sur une période d'écoute RADAR donne des performances d'antibrouillage meilleures que la version standard en configuration antenne tournante. Nous avons établi une expression du critère de puissance résiduelle de nuisance dans le cas où l'on dispose de deux voies de réception et montré que les performances de la version variable dans le temps ne dépendaient pas du choix de t_0.

Références

[1] S.D. Hayward, "Effects of motion on adaptive arrays", IEEE Proc. RSN, February 1997.

[2] S.D. Hayward, "Adaptive beamforming for rapidly moving arrays", Int. Conf. on Radar, Beijing, 1996.

[3] I.S. Reed, J.D. Mallet, L.E. Brennan, "Rapid convergence rate in adaptive arrays", IEEE Trans. Aerospace and Electronic Systems, Vol. AES-10, no 6, Nov. 74.

[4] M. Zatman, "The properties of adaptive algorithms with time varying weights", IEEE SAM 2000 Workshop, Boston 2000.

Analyse de la Texture dans les Images SAR Appliquée au Filtrage de Speckle

Olivier D'HONDT, Laurent FERRO-FAMIL, Eric POTTIER

I.E.T.R - Groupe Image et Télédétection
Campus de Beaulieu - Bat 11D, 236 Avenue du Général Leclerc, CS 74205, 35042, Rennes, France
olivier.dhondt@univ-rennes1.fr, laurent.ferro-famil@univ-rennes1.fr,
eric.pottier@univ-rennes1.fr.

Résumé – Cet article propose un nouveau modèle pour les statistiques d'ordre deux de la texture spatiale dans les images SAR. La fonction d'auto-covariance est approximée localement par un noyau gaussien anisotrope (AGK) 2-D dans le but de décrire la texture en termes d'orientation et d'anisotropie locales. L'estimation des paramètres de la texture à une échelle donnée est basée sur l'opérateur nommé tenseur de structure et ne requiert pas l'estimation explicite de l'auto-covariance. Enfin, un nouveau filtre nommé AGK-MMSE prenant en compte cette information spatiale est introduit et comparé au filtre MMSE géometrique classique. Le filtre proposé présente de meilleures performances en termes de préservation de la texture et d'amélioration des structures.

Abstract – This paper proposes a new model for the two-point statistics of spatial texture in SAR images. The autocovariance function is locally approximated by a 2-D Anisotropic Gaussian Kernel (AGK) in order to characterise texture by its local orientation and anisotropy. The estimation of texture parameters at a given scale is based on the Gradient Structure Tensor (GST) operator and does not require the explicit computation of the autocovariance. Finally, a new filter called AGK-MMSE that takes into account this spatial information is introduced and compared to the refined MMSE filter. The proposed filter shows better performances in terms of texture preservation and structure enhancement.

1 Introduction

Cet article présente un nouveau modèle pour les statistiques d'ordre deux de l'intensité des images SAR. Les notions d'orientation et d'anisotropie locales sont introduites par l'intermediaire d'une forme paramétrique de l'auto-covariance bidimensionnelle. Ce modèle simple permet de décrire aussi bien les structures déterministes que l'auto-correlation spatiale d'un foullis non-homogène. Il est enfi n montré que ce type de description peut être utilisée afi n d'améliorer les performance des fi ltres de speckle conventionnels.

Le modèle théorique est introduit dans la section 2, puis une méthode d'estimation des paramètres du modèle est présentée section 3. Une amélioration des fi ltres de speckle basée sur l'utilisation de l'information spatiale est présentée section 4. La section 5 montre l'application aux données réelles. Enfi n, la section 6 présente les conclusions de cet article.

2 Modèle de Texture Spatiale

2.1 Le Modèle Multiplicatif

Il est maintenant bien connu que l'intensité des images SAR suit un modèle multiplicatif. C'est pourquoi nous adopterons la décomposition donnée dans [8], qui suppose que pour la position spatiale t, l'intensité I peut être écrite sous la forme suivante :

$$I(t) = \mu(t)T(t)F(t) \quad (1)$$

où μ est la moyenne locale de la refléctivité liée aux propriétés physiques du terrain, T et F sont deux processus aléatoires décrivant respectivement la texture et le speckle. A cause du speckle, la reflectivité a l'intérieur d'un pixel n'est pas directement observable. Cependant, une zone est considerée homogène si ses paramètres statistiques ne varient pas spatiallement, ce qui permet d'estimer la réflectivité réelle par un moyennage spatial. Si une zone présente des variations importantes de réflectivité, une moyenne spatiale sur une échelle plus grande que les fluctuations en question entraîne une perte d'information sur les propriétés structurelles du milieu. Afi n de décrire au mieux ces propriétés, la variable T représentant ces fluctuations doit être prise en compte. En particulier, les fluctuations spatiales dues à la texture pourront être décrites par des statistiques d'ordre deux, telles que la fonction d'auto-covariance.

2.2 Statistiques d'ordre deux

2.2.1 Descripteurs 1-D

Le modèle multiplicatif inroduit par l'equation (1) peut être vu localement comme un modèle contenant trois échelles distinctes : l'échelle la plus fi ne correspond au speckle alors que la plus grossière correspond a la réflectivité moyenne, localement constante. La texture est alors considérée comme une variation spatiale à une échelle intermédiaire entre les deux premières. Si l'on considère que l'information spatiale est décrite par les statistiques d'ordre deux, l'intensité peut etre caractérisée par sa fonction d'auto-corrélation, défi nie pour un vecteur 2-D τ par :

$$R_I(\tau) = E\left[I(t+\tau)I(t)\right]. \quad (2)$$

On peut aussi considérer l'auto-covariance, qui est indépendante de la moyenne :

$$C_I(\tau) = R_I(\tau) - \mu_I^2 \tag{3}$$

où encore le coeffi cient de corrélation : $\rho_I(\tau) = C_I(\tau)/\sigma_I^2$ dont les valeurs sont comprises entre zéro et l'unité. Différents types de textures peuvent ainsi être discriminés par le type de la fonction d'auto-corrélation (exponentielle, gaussienne, etc) et par la longeur de corrélation l, défi nie par la relation $\rho_I(l) = e^{-1}$. Pareillement, le signal peut etre representé dans le domaine fréquentiel par sa densité spectrale de puissance :

$$S_I(f) = \int_{-\infty}^{+\infty} R_I(\tau) e^{-2\pi j f \tau} d\tau. \tag{4}$$

2.2.2 Le modèle AGK, une nouvelle déscription de la texture SAR 2-D

La fonction d'auto-covariance et le spectre peuvent prendre de multiples formes selon les propriétés structurelles du milieu rencontré. C'est pourquoi l'on choisit une représentation paramétrique simple pour ces fonctions, afi n de généraliser la notion de longueur de correlation au cas 2-D. Nous avons montré dans de précédentes publications [3, 2] qu'un modèle permettant ce type de description est le noyau gaussien anisotrope ou AGK (Anisotropic Gaussian Kernel), qui suppose que l'auto-covariance est de la forme :

$$C_T(\mathbf{d}) = \sigma_T^2 \exp\left(-\mathbf{d}^T \mathbf{\Sigma}^{-1} \mathbf{d}\right) \tag{5}$$

où $\mathbf{d} = [x, y]^T$ est la position spatiale, $\mathbf{\Sigma}$ la matrice de covariance des coordonnées spatiales. Cette matrice peut être parametrée sous la forme $\mathbf{\Sigma} = \mathbf{R}_\theta^T \mathbf{\Lambda}\, \mathbf{R}_\theta$ où

$$\mathbf{\Lambda} = \begin{bmatrix} l_u^2 & 0 \\ 0 & l_v^2 \end{bmatrix} \tag{6}$$

est la matrice de covariance exprimée dans la base de ses vecteurs propres $[u, v]^T$. Les valeurs l_u et l_v sont les longueurs de corrélation principale et secondaire dans les directions des vecteurs propres et

$$\mathbf{R}_\theta = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{7}$$

est une rotation unitaire d'angle θ déterminant l'orientation privilegiée de la texture. La transformée de Fourier d'une gaussienne étant une gaussienne, et puisque les rotations spatiales sont conservées, la densité spectrale de puissance a une expression très similaire :

$$S_T(\mathbf{f}) = \sigma_T^2 \pi \ \det(\mathbf{\Sigma})^{1/2} \exp\left(-\pi^2 \mathbf{f}^T \mathbf{\Sigma} \mathbf{f}\right) + \mu_T^2 \delta(\mathbf{f}) \tag{8}$$

où $\mathbf{f} = [f_x, f_y]^T$ est le vecteur des fréquences 2-D, μ_T est la valeur moyenne de la texture et $\delta(\mathbf{f})$ est la distribution de Dirac.

3 Estimation des paramètres

3.1 Le tenseur de structure

La présence de speckle dans l'image fait que les statistiques liées à la texture ne sont pas directement accessibles. On va donc considérer qu'une image préfi ltrée contient l'information spatiale à une échelle donnée, et peut être vue comme une approximation de la scène non bruitée. Nous choisissons en particulier de représenter une échelle de la scène selon la théorie du scale-space gaussien [7] par la convolution de l'image bruitée avec un noyau gaussien isotrope. Ainsi, la scène à l'échelle σ est donnée par $I_\sigma = K_\sigma * I$, où K_σ est un noyau gaussien d'écart-type σ. Cette représentation assure la dérivabilité de l'image et l'opérateur nommé tenseur de structure peut ainsi être introduit [1] :

$$\mathbf{J} = K_\rho * (\boldsymbol{\nabla} I_\sigma \boldsymbol{\nabla} I_\sigma^T) \tag{9}$$

où $\boldsymbol{\nabla}$ est l'opérateur gradient. Le noyau gaussien K_ρ est une fenêtre de moyennage spatial jouant le rôle d'estimateur de l'espérance mathématique. En remplaçant K_ρ par l'opérateur espérance mathématique et en passant dans le domaine spectral, on peut montrer [1] que le tenseur de structure contient l'information statistique de second-ordre puisque sa valeur est proportionnelle aux moments de second-ordre du spectre $S_{I_\sigma}(\mathbf{f})$:

$$\mathbf{J} = 4\pi^2 \int_{R^2} \mathbf{f}^T \mathbf{f}\ S_{I_\sigma}(\mathbf{f})\ d\mathbf{f}. \tag{10}$$

Ainsi, en supposant le modèle AGK avec la covariance spatiale $\mathbf{\Sigma}$ pour les statistiques de second-ordre de I_σ, il est possible de retrouver les longueurs de corrélation du signal à l'echelle σ, puisque :

$$\mathbf{J} = 2\sigma_{I_\sigma}^2 \mathbf{\Sigma}^{-1} \tag{11}$$

où $\sigma_{I_\sigma}^2$ est la variance du signal à l'échelle σ.

3.2 Extraction des paramètres caractéristiques

En calculant la décomposition en valeurs/vecteurs propres du tenseur :

$$\mathbf{J} = \lambda_1 \mathbf{k_1} \mathbf{k_1}^T + \lambda_2 \mathbf{k_2} \mathbf{k_2}^T \tag{12}$$

il est possible d'estimer les paramètres du modèle sans estimation directe de la fonction d'auto-corrélation. Les longueurs de corrélation l_u et l_v peuvent etre déduites des valeurs propres de $\mathbf{J}$

$$\lambda_{1,2} = \frac{1}{2}(J_{22} + J_{11} \pm \sqrt{(J_{22} - J_{11})^2 + 4J_{12}^2}) \tag{13}$$

puisque les équations (6), (8) et (11) indiquent que $\lambda_1 = 2\sigma_{I_\sigma}^2/l_v^2$ and $\lambda_2 = 2\sigma_{I_\sigma}^2/l_u^2$. L'orientation dominante est déterminée par le vecteur propre $\mathbf{k_2}$ correspondant a la plus petite valeur propre, et son angle est donné par :

$$\theta = \tan^{-1}(k_{2,y}/k_{2,x}). \tag{14}$$

Afi n de quantifi er l'importance de l'orientation principale, il est utile de défi nir un paramètre nommé anisotropie spatiale :

$$A = 1 - \lambda_2/\lambda_1. \tag{15}$$

Ce paramètre, qui est indépendant de la variance du signal $\sigma_{I_\sigma}^2$, est compris entre 0 (texture isotrope) et 1 (orientation secondaire nulle).

4 Filtrage basé sur la texture

Dans cette section, nous introduisons une méthode de fi ltrage exploitant les paramètres de texture présentés précédemment et pouvant être utilisée pour améliorer les performances des fi ltres basés sur l'estimation des statistiques locales. Nous prendrons comme exemple les fi ltres de type MMSE.

4.1 Les filtre MMSE

La classe des filtres MMSE [4, 6] est l'une des plus utilisée pour le débruitage des images SAR. L'estimation de la réflectivité locale est donnée par le critère :

$$\hat{R}(x,y) = k(x,y)I(x,y) + \hat{\mu}_I\left(1 - k(x,y)\right). \qquad (16)$$

La variable $k(x,y)$ depend de l'homogénéité de la zone considerée et permet de combiner la valeur de la moyenne locale avec la valeur observée du signal bruité. La fonction de pondération $k(x,y)$ dépend du coefficient de variation et autorise un moyennage maximal sur les zones homogènes ainsi que la préservation des hétérogeneités. Une limitation importante d'un tel filtre est l'utilisation d'une fenêtre d'estimation fixe, donc ne s'adaptant pas à la structure spatiale de l'image. Pour pallier cet inconvénient, des versions améliorées [5] estiment la valeur moyenne pour un nombre fini de masques orientés, et retiennent la valeur correspondant au coefficient de variation le plus faible (filtre de Lee géométrique). Si le résultat est nettement amélioré, l'introduction de cet ensemble discret de décisions entraîne des transitions brusques dans l'estimée de la moyenne, ce qui permet une bonne préservation des contours, mais n'autorise pas la restauration de la texture a l'intérieur d'une région. Cela entraîne de plus la dégradation des structures dont l'orientation n'est pas alignée avec celle de l'un des masques.

4.2 Le filtre AGK-MMSE

Pour améliorer la restauration de la réfléctivité dans les zones hétérogènes, nous proposons un nouveau filtre qui combine le critère MMSE avec une estimation adaptative des statistiques locales, exploitant les paramètres de texture. Par la suite, ce filtre sera appelé "AGK-MMSE". Dans un premier temps, les paramètres spatiaux A et θ sont estimés pour chaque pixel au moyen de la décomposition valeurs/vecteurs propres du tenseur de structure (12) à l'echelle σ, sur un voisinage déterminé par la fenêtre K_ρ. Les statistiques locales du filtre MMSE sont ensuite estimées dans une fenêtre gaussienne anisotrope d'orientation $\hat{\theta}$ et d'anisotropie $\hat{A}$:

$$H(\mathbf{d}) = \exp\left(-\mathbf{d}^T \mathbf{\Sigma}^{-1}_{\hat{A},\hat{\theta}} \mathbf{d}\right). \qquad (17)$$

L'expression de la matrice de covariance de ce noyau gaussien anisotrope est :

$$\mathbf{\Sigma}_{\hat{A},\hat{\theta}} = \mathbf{R}^T_{\hat{\theta}} \begin{bmatrix} \rho^2 & 0 \\ 0 & \rho^2(1-\hat{A}) \end{bmatrix} \mathbf{R}_{\hat{\theta}} \qquad (18)$$

où $\mathbf{R}_{\hat{\theta}}$ est la matrice de rotation d'angle $\hat{\theta}$. Les écart-types de ce noyau sont limités par le paramètre ρ de la fenêtre d'estimation des paramètres de texture, afin d'éviter d'inclure dans l'estimation des pixels exterieurs a cette fenêtre. Enfin, le critère MMSE (16) est appliqué afin de préserver les diffuseurs isolés, la moyenne locale à la position $\mathbf{u_0}$ étant donnée par :

$$\hat{\mu_I}(\mathbf{u_0}) = \frac{\int_{\mathbf{u}\in R^2} H(\mathbf{u}-\mathbf{u_0}) I(\mathbf{u}) d\mathbf{u}}{\int_{\mathbf{u}\in R^2} H(\mathbf{u}-\mathbf{u_0}) d\mathbf{u}}. \qquad (19)$$

Dans cet article, on choisit d'appliquer les statistiques du filtre de Lee [6].

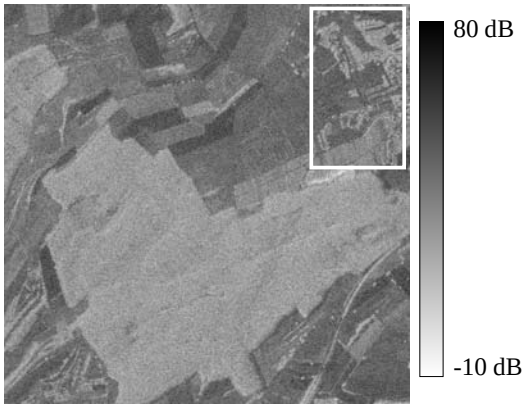


FIG. 1 – Zone test mono-vue extraite de "Weiherbachtal" (ESAR, DLR), et sous-image retenue pour l'évaluation visuelle (rectangle blanc).

TAB. 1 – Resultat de filtrage sur une zone homogène

	Image Bruitée	Filtre MMSE	Filtre AGK-MMSE
ENL théorique	1	45	45.36
ENL estimé	0.95	15.8	17.9
Moyenne estimée	45.0 dB	45.1 dB	45.0 dB

5 Résultats

Le filtre AGK-MMSE a été validé sur les données mono-vue en bande L de la scène "Weiherbachtal" acquise par le capteur ESAR du DLR (Centre Aerospatial Allemand). Ses performances ont été comparées à celles du filtre MMSE géométrique. La figure 1 représente la zone test utilisée pour l'évaluation.

L'anisotropie A étant très sensible au bruit, un seuil a été introduit sur ce paramètre, en fonction des statistiques du coefficient de variation local CV_I^2, calculées sur une zone homogène de l'image. Ainsi, la valeur de A est forcée à zéro pour les pixels qui remplissent la condition $CV_I^2 < MEAN(CV_I^2) + STD(CV_I^2)$. Si l'on approxime la distribution de CV_I^2 par une loi normale, cette condition correspond à un intervalle de confiance de 68 % pour l'hypothèse de la zone homogène.

Les deux paramètres du filtre AGK-MMSE sont l'échelle du préfiltrage σ et l'échelle de la fenêtre d'estimation ρ. Nous avons déterminé de façon expérimentale que le choix $\rho = \sqrt{2}\sigma$ donnait de bons résultats et avait l'avantage de réduire le nombre de paramètres d'entrée à un. L'anisotropie et l'angle d'orientation estimés sur la zone test sont montrés figure 2. Le moyennage maximal de notre filtre est atteint sur une zone homogène et sa valeur théorique est $ENL_{AGK-MMSE} = 2\pi\rho^2$. Par conséquent, pour comparer les performances réelles des deux filtres sur une zone homogène, nous avons fixé σ à 1.9 afin d'obtenir un moyennage proche de celui du filtre de Lee géométrique pour une fenêtre 9×9.

Les résultats quantitatifs sur une zone homogène sélectionée manuellement sont reportés dans le tableau 1 et l'on peut observer que le filtre AGK-MMSE atteint un nombre de vues équivalentes comparable a celui du filtre MMSE géométrique, alors que la valeur du biais est faible pour les deux filtres. La comparaison visuelle des résultats de filtrage sur une sous-zone

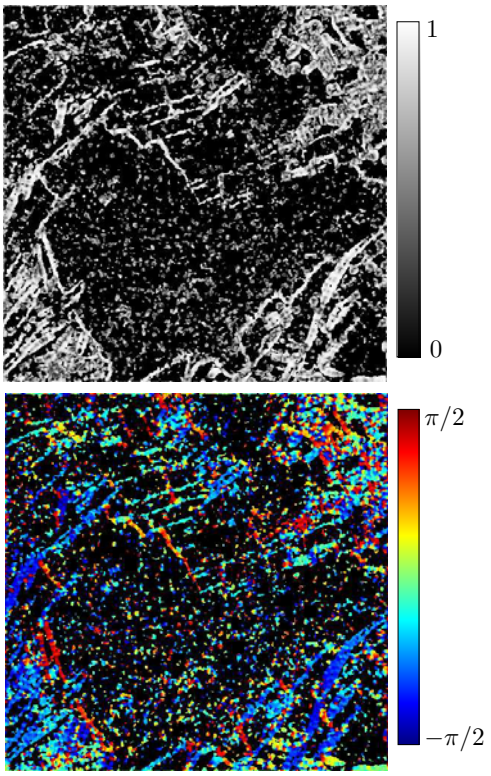


FIG. 2 – Anisotropie A (haut) and orientation θ (bas) estimées pour $\sigma = 1.9$ and $\rho = \sqrt{2}\sigma$. A et θ sont représentés pour les valeurs signifi catives du coeffi cient de variation (voir texte). La référence pour θ est verticale.

ainsi que les rapports entre l'image bruitée et l'image fi ltrée (fi - gure 3) montrent que le fi ltre AGK-MMSE préserve mieux les structures que le fi ltre MMSE géométrique.

6 Conclusion

Nous avons présenté un nouveau modèle pour la déscription locale des statistiques d'ordre deux de la texture dans les images SAR, qui permet de représenter les différentes formes de texture, c'est à dire le fouillis hétérogène et les sturctures déterministes. Ce modèle est basé sur une forme paramétrique de la fonction d'auto-covariance de la texture, appelée noyau gaussien anisotrope. Ainsi, chaque pixel peut être caractérisé par une orientation locale et une anisotropie, relatives à son voisinage. Il à ensuite été montré que l'opérateur nommé tenseur de structure permettait l'estimation des paramètres du modèle à une échelle donnée. Cette modélisation a ensuite été appliquée au domaine du fi ltrage de speckle. Nous avons en effet proposé une amélioration du fi ltre MMSE géometrique, prenant en compte ces paramètres de texture. Ce fi ltre nommé AGK-MMSE est fi nallement comparé au fi ltre MMSE traditionnel et montre de meilleures performances en terme de préservation des structures présentes dans l'image.

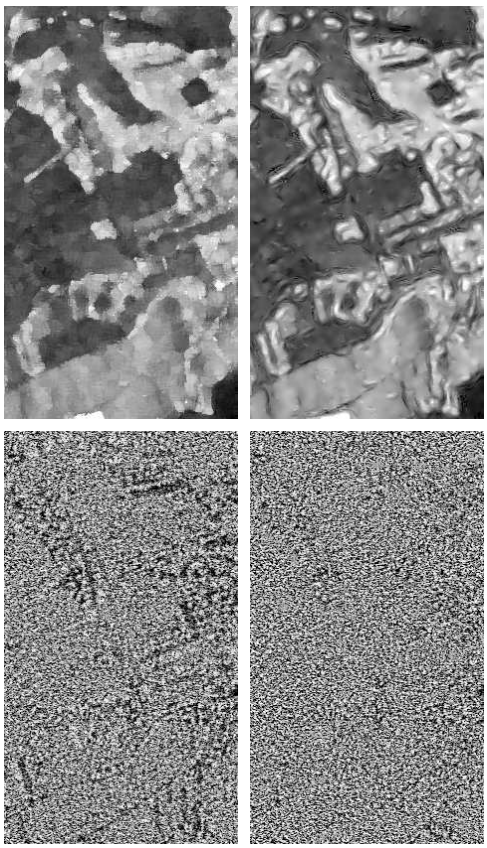


FIG. 3 – Résultats de fi ltrage : fi ltre MMSE (haut gauche), fi ltre AGK-MMSE (haut droite), rapport entre les images originale et fi ltrée pour : le fi ltre MMSE (bas gauche), le fi ltre AGK-MMSE (bas droite).

Références

[1] J. Bigun, G. H. Granlund, and J. Wiklund. Multidimensional orientation estimation with applications to texture analysis and optical flow. *IEEE Trans. Pattern Anal. Machine Intell.*, 13(8) :775–790, 1991.

[2] O. D'Hondt, L. Ferro-Famil, and E. Pottier. Local orientation analysis of spatial texture from polarimetric SAR data'. In *IGARSS Symposium*, September 2004.

[3] O. D'Hondt, L. Ferro-Famil, and E. Pottier. Non stationary texture analysis from polarimetric SAR data. In *EUSAR Symposium*, May 2004.

[4] D. T. Kuan, A. A. Sawchuk, T. C. Strand, and P. Chavel. Adaptive noise smoothing fi lter for images with signal-dependent noise. *IEEE Trans. Pattern Anal. Machine Intell.*, PAMI-7(2) :165–177, 1985.

[5] J.-S. Lee. Refi ned fi ltering of image noise using local statistics. *Comput. Vision, Graphics, Image Process*, 15(2) :380–389, 1981.

[6] J.-S. Lee. Speckle suppression and analysis for synthetic aperture radar images. *Optical Engineering*, 25(5) :636–645, 1986.

[7] T. Lindeberg. Scale-space theory : A basic tool for analysing structures at different scales. *J. of Applied Statistics*, 21(2) :224–270, 1994.

[8] F. T. Ulaby, F. Kouyate, B. Brisco, and T. Williams. Textural Information in SAR Images. *IEEE Trans. Geosci. Remote Sensing*, 24(2) :235–245, 1986.

Quantification vectorielle algébrique modulée : codage de source/tatouage conjoints à débit variable

Ludovic GUILLEMOT, Jean-Marie MOUREAUX

Université Henri Poincaré, Nancy 1, CRAN, CNRS UMR 7039,
BP 239, F-54506 Vandœuvre-lès-Nancy Cedex
{ludovic.guillemot,jean-marie.moureaux}@cran.uhp-nancy.fr

Résumé – Nous proposons ici un schéma dit de quantification vectorielle algébrique modulée (QVAM). Il s'agit d'un codeur de source à débit variable permettant l'insertion d'information. Ses bonnes performances en terme de compromis débit distorsion rendent possible son utilisation dans des domaines où la compression avec perte est incontournable, en particulier celui du codage par transformée reconnu comme le plus efficace, notamment pour les images. La QVAM est essentiellement basée sur, d'une part l'utilisation d'un quantificateur vectoriel permettant de diminuer de manière significative la limite de la borne entropique du débit, et d'autre part, sur l'emploi d'une zone morte vectorielle pour exploiter la représentation creuse de la source. En outre, la phase d'estimation de la taille des cellules de quantification est un problème clef des méthodes d'insertion par quantification qui a fait l'objet, à notre connaissance, de travaux uniquement dans le cas scalaire. Nous proposons ici une méthode d'estimation adaptée aux quantificateurs vectoriels algébriques.

Abstract – The scheme we propose here is called modulated lattice vector quantization (MLVQ). It is a variable rate coder dedicated to joint compression and information embedding. Thanks to its good performances in terms of rate distortion trade-off, it can be advantageously used in a transform coding context, for exemple in image coding. MLVQ is based on one hand, on the use of a vector quantizer permitting to decrease significantly the limit of the entropic bound, and on the other hand, on a vector dead zone which allows to take profit of the sparsity of the source. Moreover, the need to know quantization parameters previous at the extraction step is a key point of quantization based methods. Existing methods are dedicated to scalar quantizers. Here we propose an estimation procedure of the size of the quantization cells adapted to the multidimensionnal case.

1 L'approche tatouage/compression conjoints

Du fait de l'explosion des communications numériques, la dernière décennie a vu de nombreux travaux de recherche, tant dans le domaine de la compression que dans celui du tatouage. La compression étant une attaque redoutable pour le tatouage, l'approche tatouage/compression conjoints offre de nombreux avantages, parmi lesquels celui pour la compression de ne plus être une source d'interférence lors de l'extraction du tatouage.

Cette approche a été cependant peu abordée dans la littérature ou proposée dans un contexte ad'hoc [5] [9]. Ainsi, par exemple Wu et *al.* ont développé dans [8] un schéma conjoint effectif basé sur la méthode de quantification par modulation d'index (QIM) [1]. Cependant le codage de source mis en oeuvre est à pas fixe, et par conséquent n'exploite pas les redondances présentes à l'intérieur des signaux. En outre, le cadre scalaire semble peu adapté à ce type d'application en raison de l'existence d'une limite inférieure de la borne entropique (dû au fait que l'on ajoute 1 bit par échantillon). A noter que cette méthode ne permet pas d'exploiter les propriétés de robustesse liées à l'emploi d'un quantificateur vectoriel. Enfin, le problème de l'estimation des paramètres de quantification est crucial dans une approche conjointe qui nécessite de pouvoir faire varier la taille des cellules de quantification, en d'autres termes le taux de compression. A notre connaissance, des solutions à ce problème a été développées dans [2] et [6] mais elles se limitent au cas scalaire.

Le schéma hybride que nous proposons ici est également basé sur le principe de la QIM, mais dans son cadre vectoriel. Il s'agit d'un quantificateur vectoriel algébrique modulé (QVAM) de faible complexité[1] qui permet tatouage et codage à débit variable. L'originalité de nos travaux réside dans le fait que notre méthode permet à la fois d'exploiter les redondances d'une source grâce à l'emploi d'une zone morte vectorielle [7] et d'estimer les paramètres du quantificateur dans le cas vectoriel.

Le plan de cet article est le suivant. Dans le paragraphe 2 nous décrivons l'algorithme QVAM proposé. Le paragraphe 3 est consacré à l'estimation du paramètre de quantification et est suivi par une conclusion.

2 Quantification vectorielle algébrique modulée

Le principe de l'insertion d'information par QIM repose sur le partitionnement d'un dictionnaire en m sous-dictionnaires permettant l'insertion d'un message m-aire[2]. Soit $X \in \mathbb{R}^n$ un vecteur de la source et γ le facteur d'échelle, *i.e.* le paramètre déterminant la taille des cellules de quantification.

On définit le **quantificateur vectoriel algébrique modulé** Q_i à partir du quantificateur uniforme Q dans $\mathbb{Z}^n$ de la manière

[1] A la fois du point de vue du codage de source (grâce à la structure du quantificateur) et du tatouage (effectué conjointement).

[2] Nous nous limiterons dans cet article à l'insertion d'un message binaire.

suivante :

$$Q_i(X) \triangleq mQ\left(\frac{X - i\frac{\gamma}{m}}{\gamma}\right) + i, \text{avec } i \in \{0, 1, ..., m-1\}. \quad (1)$$

L'ensemble des vecteurs obtenus à partir des m quantificateurs Q_i est l'union de m cosets S_i de $\mathbb{Z}^n$: $S_i \triangleq \{m\mathbb{Z}^n + [i]\}$.

Nous appelerons cet ensemble **réseau modulé** $\mathbb{Z}^n_m$:

$$\mathbb{Z}^n_m \triangleq \bigcup_{i=0}^{m-1} S_i, \text{ avec } [i] = \begin{pmatrix} i \\ ... \\ i \end{pmatrix}. \quad (2)$$

Afin d'exploiter la structure du réseau nous avons proposé une méthode d'indexage adaptée au réseau modulé [4].

Nous allons à présent expliquer pourquoi coder une source sur ce type de réseau conduit à de faibles performances, en dehors des hauts débits, en terme de compromis débit-distorsion.

Pour des vecteurs de taille n (la population de la couche n du dictionnaire se note $N(n)$) le débit minimal atteignable $\mathcal{H}^\infty$ est donné par la formule suivante :

$$R_C(\gamma) \underset{\gamma\to\infty}{\to} \mathcal{H}^\infty = 1 + \frac{log_2(N(n))}{2} \text{ bits par vecteur.} \quad (3)$$

avec R_C le débit entropique de la source.

Preuve. On suppose ici que le message à insérer est équiprobable. Lorsque le facteur d'échelle tend vers l'infini, seulement deux couches du réseau modulé sont utilisées, la couche 0 (correspondant au bit 0) et la couche n (bit 1). D'après la formule du débit du schéma QVA disponible dans [7], on a :

$$\begin{aligned}\mathcal{H}^\infty = &- [P(\|Y\| = 0)\log_2(P(\|Y\| = 0)) + \\ &P(\|Y\| = n)\log_2(P(\|Y\| = n))] + \\ &[P(\|Y\| = 0)\log_2(N(0)) + P(\|Y\| = n)\log_2(N(n))].\end{aligned}$$

En considérant l'équiprobabilité du message on obtient : $P(\|Y\| = 0) = P(\|Y\| = n) = \frac{1}{2}$. D'où la relation :

$$\mathcal{H}^\infty = \frac{1}{n}\left\{1 + \frac{1}{2}\left[\log_2(N(0)) + \log_2(N(n))\right]\right\} \text{ bits/éch.}$$

A noter que l'on a les relation suivantes :
$N(0) = card(S_0(0)) = 1$
$N(n) = card(S_0(n)) + card(S_1(n))$
$= card\left(N_{\mathbb{Z}^n}\left(\frac{n}{2}\right)\right) + card\{Y = (\pm 1, .., \pm 1)\}$
$= card\left(N_{\mathbb{Z}^n}\left(\frac{n}{2}\right)\right) + 2^n.$
avec $N_{\mathbb{Z}^n}\left(\frac{n}{2}\right)$ la population de la couche $\frac{n}{2}$ du réseau $\mathbb{Z}^n$. ■

La formule (3) plaide tout d'abord en faveur de l'approche vectorielle, la limite $\mathcal{H}^\infty$ étant de 1,5 bits par éch. pour le cas scalaire contre 0,8 bits/éch., par exemple, pour la QVAM avec des vecteurs de taille 8. Sur la figure 1 sont représentées les courbes distorsion en fonction du débit entropique du quantificateur scalaire uniforme et du quantificateur modulé. Comme on peut le constater, hormis pour les haut débits, la quantification scalaire modulée est un piètre codeur de source. Un autre facteur entre également en jeu ici : la représentation creuse du signal quantifié est entravée par la modulation.

La contribution majeure de nos travaux réside dans l'introduction du *principe d'exclusion* dans la stratégie d'insertion par QIM : les vecteurs appartenant à une région appelée zone morte vectorielle (ZMV) sont quantifiés par le vecteur nul et exclus du processus d'insertion. De cette manière, la limite $\mathcal{H}^\infty$ est d'une part abaissée et, d'autre part, la représentation creuse du signal - exploitable par un codeur entropique - est maintenue en sortie du quantificateur.

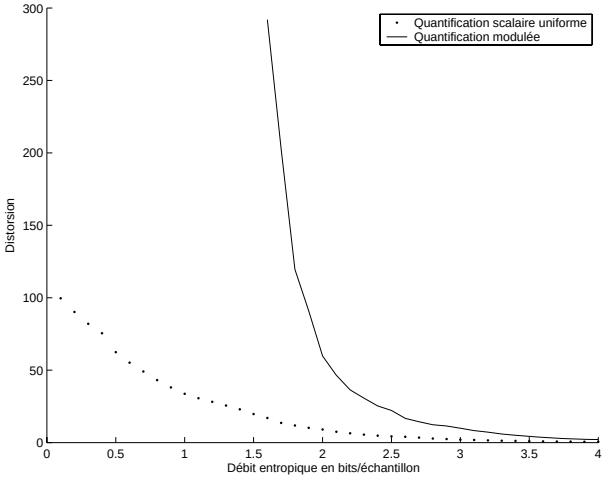


FIG. 1 – Comparaison des courbes débit-distorsion des schémas Quantification Scalaire Uniforme et Quantification Scalaire Modulée : cas d'une source laplacienne centrée d'écart-type $\sigma = 10$.

La figure 2 illustre le principe d'exclusion par zone morte vectorielle pyramidale, dans le cas de $\mathbb{Z}^2_2$. Les vecteurs de la source appartenant à la zone morte sont quantifiés par zéro et ne contribuent pas à l'insertion du message. Les figures 3 et 4 montrent la superiorité des performances de la QVAM par rapport à la QIM (c'est-à dire la QVAM sans zone morte), en terme de compromis débit-distorsion. Sur la figure 3 la taille de la zone morte est déterminée de manière à minimiser la distorsion pour un débit donné. Sur la figure 4, la taille est fixée en fonction du taux d'insertion souhaité.

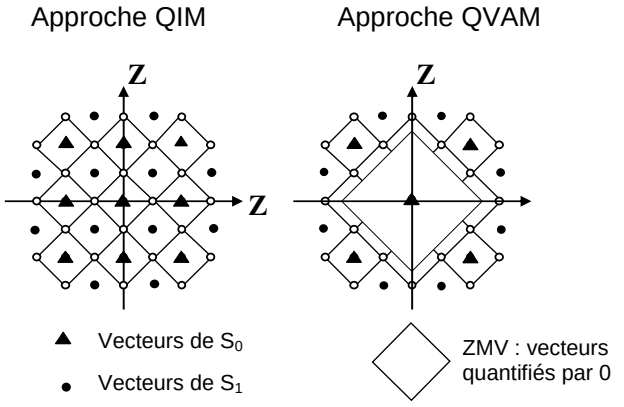


FIG. 2 – Zone morte vectorielle (ZMV) pyramidale sur le réseau $\mathbb{Z}^2_2$.

Les figures 5 et 6 permettent de mieux se rendre compte des performances de la QVAM. On peut y voir deux versions de l'image Lena codées respectivement sans et avec indexage modulé et zone morte vectorielle (l'allocation des débits est la même pour les deux schémas). Le résultat est sans appel : tandis que l'image compressée par QVAM affiche un PSNR correct (33,1 dB), la première montre que le codage sans indexage modulé ni zone morte vectorielle est totalement inadapté à la

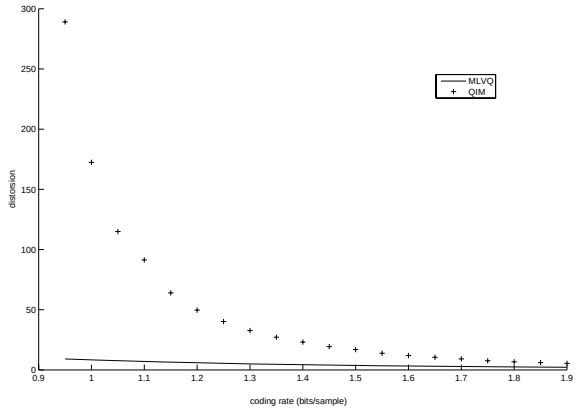


FIG. 3 – Courbes débit-distorsion des schémas QVAM avec zone morte optimale et QIM : cas d'une sous-image d'ondelettes de l'image Lena (filtre 9-7, niveau 3, détails verticaux).

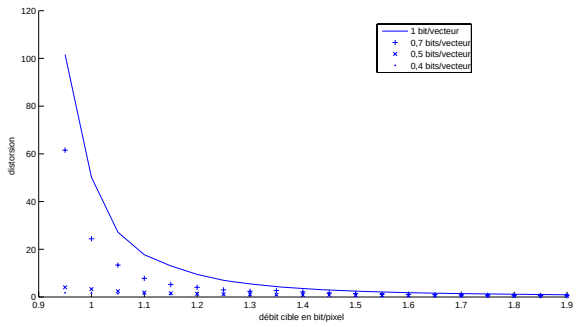


FIG. 4 – Courbe débit-distorsion de l'approche QVAM pour quatre taux d'insertion.

compression.

3 Estimation de la taille de la cellule de quantification

Une des principales limitations des méthodes d'insertion par quantification réside dans le fait que la taille des cellules est un paramètre qui doit être connu à l'extraction du schéma de tatouage. Dans le cadre de l'approche tatouage/compression de source à débit variable il est impossible de fixer cette donnée, puisque de la taille des cellules dépend le débit.

Nous proposons ici une méthode d'estimation de la taille des cellules de quantification (ETCQ) basée sur la requantification de la source tatouée/compressée lors de la phase d'extraction. Cette technique d'estimation est la généralisation aux réseaux de l'approche que nous avons proposé dans [3] pour un quantificateur scalaire uniforme. Soit $\gamma \in \mathbb{R}_+$, on définit $\gamma\Lambda$ comme le réseau dilaté Λ d'un facteur γ. Les deux propriétés du réseau $\gamma\Lambda$ ci-dessous constituent le point central de notre méthode d'estimation :

$$\text{Si } \gamma = k,\ k \in \mathbb{N},\ \text{alors } \gamma\Lambda \subset \Lambda \tag{4}$$

$$\text{Si } \gamma = \frac{1}{k}, k \in \mathbb{N},\ \text{alors } \gamma\Lambda \supset \Lambda \tag{5}$$

FIG. 5 – Image Lena compressée sans indexage modulé ni zone morte vectorielle, avec un débit de 0,36 bits/pixel et un PSNR de 16,8 dB ; longueur du message inséré : 8,6 kbits.

La fonction distorsion après requantification par un facteur d'échelle α : $\alpha \mapsto d_{Y_{\gamma\Lambda}}(\alpha) = \frac{1}{n}\left\|Y_{\gamma\Lambda} - Q_{\alpha\Lambda}\left(Y_{\gamma\Lambda}\right)\right\|_2^2$ admet alors les deux caractéristiques suivantes :

1. $D_{Y_{\gamma\Lambda}}(\alpha)$ est nulle (et minimale) en $\frac{\gamma}{k}$, $k \in \mathbb{N}^*$.
2. $D_{Y_{\gamma\Lambda}}(\alpha)$ admet des minimas locaux en $\alpha = k\gamma$, $k \in \mathbb{N}^*$.

L'estimation du facteur d'échelle consiste à déterminer le dernier minimum avant une forte pente. Les figures 7 et 8 illustrent le fonctionnement de l'ETCQ : la source quantifiée/tatouée est requantifiée avec un pas variable. Le facteur d'échelle correspond au dernier minimum local. Comme on peut le voir, ce minimum existe toujours en dépit d'une attaque telle qu'un bruit blanc gaussien additif, ou encore un gain. Dans le cas d'un bruit additif d'écart-type inférieur à 1,5, le taux d'erreur d'extraction est négligeable. Il est de 10% lorsque l'écart-type est de 1,5. Dans le cas du gain, les minimas correspondent bien au facteur d'échelle multiplié par le gain utilisé et les erreurs d'extraction sont nulles à chaque fois.

4 Conclusion

Nous avons présenté ici un schéma original de compression et tatouage conjoints : le quantificateur vectoriel algébrique modulé. L'approche conjointe offre les avantages suivants : faible complexité, aucune erreur d'extraction due à la compression avec perte (en l'absence d'autres attaques). Tenant compte des caractéristiques des schémas de compression les plus performants à l'heure actuelle, la QVAM a été mis au point afin de permettre un codage à longueur variable, et surtout de tirer profit de la parcimonie des sources obtenues après transformation.

Cette dernière propriété n'est à notre connaissance pas représentée dans la littérature. Pour ce faire, la QVAM repose d'une part, sur les bonnes propriétés en terme de débit entropique des quantificateurs vectoriels algébriques et, d'autre part sur l'exclusion du processus d'insertion de certains éléments de

FIG. 6 – Image Lena compressée avec indexage modulé et zone morte vectorielle, avec un débit de 0,36 bits/pixel et un PSNR de 33,1 dB ; longueur du message inséré : 3,8 kbits.

la source à travers une zone morte vectorielle. Ce dernier point permettant de maintenir une représentation creuse de la source en sortie du quantificateur.

Enfin, un problème clé relatif à l'approche conjointe a été également abordé : la taille des cellules de quantification ne peut être figée dans l'optique d'un codeur de source performant. L'efficacité de notre méthode d'estimation ETCQ permet de résoudre ce problème, y compris en cas d'attaques telles qu'un bruit additif, ou encore un gain.

Les performances de codage de notre schéma sont encourageantes et démontrent l'intérêt de l'approche conjointe compression / tatouage dans un contexte où la compression constitue la principale attaque. Les questions liées à l'amélioration de la robustesse avec l'emploi de réseaux mieux adaptés ou de codes correcteurs d'erreurs offrent de nombreuses perspectives de recherche. Enfin, la transposition de notre approche à d'autres types de codeurs semble être également un champ d'étude très prometteur.

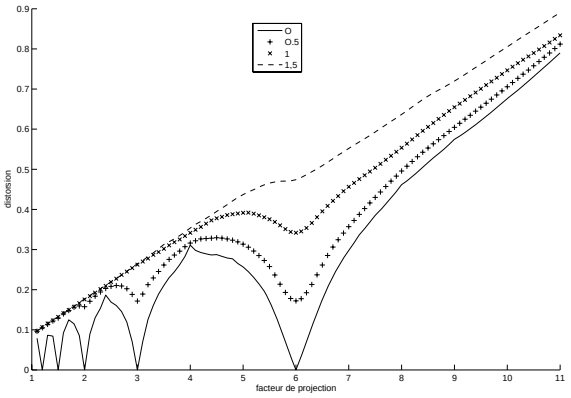


FIG. 7 – Courbes distorsion en fonction du facteur d'échelle en l'absence d'attaques et en présence d'un bruit blanc gaussien additif d'écart-type 0,5 ; 1 et 1,5 (facteur d'échelle initial égal à 6).

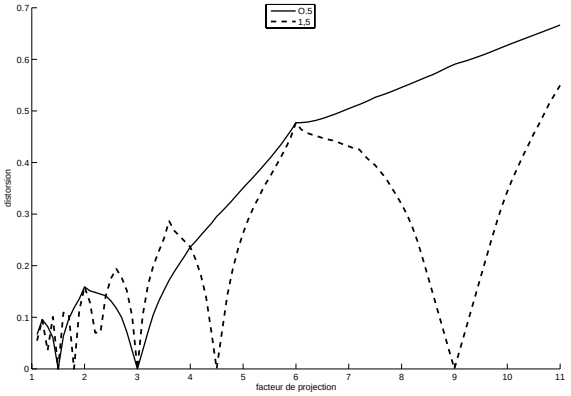


FIG. 8 – Courbes distorsion en fonction du facteur d'échelle en présence d'un gain de 0,5 et 1,5 (facteur d'échelle initial égal à 6).

Références

[1] B. Chen et G. Wornell, "Quantization Index Modulation : A Class of Provably Good Methods for Digital Watermarking and Information Embedding", IEEE Trans. on Information Theory, vol. 47, NO. 4, May 2001.

[2] J. J. Eggers, R. Bäuml, R. Tzschoppe et B. Girod, "Scalar Costa Scheme for Information Embedding", IEEE Trans. on Signal Processing, vol. 51, No. 4, pp. 1003-1019, avril 2003.

[3] L. Guillemot et J. M. Moureaux "Bite-Rate adapted Watermarking algorithm for compressed images", IEEE ICME 2002, Lausanne, août 2002.

[4] L. Guillemot et J. M. Moureaux, "Hybrid transmission, compression and data hiding : quantisation index modulation as source coding strategy", Electronics letters Vol. 40, No. 17, pp. 1053-1055, août 2004.

[5] J. M. Moureaux et L. Guillemot, "Image Compression and Watermarking using Lattice Vector Quantization", SPIE security and watermarking of multimedia contents, pp. 600-610, San Jose, janvier 2002.

[6] I. D. Shterev, I. L. Lagendijk et R. Heusdens, "Statistical Amplitude Scale Estimation for Quantization-based Watermarking", SPIE security, steganography and watermarking of multimedia contents VII, San José, janvier 2004.

[7] T. Voinson, L Guillemot et J.M. Moureaux, "Image compression using Lattice Vector Quantization with code book shape adapted thresholding", IEEE 2002 International Conference on Image Processing, Rochester, septembre 2002.

[8] G. Wu et E.H. Yang, "Joint watermarking and compression using scalar quantization for maximizing robustness in the presence of additive gaussian attacks", IEEE Trans. on Signal Processing, vol. 53, NO. 2, pp. 834-844, février 2005.

[9] L. Xie et G.R. Arce, "A Class of Authentication Digital Watermarks for Secure Multimedia Communication," IEEE Trans. on Image Processing, vol. 10, NO. 11, 2001.

Amélioration de l'invisibilité par adaptation de la quantification aux données à insérer.

Jean - Luc TOUTANT, William PUECH, Christophe FIORIO

LIRMM
161 rue Ada, 34392 Montpellier Cedex 5, France
toutant@lirmm.fr, puech@lirmm.fr, fiorio@lirmm.fr

Résumé – Cet article présente une amélioration pour les méthodes d'insertion de données basées sur la DCT : l'adaptation de la quantification à l'insertion de données cachées. Les méthodes actuelles n'intègrent l'information secrète qu'après l'étape de quantification de la compression JPEG. La robustesse au format JPEG est bien obtenue, mais l'image subit deux pertes d'information successives : la quantification et l'insertion des données. En adaptant la quantification au message secret, nous réduisons les variations induites et donc, améliorons la qualité de l'image compressée et marquée vis-à-vis de l'originale.

Abstract – This paper presents an improvement for DCT-based data-hiding methods : embedding bits during quantization. Current methods embed the secret information after the quantization of the JPEG compression. The robustness to JPEG is obtained but the image undergoes two successive losses of information: the quantization and the data-hiding. By taking into account the hidden bit in the quantization, we reduce the induced variation and so, improve the quality of the stego-image relatively to the original one.

1 Introduction

Dans cet article, nous proposons une amélioration pour les méthodes d'insertion de données basées sur la compression JPEG. Elle s'applique aux méthodes procédant par substitution des bits de poids faibles (LSBs) des coefficients DCT.

Actuellement, la quantification et l'insertion de données sont des étapes bien distinctes. L'image subit donc deux dégradations successives.

En intégrant l'information pendant la quantification, nous réduisons la dégradation finale. Pour cela, nous minimisons l'approximation faite par la quantification en fonction des données à insérer. La qualité visuelle de l'image compressée et marquée est ainsi améliorée.

Nous commencerons par nous intéresser à la compression JPEG et aux méthodes d'insertions de données basées sur la DCT. Nous présenterons ensuite l'amélioration de la qualité visuelle avant d'illustrer notre propos par une application sur la méthode Jpeg/Jsteg.

2 La compression JPEG

Avec le développement des réseaux, la compression est devenue un point essentiel pour garantir des transmissions rapides. Le format JPEG [1], basé sur la Transformée Cosinus Discrète (DCT), est le standard actuel de compression pour les images. Les méthodes d'insertion de données se doivent d'y être robustes.

La compression JPEG se décompose en différentes étapes présentées figure 1 : la transformation du domaine spatial à un domaine fréquentiel, la quantification et le codage entropique. C'est une méthode de compression avec pertes puisque la quantification dégrade le contenu de l'image.

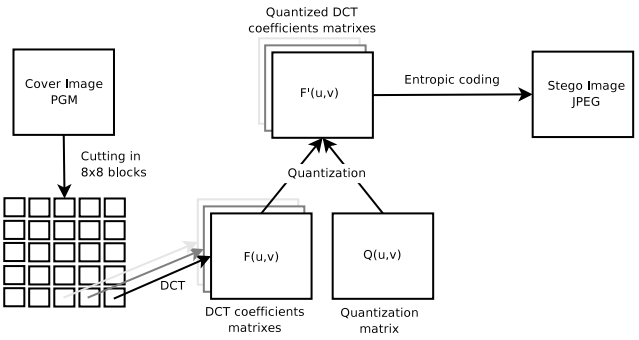


FIG. 1 – Le processus de compression JPEG.

Au préalable, pour réduire la complexité des calculs en espace et en temps, l'image est divisée en blocs de 8×8 pixels qui seront traités séparément.

Ensuite, l'image est transformée de sa représentation initiale, une matrice de pixels, à sa décomposition en une somme de fréquences. À partir de chaque bloc de 8×8 pixels, une matrice 8×8 est déterminée par la DCT. Un couple de coordonnées (u, v) définit alors une fréquence et la valeur associée dans la matrice, appelée coefficient DCT et notée $F(u, v)$, correspond à son amplitude. L'ensemble de ces valeurs constituent les coefficients DCT.

Les basses fréquences sont localisées dans le coin supérieur gauche et les hautes fréquences dans le coin inférieur droit.

La deuxième étape, où se produit la perte d'information, est la quantification. Chaque coefficient DCT, $F(u, v)$, est réduit par un facteur de quantification, $Q(u, v)$, dépendant de ses coordonnées dans la matrice, c'est-à-dire de la fréquence. Ces facteurs forment la matrice de quantification. La même est utilisée pour le traitement de tous les blocs 8×8 issus de l'image.

Le résultat est une nouvelle matrice, d'entiers, les coefficients DCT quantifiés, $F'(u,v)$. Le but est de créer un maximum de redondance, notamment un grand nombre de valeurs nulles. La matrice de quantification s'appuie sur une matrice standard et le facteur de qualité de la compression JPEG. Ce facteur détermine la qualité relativement à l'image originale. Il amplifie ou réduit les valeurs des facteurs de quantification pour augmenter ou diminuer les dégradations.

Enfin, le codage entropique compresse l'image. En utilisant le code de Huffman, les coefficients DCT quantifiés sont transformés en une forme plus compacte. En particulier, les coefficients nuls créés par la quantification améliore grandement le taux de compression.

3 Insertion de données basée sur la DCT

La stéganographie désigne la dissimulation d'une information dans un autre support. Elle possède trois grandes caractéristiques qui dirigent ses utilisations : la robustesse, l'invisibilité et la capacité. La robustesse assure que l'information secrète ne peut pas être détruite sans dégrader fortement l'image. L'invisibilité vise à ce que l'image ne soit pas perturbée par l'information secrète. La capacité définit la quantité d'information qui peut être intégrée dans le support.

Ces trois caractéristiques sont en relation étroite et inverse. Par exemple, l'amélioration de la capacité a généralement une influence négative sur la robustesse.

L'insertion de données nécessite une bonne invisibilité et une grande capacité. Elle s'oppose au tatouage (watermarking), par exemple, qui propose une grande robustesse pour un message très court, d'une centaine de bits. Elle ne peut pas prétendre à la même robustesse et à répondre à des attaques ciblées. Par contre, il est intéressant qu'elle supporte des manipulations plutôt courantes sur les images, comme la compression.

En général, la robustesse à la compression JPEG s'obtient en travaillant dans le même domaine fréquentiel, celui de la DCT. Les méthodes d'insertion de données existantes suivent ce principe.

3.1 Méthodes existantes

L'insertion de données dans les composantes continues (DC) se base sur la connaissance de l'influence de cette composante : elle est proportionnelle à la moyenne de l'intensité des pixels du bloc. L'insertion d'un bit secret par substitution du LSB de la composante continue conduit donc à des modifications bien définies.

Cette méthode possède deux avantages : sa simplicité et sa robustesse. Au contraire, sa capacité est médiocre puisque seul un bit est intégré dans chaque bloc de 8×8 pixels. Les données insérées ne peuvent dépasser 0.19% de la taille de l'image. C'est ce point que les méthodes suivantes cherchent à améliorer.

La méthode Jpeg/Jsteg [2] étend l'idée précédente à tous les coefficients DCT de valeur absolue supérieure à 1. La restriction sur le choix des coefficients préserve les valeurs nulles pour conserver une bonne compression.

Cette méthode est basée sur le contenu de l'image. En effet, un bloc homogène contient peu de coefficients non nuls et intégrera moins de données qu'un bloc hétérogène qui contient de tels coefficients en plus grand nombre. Ce principe coincide avec un constat évident sur le système visuel humain (SVH) : les modifications sont moins visibles dans les zones hétérogènes que dans les zones homogènes.

Le principale point négatif de cette méthode reste l'incertitude sur la capacité. Elle peut varier de 0% (pour une image homogène de gris moyen) à 12,5% de la résolution de l'image. Elle reste néanmoins supérieure à la capacité de l'insertion de données dans la composante DC pour des images réelles. Comme la quantitié d'information insérée est plus importante, la méthode s'avère aussi moins robuste.

L'insertion de données par modification de la matrice de quantification [3]. En fonction du facteur de qualité choisi pour la compression JPEG, la variation d'un coefficient DCT quantifié peut conduire à d'importants changements dans l'image. Cette méthode d'insertion de données propose de choisir des coefficients qui ne seront pas quantifiés et pourront intégrer l'information sans dégrader l'image. La capacité est connue à l'avance et dépend simplement du nombre de coefficients non quantifiés. La méthode originale s'appuie sur 26 coefficients choisis parmi les moyennes fréquences dans chaque bloc. Elle insère deux bits secrets par substitution sur les deux LSBs. La capacité atteint alors 10.16% de la taille de l'image.

La matrice de quantification est présente dans le fichier compressé JPEG pour reconstruire l'information. L'invisibilité n'est pas entièrement satisfaisante puisqu'il suffit de relire ce fichier pour détecter l'utilisation de la méthode.

L'insertion de données adaptée au SVH expoite plus significativement que la méthode Jpeg/Jsteg les carences visuelles de l'oeil humain [4]. L'idée est, là encore, d'insérer plus de données dans les régions hétérogènes. La première étape consiste justement à trier les blocs 8×8 de l'image selon ce critère. Ensuite, les données sont intégrées dans les mêmes coefficients que pour le Jpeg/Jsteg. Seulement, ce n'est plus un seul, mais plusieurs bits qui peuvent être intégrés en fonction de l'homogénéité du bloc. La capacité est améliorée comparativement au Jpeg/Jsteg, mais reste dépendante de l'image.

3.2 Point commun de ces méthodes : insertions de données par substitution après la quantification

Les méthodes présentées sont assez dissemblables, elles reflètent différentes approches, basées ou non sur le SVH, s'appuyant sur des coefficients DCT différents. Néanmoins, toutes ces méthodes insèrent les données après la quantification, comme présenté figure 2, et par substitution des LSBs. Cette substitution peut être formalisée par l'équation suivante (uniquement pour le LSB) :

$$F'_1(u,v) = F'(u,v) - F'(u,v)\%2 + b_t, \qquad (1)$$

où $x\%y$ est le reste de la division entière de x par y, et b_t, le bit à insérer.

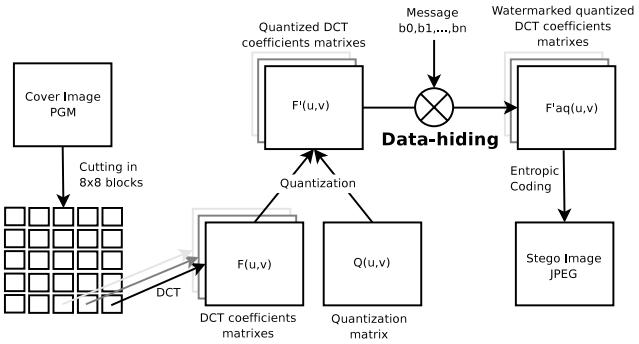


FIG. 2 – Insertion de données après quantification.

Cette approche est la plus simple, mais elle ajoute une perte d'information supplémentaire au processus de compression : l'insertion des données cachées.

4 Quantification adaptée aux données à insérer

En groupant les deux pertes d'information, la quantification et l'insertion de données, nous contrôlons mieux la perte finale et pouvons la réduire substantiellement.

4.1 Principes

Après la quantification, si le bit secret est différent du LSB, deux entiers sont à considérer : celui juste supérieur et celui juste inférieur au coefficient DCT quantifié. Il est impossible de les départager puisqu'ils sont à même distance de la valeur entière actuelle. La substitution est alors le choix arbitraire de l'un d'entre eux, en l'occurence, celui qui conduira à ne pas modifier les autres bits du coefficient.

Au contraire, avant quantification, un coefficient DCT est une valeur flottante. Nous pouvons alors évaluer l'entier le plus proche qui conduira à un coefficent DCT quantifié possèdant le bit secret comme LSB. Ce choix va réduire l'approximation globale.

Cette explication est valable pour le LSB et s'étend facilement à des substitutions sur plusieurs LSBs. Elle se formalise par les équations suivantes (toujours pour le LSB) :

$$F_2'(u,v) = \begin{cases} \left\lfloor \frac{F(u,v)}{Q(u,v)} \right\rfloor \text{ si } \left\lfloor \frac{F(u,v)}{Q(u,v)} \right\rfloor \%2 = b_t \\ \left\lceil \frac{F(u,v)}{Q(u,v)} \right\rceil \text{ si } \left\lceil \frac{F(u,v)}{Q(u,v)} \right\rceil \%2 = b_t \end{cases} . \quad (2)$$

Exemple

- *Bit secret :* $b_t = 1$,
- *valeur du coefficient :* $F(u,v) = 96$,
- *facteur de quantification associé au coefficient :* $Q(u,v) = 10$,
- *Valeur du coefficient quantifié :* $F'(u,v) = \left[\frac{96}{10}\right] = 10$.

Avec l'approche usuelle, la valeur du coefficient quantifié intégrant le bit serait : $F_1'(u,v) = 11$.

Les deux valeurs les plus proches du coefficient DCT qui conduisent à des valeurs entières après quantification sont 90 et 100. Nous choisissons la valeur qui, après quantification, aura son LSB égal à b_t*, le bit secret. Nous avons alors pour l'approche par adaptation de la quantification :* $F_2'(u,v) = 9$.

Les deux approches conduisent à des résultats différents. Adapter la quantification conduit à une valeur plus proche de la valeur flottante du coefficient DCT quantifié, 9.6.

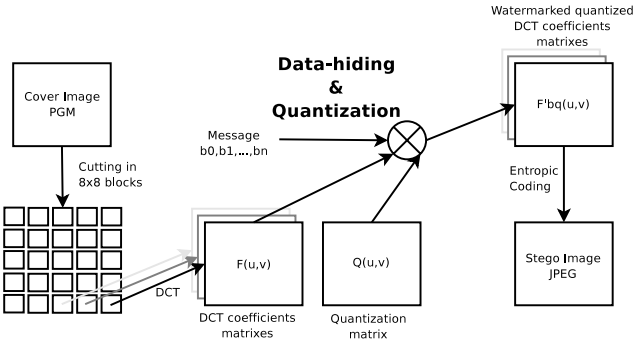


FIG. 3 – Adaptation de la quantification aux données à insérer.

L'adaptation de la quantification aux données à insérer est illustrée figure 3.

4.2 Résultats théoriques

4.2.1 Domaine fréquentiel

Encore une fois, nous ne prenons en compte que le LSB. Les différents cas sont présentés dans le tableau 1 : l'adaptation de la quantification est toujours au moins aussi performante que l'insertion après quantification. Les coefficients DCT quantifiés sont moins modifiés et l'image moins perturbée.

4.2.2 Domaine Spatial

Les résultats dans le domaine fréquentiel attestent d'une meilleure conservation de l'image avec l'adaptation de la quantification, mais sans information quantitative sur ce gain. Les résultats dans le domaine spatial, c'est-à-dire, sur les variations d'intensités des pixels, vont combler cette lacune. Nous nous intéressons donc à deux indicateurs, l'Erreur Quadratique Moyenne (EQM) et le Pic du Rapport Signal à Bruit (PSNR). C'est principalement le PSNR qui nous intéresse, mais la EQM est nécessaire pour l'utiliser.

L'EQM, pour un support de taille $N \times N$, peut être exprimée par :

$$EQM = \frac{\sum_{i=0}^{N-1} \sum_{j=0}^{N-1} (p(i,j) - p'(i,j))^2}{N^2}, \quad (3)$$

où $p(i,j)$ et $p'(i,j)$ sont respectivement l'intensité du pixel (i,j) de l'image originale et de l'image marquée,

TAB. 1 – Comparaison des variations induites entre l'insertion de données après ou pendant la quantificiation.

$\left\lfloor \frac{F(u,v)}{Q(u,v)} \right\rfloor$			$\left\lfloor \frac{F(u,v)}{Q(u,v)} \right\rfloor$		$\left\lceil \frac{F(u,v)}{Q(u,v)} \right\rceil$			
$\left\lfloor \frac{F(u,v)}{Q(u,v)} \right\rfloor \%2$	0		1		0		1	
b_t	0	1	0	1	0	1	0	1
Après la quantification : $\lvert\Delta F'_1(u,v)\rvert$	δ	$1-\delta$	$1+\delta$	δ	$1-\delta$	$2-\delta$	δ	$1-\delta$
Pendant la quantification : $\lvert\Delta F'_2(u,v)\rvert$	δ	$1-\delta$	$1-\delta$	δ	$1-\delta$	δ	δ	$1-\delta$
Gain : $\lvert\Delta F'_1(u,v)\rvert - \lvert\Delta F'_2(u,v)\rvert$	0	0	$\mathbf{2}\delta$	0	0	$\mathbf{2}-\mathbf{2}\delta$	0	0

δ, partie flottante de $\frac{F(u,v)}{Q(u,v)}$: $\delta = \frac{F(u,v)}{Q(u,v)} - \left\lfloor \frac{F(u,v)}{Q(u,v)} \right\rfloor$

et le PSNR par :

$$PSNR = 10 \log_{10} \frac{255^2}{EQM}. \quad (4)$$

Le PSNR donne une idée de la qualité objective par rapport à l'image originale. Plus il est élevé, plus l'image marquée et l'originale sont proches.

Des résultats théoriques sont difficiles à obtenir. Ils sont très dépendants du contenu de l'image et nécessitent de trouver les modélisations probabilistiques adaptées. Nous ne nous attarderons pas sur cette partie et l'illustrerons plutôt par des résultats pratiques.

4.3 Application à la méthode Jpeg/Jsteg

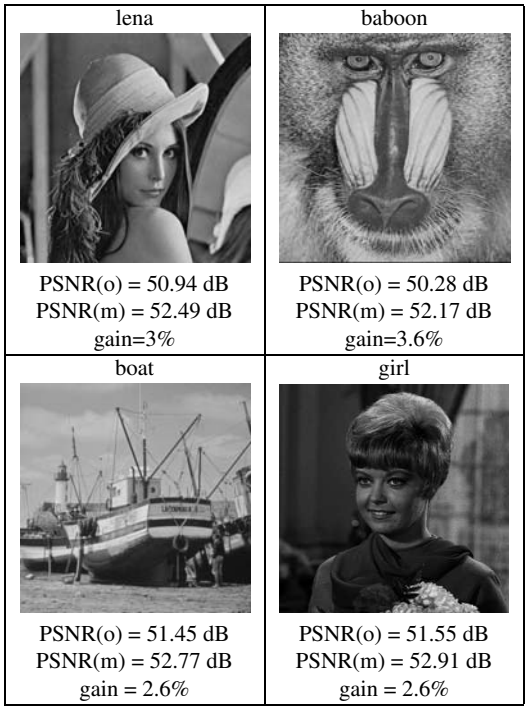


FIG. 4 – Comparaison de la qualité pour des images marquées obtenues avec Jpeg/Jsteg (o) et Jpeg/Jsteg adaptatif (m) pour un facteur de qualité de 100%.

Nous avons implémenté les différentes méthodes en adaptant la quantification aux données à insérer. Les résultats pour le Jpeg/Jsteg sont présentés figure 4. Pour chaque image, avec un facteur de qualité de 100%, le gain est conséquent et dépasse les $2,5\%$.

Les résultats sont similaires pour les autres méthodes : la qualité visuelle objective est améliorée.

Il est important de noter que le facteur de qualité de la compression JPEG joue un rôle important. Si il est élevé, les erreurs induites par la quantification reste faible et le gain obtenu est visible. Au contraire, si il est faible, le gain est réduit par le bruit trop important généré par la quantification : pour la méthode Jpeg/Jsteg, avec un facteur de qualité de 50%, le gain chute à moins de 1%.

5 Conclusion

Nous proposons une amélioration pour les méthodes d'insertion de données basées sur la DCT. Au lieu d'être intégrée après la quantification et par simple substitution, l'information à insérer guide la quantification. La qualité visuelle est améliorée dans des proportions dépendantes du facteur de qualité de la compression JPEG.

Nous avons présenté des résultats et des applications uniquement pour la substitution du LSB, mais ils peuvent facilement être étendus à la substitution sur les LSBs.

Références

[1] G. Wallace. *The JPEG still picture compression standard*. Communication of the ACM, Apr. 1991, volume 34(4), pp 31-44.

[2] C.-T. Hsu et J.-L. Wu. *Hidden Digital Watermarks in Images*. IEEE Transaction on Image Processing, 1999, volume 8, pp 58-68.

[3] C.-C. Chang, T.-S. Chen et L.-Z. Chung. *A Steganographic Method Based upon JPEG and Quantization Table Modification*. Information Sciences, 2002, volume 141, pp 123-138.

[4] H.-W. Tseng and C.-C. Chang. *High Capacity Data Hiding in JPEG-Compressed Images*. Informatica, 2004, volume 15(1), pp 127-142.

Stéganographie d'images basées sur le théorème chinois des restes

Dinu COLTUC[1], Alain TRÉMEAU[2]

[1]FIE, Université Valahia de Targoviste
BP 16-296, Bucarest, Roumanie

[2]LIGIV, Université Jean Monnet
10, rue Barrouin, 42000 Saint Etienne, France
coltuc@valahia.ro, tremeau@ligiv.org

Résumé – Dans ce papier nous nous intéressons au problème de cacher dans une image, moyennant une seule opération, de multiple messages provenant de multiples utilisateurs. L'idée est d'utiliser le théorème chinois des restes. Sur le plan théorique, il est établi qu'un nombre illimité de messages peuvent être insérée dans l'image elle-même. Une des limites de l'approche vient du fait que ces multiples messages ne doivent être ni détectables, ni visibles. Par construction, la première condition est naturellement satisfaite, par contre toute la difficulté consiste à satisfaire la seconde condition. Un atout de cette approche est que chaque utilisateur peut disposer d'une clé différente pour détecter une des marques superposée à l'image.

Abstract – This paper investigates the problem of embedding, in the same cover image, multiple messages for multiple receiving parties. A solution based on the Chinese Reminder Theorem is given. Theoretically, an unlimited number of messages can be inserted. The security of the scheme is provided by distinct secret key for each receiver. Experimental results are provided.

1 Introduction

Dans ce papier nous nous intéressons au problème de cacher dans une image, moyennant une seule opération, de multiple messages provenant de multiples utilisateurs. Chaque utilisateur dispose d'une clé secrète qui lui est propre.

Pour illustrer l'insertion de multiple messages prenons tout d'abord un des cas d'étude les plus simples; celui de deux messages (*codewords*) que l'on veut superposer à un couple de pixels a et b. Soit m_1 et m_2 les deux clés secrètes considérées, et r_1 et r_2 les deux messages associés. Considérons que $r_1 < m_1$ et $r_2 < m_2$. L'idée est d'ajouter une certaine quantité à l'un des pixels (e.g. a) de façon que le reste de la différence entre les deux pixels (i.e. a - b) divisé par m_1 soit égal à r_1. Ensuite, il suffit d'ajouter une certaine quantité à chacun des deux pixels (e.g. a et b) de façon que le reste de la somme des deux pixels (i.e. a + b) divisé par m_2 soit égal à r_2. Un tel approche permet l'insertion simultanée de deux messages dans tout groupe de deux pixels.

Ce papier propose une méthode pour superposer K messages à une image, avec $K \geq 2$. Notre approche est basée sur le principe du théorème chinois des restes [1]. Reste ensuite, comme dans tout processus de stéganographie sécurisé, à s'assurer que les messages transmis n'ont pas été altérés par un processus donné. Nous verrons que compte tenue de la stratégie de marquage que nous avons utilisé, la résolution de ce dernier point est par définition immédiate.

Le plan de ce papier et le suivant. Le théorème chinois des restes est brièvement rappelé dans la deuxième section. Dans la troisième section, nous donnons les principes de base de notre approche. Des résultats expérimentaux sont fournis dans la quatrième section. Finalement, dans la cinquième section nous présentons nos conclusions.

2 Théorème chinois des restes

Le théorème chinois des restes démontre l'existence d'un nombre inconnu qui, divisé par un ensemble de diviseurs, donne un ensemble de restes souhaités. Soit N ce nombre, m_i les valeurs de chacun des diviseurs, et r_i, $i = 1, \ldots, K$ chacun des restes modulo les valeurs des diviseurs, on a alors :

$$N \bmod m_i = r_i,\ i = 1, \ldots, K \tag{1}$$

Selon le théorème chinois des restes, si l'ensemble des valeurs de modulo (i.e. les m_i) forme un ensemble de nombres premiers entre eux alors le système d'équations donné par (1) possède une solution unique N, $1 \leq N \leq M$ où :

$$M = m_1 m_2 \ldots m_K \tag{2}$$

L'existence et l'unicité de N peut être facilement démontrée mathématiquement en trouvant un ensemble de K nombres M_i ayant les deux propriétés suivantes :

$$M_i \bmod m_i = 1,\ i = 1, \ldots, K \tag{3}$$

et

$$M_i \bmod m_j = 0,\ i \neq j,\ i, j = 1, \ldots, K, \tag{4}$$

Donc, N est calculé par :

$$N = \left(\sum_{i=1}^{K} M_i r_i \right) \bmod M \tag{5}$$

La preuve est immédiate: les équations (1) peuvent être établies en prenant $N \bmod m_i$, $i = 1, \ldots, K$, et en utilisant les équations (3) et(4).

Dans le but de trouver l'ensemble M_i, on peut remarquer que M_i est divisible par tous les modulo à l'exception de m_i, c'est à dire :

$$M_i = A\frac{M}{m_i} \tag{6}$$

De plus, la division de M_i par m_i donner un reste égal à 1 :

$$M_i = Bm_i + 1 \tag{7}$$

Les équations (6) et (7) entraînent que :

$$A\frac{M}{m_i} - Bm_i = 1 \tag{8}$$

ou, évidement, $\frac{M}{m_i}$ est un entier.

L'équation (8) est équivalente à l'équation classique :

$$Ap + Bq = 1. \tag{9}$$

L'algorithme Euclidien décrit dans [1] peut être ensuite utilisé pour trouver les valeurs A et B. Enfin, en utilisant l'équation (6), (ou l'équation (7)), l'ensemble M_i est trouvé et, par conséquence, on détermine N.

3 Principes de base de l'algorithme

Soit K le nombre de messages. Supposons que chaque message R_i, $i = 1, \ldots, K$, soit de même longueur P. Soit r_i^j la $j^{\text{ème}}$ valeur ($j = 1, \ldots, P$) du $i^{\text{ème}}$ ($i = 1, \ldots, K$) message. Pour chaque message, un diviseur m_i est associé. Afin de satisfaire le théorème chinois des restes, tous les diviseurs m_i doivent former des nombres premiers entre eux et tous les restes doivent vérifier $r_i^j < m_i$, $j = 1, \ldots, P$.

Partitionnons tout d'abord l'image en groupes de p pixels, avec $p \geq 1$. Considérons, pour des raisons de simplicité, que le nombre de groupes que l'on va définir est égal à la longueur des messages P que l'on veut appliquer. Pour chaque message (et chaque utilisateur) une fonction de permutation secrète Π_i est définie, $\Pi_i : \{1, 2, \ldots, P\} \rightarrow \{1, 2, \ldots, P\}$. Considérons que Π_0 corresponde à la permutation identité.

3.1 Clé secrète

L'algorithme de marquage que l'on propose repose sur l'utilisation d'une clé secrète qui porte sur trois éléments:

1. le partitionnement du plan image en P groupes de p pixels, quand $p \geq 1$;
2. la fonction de permutation Π_i;
3. les diviseurs m_i.

La partition est identique quelque soit l'utilisateur. Ce dispositif présente deux niveaux de sécurité. Le premier, global au système, vient du fait que l'image est partitionnée en groupes de p pixels (quand $p \geq 1$), l'autre spécifique à l'utilisateur, vient des valeurs m_i et Π_i.

Les valeurs m_i ne permettent pas un haut degré de sécurité puisque le nombre de possibilité quant au choix de ces valeurs est extrêmement limité. Par contre, la fonction de permutation Π_i et la fonction de partition renforcent nettement cette sécurité. Pour le moment, le nombre de permutations est en $P!$ ou P est le nombre de groupes pixels dans l'image. De même, un très grand nombre de partitions en groupes de p pixels peut être défini pour un ensemble donné de P groupes. L'ordre dans lequel sont traités ces p pixels à l'intérieur du groupe a aussi une répercussion sur le nombre de combinaisons.

FIG. 1: Image de test.

3.2 Marquage

L'algorithme de marquage est alors le suivant :

1. générer les messages à partir des permutations secrètes $\Pi_i(R_i)$;
2. générer le $j^{\text{ème}}$ élément de la séquence N_j, $j = 1, \ldots, P$, à partir du théorème chinois des restes, i.e. en fonction de l'ensemble des valeurs de modulo précédemment utilisées;
3. insérer chaque message N_j dans chacun des groupes de p pixels à partir de la permutation identité Π_0.

Etant donné que $r_i^j < m_i$ et que le message N_j est inséré dans p pixels, la capacité du schéma est de :

$$C = \frac{\sum_{i=1}^{K} \log_2 m_i}{p} \text{ bpp} \tag{10}$$

On peut remarquer que M croit avec K. Donc, pour un p fixé, C croit également. L'effet de la croissance du taux d'information conduit évidement à une perte de qualité de l'image stéganographiée. Dans le but de contrôler cette qualité on peut restreindre le taux d'information par pixel a une certaine valeur q imposée et déterminer la taille de p :

$$p \geq \left\lceil \frac{\log_2 M}{q} \right\rceil \tag{11}$$

ou $\lceil x \rceil$ est le plus petit entier plus grand ou egal à x.

FIG. 2: Image stéganonographiée avec 2 messages par pixel, 2.58 bpp, 43 dB.

FIG. 3: Image stéganographiée avec 3 messages par pixel, 4.9 bpp, 29 dB.

3.3 Detection

L'algorithme de détection se définit assez facilement de la manière suivante pour l'utilisateur i :

1. extraire le $j^{\text{ème}}$ élément de la séquence N_j, $j = 1,\ldots,P$, à partir de la partition en groupes de p pixels (si $p > 1$) et de la clé secrète de permutation Π_i;
2. extraire le message $r_i^j = N_j \bmod m_i,\ j = 1,\ldots,P$.

Etant donné que la séquence N_j est accessible à tous les utilisateurs, on pourrait penser pouvoir détecter tous les messages destinés à tout autre utilisateur que soit même. Quand bien même un utilisateur pourrait retrouver à partir de sa valeur de modulo l'ensemble de la séquence des messages, i.e. l'ensemble des valeurs de modulo secrètes, il ne pourrait décoder celles-ci car il ne saurait pas dans quel ordre ceux-ci ont été constitués. Comme nous l'avons déjà indiqué, retrouver à partir de tests exhaustifs l'ensemble des permutations effectuées relève d'une opération d'une extrême complexité.

4 Résultats experimentaux

La méthode présentée dans la section précédente produit à partir de K messages une séquence de entiers N_j qu'on doit insérer dans un ou plusieurs pixels. Un grand nombre des méthodes ont été développées pour l'insertion des données [2, 3, 4]. Parmi ces méthodes, les méthodes spatiales qui remplacent les bits de poids faible sont les plus simples. Pour la mise en oeuvre de notre méthode nous allons utiliser une telle méthode. Dans ce papier nous avons choisi de présenter seulement quelques résultats illustratifs, ceux-ci ont été obtenus à partir de l'image test *Lena* de taille 512×512. (Fig. 1).

Dans un premier temps nous avons considéré l'insertion simultanée de deux messages en chaque pixel. Dans le premier exemple que nous donnons, nous avons utilisé les paramètres suivants : $m_1 = 2$, $m_2 = 3$ et $p = 1$. Afin d'insérer les N_j séquences dans chaque pixel de l'image une procédure assez simple a été utilisée. Les pixels ont été quantifiés de façon à avoir des multiples de M, on a ensuite ajouté à chaque pixel la valeur correspondante à N_j. Comme on peut le constater sur la Fig. 2, l'image stéganographiée résultante est de bonne qualité; son PSNR est de 43.1 dB. Le taux d'information cachée est ici de 2.58 bits par pixel.

Le second exemple que nous donnons porte sur l'insertion simultanée de trois messages. Dans cet exemple nous avons utilisé les paramètres suivants : $m_1 = 2$, $m_2 = 3$, $m_3 = 5$ et $p = 1$. En ajoutant un troisième message on a augmenté du taux d'information cachée, passant ainsi de 2.58 bits par pixel à 4.9 bits par pixel. Comme cela était attendu la qualité de l'image stéganographiée résultante a diminué, l'image est devenue de moins bonne qualité (cf. Fig. 3); son PSNR est tombé à 29.16 dB.

Les premiers tests que nous avons réalisés ont porté sur l'insertion de deux ou de plusieurs messages en chaque pixel. Afin d'améliorer la qualité de l'image stéganographiée ou d'augmenter le nombre de messages insérés dans l'image, nous avons ensuite étendu notre schéma d'insertion à un groupe de p pixels, $p > 1$.

Dans l'exemple que nous donnons, nous avons considéré l'insertion simultanée de trois messages pour tous les groupes de 2 pixels. Nous avons utilisé les paramètres suivants : $m_1 = 2$, $m_2 = 3$, $m_3 = 5$. Ces paramètres sont exactement les mêmes que ceux utilisés dans l'exemple précédent pour l'insertion de trois messages en chaque pixel. Etant donné que l'insertion porte, maintenant non

FIG. 4: Image stéganographiée avec 3 messages groupe de 2 pixels, 2.45 bpp, 40.72 dB.

FIG. 5: Image stéganographiée avec 4 messages par groupe de 2 pixels, 3.85 bpp, 32.57 dB.

plus sur chaque pixel, mais sur un groupe de 2 pixels, le taux d'information cachée a *a fortiori* diminué d'autant et la qualité de l'image stéganographiée a augmenté. Nous avons obtenu un taux d'information cachée de 2.45 bpp et un PSNR de 40.72 dB (cf. Fig. 4).

Nous pouvons augmenter le nombre de messages distincts que l'on peut superposer à des groupes de 2 pixels. L'exemple de la Fig. 5 est obtenu pour les paramètres suivants : 4 messages ($m_1 = 2$, $m_2 = 3$, $m_3 = 5$ et $m_4 = 7$). Dans ce conditions, pour une insertion dans des groupes de 2 pixels, nous avons obtenu un taux d'information cachée de 3.85 bpp et un PSNR de 32.57 dB.

5 Conclusion

Dans ce papier nous avons proposé une méthode d'insertion de multiple messages pour différents utilisateurs qui repose sur le principe du théorème chinois des restes. Afin d'illustrer le principe de notre méthode, un schéma d'insersion simple a été proposé. La sécurité de ce schéma est principalement assurée par l'utilisation d'une fonction de permutations secrète d'un ensemble de P éléments, ou P représente le nombre de pixels de l'image (ou une fraction de l'image). Un niveau supplémentaire de sécurité est assuré par l'utilisation d'une fonction de partition secrète de l'image pour le cas ou l'insertion porte non pas sur chaque pixel mais sur des groupes de p pixels. Différents exemples d'insertion simultanée de deux ou trois messages, en chaque pixel ou sur des groupes de 2 pixels ont été donné. Les résultats ainsi obtenus semblent prometteurs. D'autres travaux sont en cours, ceux-ci portent notamment sur la conception de schémas d'insertion plus complexes. L'objectif est d'optimiser le rapport capacité d'insertion / nombre de messages insérés, sous la contrainte de ne pas trop dégrader la qualité de l'image résultante, d'où la nécessité d'affiner la procédure d'insertion à partir de groupes de p pixels. L'extension aux images couleur constitue en ce sens une piste intéressante, en effet grâce à l'information couleur on peut plus facilement jouer sur une minimisation de la perte de qualité d'une image, compte tenu de la sensibilité du système visuel humain à percevoir plus facilement certaines différences images que d'autres.

References

[1] D. E. Knuth, *The Art of Computer Programming. Seminumerical Algorithms*, Third Edition, Addison Wesley, 1998.

[2] F. A. P. Petitcolas, R. J. Anderson, M. G. Kuhn, "Information Hiding: a Survey", *Proceedings of the IEEE*, vol. 87 no. 7, pp. 1062-1078, July 1999.

[3] F. Hartung, M. Kutter, "Multimedia Watermarking Techniques", *Proceedings of the IEEE*, vol. 87 no. 7, pp. 1079-1107, July 1999.

[4] B. Chen, G. W. Wornell, "Quantization Index Modulation: A Class of Provably Good Methods for Digital Watermarking and Information Embedding", *IEEE Trans. on Information Theory*, vol. 47 no. 4, pp. 1423-1443, May 2001.

Tatouage vidéo utilisant l'ACI et basé sur un treillis informé

Hussein JOUMAA, Franck DAVOINE

Laboratoire Heudiasyc, CNRS, Université de Technologie de Compiègne
BP 20529, 60205 Compiègne Cedex.
Joumaahu@hds.utc.fr, Franck.Davoine@hds.utc.fr

Résumé – L'Analyse en Composantes Indépendantes (ACI) a été utilisée récemment dans plusieurs schémas de tatouage. Il s'avère important d'évaluer la robustesse d'un système de tatouage dans cet espace et de la comparer par rapport à celles réalisées dans d'autres espaces, notamment celui obtenu par la Transformée en Cosinus Discrète (TCD). Ainsi, nous proposons dans cet article d'évaluer la performance d'un schéma de tatouage informé, utilisant l'ACI, et de montrer ses principaux intérêts et inconvénients par rapport à d'autres domaines. Pour ce faire, nous présentons une méthode de tatouage basée sur le treillis modifié d'un code convolutif, et opérant dans deux espaces transformés à l'aide d'une TCD ou d'une ACI. Le système ainsi proposé assure, dans les deux domaines, un niveau de robustesse important, ainsi qu'une capacité d'insertion intéressante pour les applications d'augmentation de contenu. Dans cet article, nous nous intéressons à l'application suivante : le tatouage constitue un canal caché transmis au sein d'un canal de TV numérique, compressé par MPEG-2

Abstract – Independent Component Analysis (ICA) techniques have been recently used in different watermarking schemes. However, performance of an ICA video watermarking scheme in comparison with those using classical domains, such as the Discret Fourier Transform (DCT) domain, is still not clear. In this paper, we attempt to fill this gap. Therefore, we propose a video watermarking scheme, using an informed trellis, applied in two transformed domains obtained by using respectively the DCT transform and an ICA coding technique. We show that, for both domains, the scheme offers a good robustness against MPEG-2 compression, as well as an important capacity level. We consider in this paper data hiding in digital TV channels where data are compressed using MPEG-2.

1 Introduction

L'Analyse en Composantes Indépendantes (ACI) a été utilisée dans plusieurs schémas de tatouage pour images fixes [3] [11]. Il est nécessaire d'évaluer la robustesse d'un système de tatouage dans cet espace. Nous proposons dans cet article d'évaluer la performance d'un schéma de tatouage informé, utilisant l'ACI, et de montrer ses principaux intérêts par rapport à d'autres domaines, notamment celui obtenu par la Transformée en Cosinus Discrète (TCD). Nous présentons ainsi une méthode de tatouage basée sur le treillis modifié d'un code convolutif (Section 2), et opérant dans deux espaces transformés à l'aide d'une TCD (Section 3) ou d'une ACI (Section 4).
Dans les deux schémas, l'insertion est effectuée dans les images d'erreur de prédiction de mouvement d'une séquence vidéo. Nous prenons ainsi en compte la dimension temporelle de la vidéo et nous améliorons par conséquent la performance du schéma de tatouage. En effet, les premières méthodes de tatouage ont été conçues pour des images fixes. Le domaine de tatouage vidéo a ensuite gagné une grande attention, et beaucoup de chercheurs se sont intéressés à définir des schémas de tatouage qui lui sont propres. La plupart d'entre eux avaient déjà développé des algorithmes pour les images fixes, il était donc naturel qu'ils appliquent, dans un premier temps, leurs algorithmes sur la vidéo en considérant celle-ci comme une série d'images indépendantes [4] [1].
Une deuxième classe de méthodes de tatouage vidéo prend en compte l'information temporelle présente dans la vidéo et l'exploite [14] [6]. La troisième classe effectue le tatouage directement dans le flux compressé [10] [2]. Elle nécessite une maîtrise de la structure du codeur utilisé dans la compression, afin de pouvoir modifier quelques paramètres du flux d'une façon appropriée.
Nous avons donc choisi de travailler dans le cadre de la première classe de méthodes, tout en exploitant la dimension temporelle de la vidéo : le message est inséré dans les images d'erreur de prédiction de mouvement afin d'exploiter la dimension temporelle de la séquence.
Dans cet article, nous nous intéressons à l'application suivante : le tatouage constitue un canal caché transmis au sein d'un canal de TV numérique, compressé par MPEG-2. L'article se termine par la fourniture de résultats de comparaison de robustesse entre les deux schémas utilisant respectivement la TCD et l'ACI.

2 Une méthode de tatouage vidéo informée

Dans cette partie, nous présentons une technique d'insertion informée inspirée du schéma proposé par Miller et al. [12] pour le tatouage d'images fixes.
Dans un système de tatouage informé, on associe à chaque message $\mathbf{m}$ un "sous-dictionnaire de codage" CB_m. Lors de la phase d'insertion, $\mathbf{m}$ est codé par le vecteur de CB_m le plus proche du signal hôte. Quant au décodage, il s'effectue au niveau du "dictionnaire principal" constitué de l'union des différents sous-dictionnaires : l'appartenance du vecteur décodé à un sous-dictionnaire particulier permet d'identifier le message incrusté dans le signal hôte.
Les codes correcteurs d'erreur permettent de construire et d'organiser le dictionnaire principal, ainsi que les différents sous-dictionnaires de codage [8]. La technique utilisée dans cet ar-

ticle repose sur le treillis pleinement connecté d'un code convolutif constituant le dictionnaire principal d'un système de tatouage informé. Les sous-dictionnaires de codage sont ensuite construits en élaguant le treillis en fonction du message à insérer.
Dans le cas d'un code convolutif traditionnel, représenté par un treillis, un message $\mathbf{m}$ de k bits est associé à un chemin particulier du treillis : $\mathbf{m}$ est ainsi codé par une suite de k branches du treillis. Dans le treillis, deux branches sont issues de chaque nœud : la première correspond à un bit 0 de $\mathbf{m}$ et la seconde à un bit 1. Une branche peut coder un vecteur aléatoire de longueur n_b et ainsi, en sortie du treillis, un message de k bits est codé par un vecteur de taille $k \times n_b$.
Pour construire un tatoueur informé à partir du système précédent, Miller et al. [12] ont associé à un message $\mathbf{m}$ plusieurs chemins du treillis. Ces associations définissent un sous-dictionnaire. Elles s'effectuent en commençant par construire un treillis pleinement connecté T qui constitue le dictionnaire principal : chaque nœud est connecté à tous les nœuds qui le précèdent et qui le suivent immédiatement. La moitié des branches porte un bit 0, l'autre un bit 1. Un message particulier $\mathbf{m}$ est ainsi représenté par plusieurs chemins dans le treillis : ces chemins constituent un treillis $T_{\mathbf{m}}$ correspondant au sous-dictionnaire $CB_{\mathbf{m}}$. Le treillis pleinement connecté T est constitué de l'union des treillis $T_{\mathbf{m}}$.

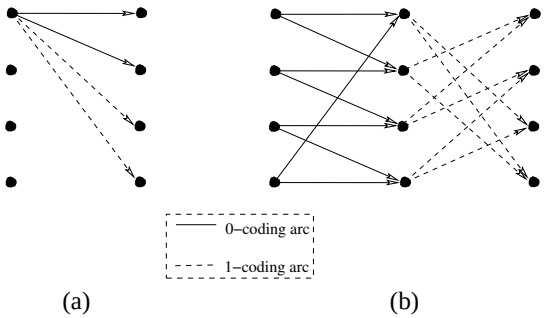


FIG. 1 – Treillis pleinement connecté à 4 états : Dans la figure (a), on représente seulement les arcs issus de l'état A. Chaque bit est codé par 2 manières différentes. Dans la figure (b), on représente le treillis T_{01} correspondant au message 01.

3 Tatouage vidéo dans le domaine TCD

Dans cette partie, nous présentons le schéma de tatouage vidéo utilisant la Transformée en Cosinus Discrète. Le message est inséré dans les images d'erreur de prédiction de mouvement afin d'exploiter la dimension temporelle de la séquence.
Pour préserver la qualité visuelle de la vidéo, nous avons choisi de ne pas tatouer toutes les images de la séquence 3. Nous avons choisi ainsi d'insérer le message dans des images d'erreur de prédiction de mouvement P_e, séparées par des images I (codées en Intra). Une image P_e tatouée est décomposée en des blocs 8×8, transformés par la TCD. Une série de n_1 coefficients est extraite de chaque bloc transformé, pour construire le signal hôte $\mathbf{x}$. On applique ensuite l'algorithme de Viterbi sur $\mathbf{x}$ pour trouver, parmi les vecteurs du sous-dictionnaire correspondant au message $\mathbf{m}$, le vecteur $\mathbf{w}$ le plus proche de $\mathbf{x}$. Le décodage de Viterbi est dans ce cas pondéré, et considère une mesure de corrélation pour trouver le chemin le plus vraisemblable. $\mathbf{w}$ est ensuite ajouté à $\mathbf{x}$ avec une pondération α permettant de régler la force du tatouage : $\mathbf{s} = \mathbf{x} + \alpha\mathbf{w}$.
Le décodage se fait en aveugle en appliquant l'algorithme de Viterbi sur un vecteur $\mathbf{y}$ extrait de la séquence reçue.

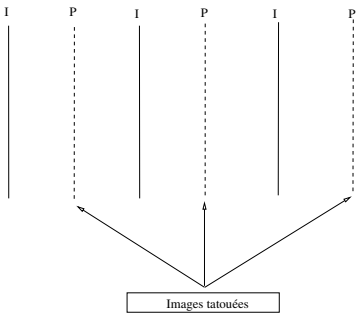


FIG. 2 – Codage par ACI

FIG. 3 – Le message est inséré dans des images d'erreur de prédiction de mouvement P_e, séparées par des images I (codées en Intra).

4 Tatouage vidéo dans le domaine ACI

Dans cette partie, nous présentons le schéma de tatouage vidéo utilisant l'Analyse en Composantes Indépendantes. Nous commençons par rappeler les principes de l'ACI, ainsi que les principaux schémas de tatouage utilisant l'ACI.
L'ACI est une technique statistique dont l'objectif est de décomposer un vecteur $\mathbf{x} \in \mathbb{R}^m$ en une combinaison linéaire de sources indépendantes, i. e. $\mathbf{x} = A\mathbf{s}$ où $\mathbf{s}$ est un signal à composantes indépendantes et A une matrice à coefficients réels. Cette technique est appliquée à des problèmes où les sources peuvent être supposées indépendantes. Le but de la technique est de trouver une matrice W telle que le vecteur obtenu par action de W sur $\mathbf{x}$ ait des composantes les plus indépendantes possibles.

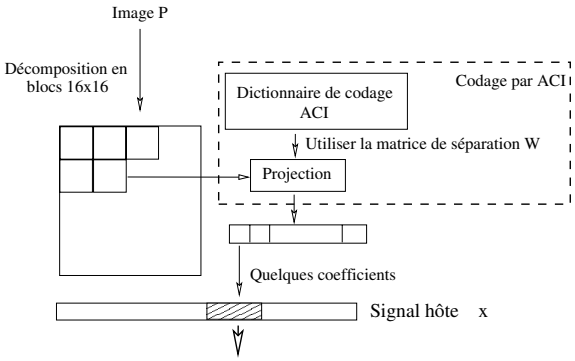


FIG. 4 – Codage d'une image P par ACI

Dans le cadre du tatouage, l'ACI a été utilisée par González-Serrano et al. [7]. Leur méthode consiste à modifier les composantes de faible amplitude du signal transformé. Elle est utilisée dans le cadre d'un schéma de tatouage fragile. Bounkong et al. [3] ont proposé une technique de tatouage basée sur la quantification des coefficients ACI d'une partition d'image. Chaque bit du message est inséré dans un seul coefficient ACI, afin de permettre une quantification forte des coefficients tatoués. Récemment, un schéma de tatouage pour images fixes, utilisant l'ACI, a été proposé par Lu et al. [11]. Une technique efficace est mise en oeuvre afin de choisir les coefficients ACI destinés à être tatoués : les coefficients sont classifiés suivant leur fréquence, en plusieurs catégories. Cette classification s'effectue en se basant sur leur amplitude. Les coefficients de faible amplitude représentent ainsi des composantes de haute fréquence de l'image.

Dans cet article, nous nous inspirons du schéma de tatouage proposé par Bounkong et al. [3], afin de proposer une méthode de tatouage vidéo, opérant dans un domaine transformé par ACI.

Pour mettre en oeuvre un algorithme de codage par ACI, il est nécessaire de construire un dictionnaire de codage formé d'un ensemble de fonctions de base, adapté au signal hôte. Cette étape a été réalisée à l'aide de l'algorithme FastICA [9]. Les images utilisées pour construire le dictionnaire (base des observations) sont extraites de séquences vidéo de scènes naturelles.

Nous avons choisi d'insérer le message dans les images d'erreur de prédiction de mouvement P_e, comme dans le cas du schéma précédent, appliqué dans le domaine TCD.

Une image P_e est ainsi décomposée en des blocs 8×8, transformés par ACI. Une série de n_1 coefficients est extraite de chaque bloc transformé, pour construire le signal hôte $\mathbf{x}$.

On a effectué différents tests pour définir une technique efficace du choix des coefficients :

- Une première approche consiste à ordonner les coefficients ACI suivant leur module décroissant, et à choisir pour le tatouage, les n_1 premiers coefficients de la série ordonnée. Cependant, l'ordre défini au codeur peut être altéré après une compression, résultant ainsi en une désynchronisation du décodeur.
- Une seconde approche, s'inspirant de l'article de Lu et al. [11], classifie les fonctions de base du dictionnaire suivant leur fréquence, afin d'appliquer le tatouage dans une bande de fréquence appropriée. Toutefois, pour les images d'erreur de prédiction de mouvement, la structure des fonctions de base est différente de celle obtenue pour les images fixes. Leur structure localisée ne permet pas de les classifier efficacement suivant le critère de fréquence.

Suite aux tests effectués, il parait plus intéressant de tatouer les coefficients correspondant à un ensemble aléatoire de fonctions de base.

Une fois le signal hôte $\mathbf{x}$ déterminé, les opérations d'insertion et de détection sont identiques à celles effectuées pour le schéma TCD (Section 3).

5 Résultats et discussion

Nous avons évalué la performance du système de décodage suite à une compression de type MPEG-2 [13].

Les tests ont été effectués sur les séquences de format QCIF : Suzie, News, Mother daughter, Hall monitor, Container, composée chacune de 150 images. Le nombre de bits insérés est de $k = 22$ bits par image. La qualité visuelle moyenne des séquences tatouées, mesurée par le PSNR, est fixée à 41 dB.

Pour le schéma appliqué dans l'espace transformé par TCD, on extrait 8 coefficients par bloc 8×8 transformé (Section 3, $n_1 = 8$). Ceci nous permet d'insérer 22 bits dans une image de taille 176×144. La performance de cette méthode, suite à une compression MPEG-2, est présentée sur la figure 5 représentant le pourcentage moyen des bits erronés au décodeur en fonction du taux de compression MPEG-2. On remarque que pour un taux de compression de 35, on est capable de décoder le message inséré avec un pourcentage d'erreur binaire de 2%.

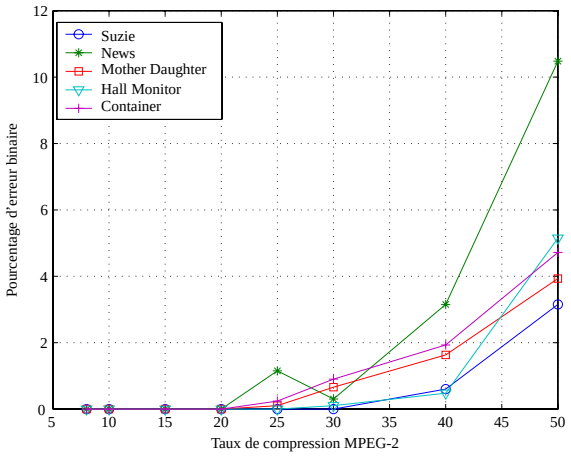


FIG. 5 – Robustesse de la méthode utilisant la TCD face à une compression MPEG-2

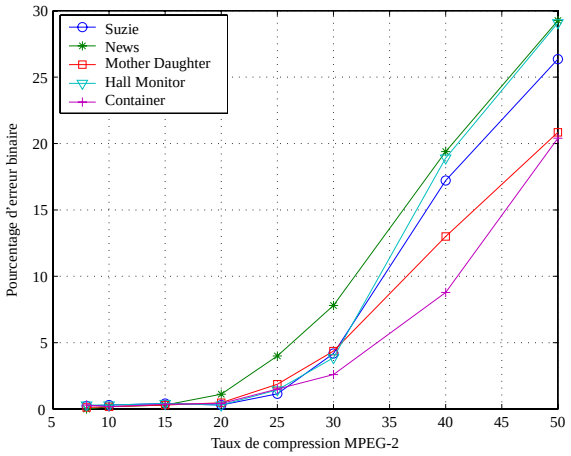


FIG. 6 – Robustesse de la méthode utilisant l'ACI face à une compression MPEG-2

Dans le second schéma, on extrait également 8 coefficients par bloc 8×8 transformé. On a pu ainsi insérer 22 bits dans une image de taille 176×144. La robustesse de cette méthode suite

à une compression MPEG-2 est présentée sur la figure 6. On remarque que pour un taux de compression de 25, on est capable de décoder le message inséré avec un pourcentage d'erreur binaire de 2%.
L'algorithme basé sur le treillis a été appliqué dans les domaines générés par les transformations TCD et ACI. On remarque que les deux schémas présentent, face à une compression MPEG-2, un niveau de robustesse important. Cependant, la méthode appliquée en TCD permet d'obtenir des niveaux de robustesse meilleurs que ceux obtenus dans le domaine généré par ACI.
Le fait que, dans le domaine TCD, on a une idée de la plage des coefficients qui va être dégradée par MPEG-2, nous permet d'obtenir une bonne performance : l'espace d'insertion est compatible avec celui de la compression par MPEG, ainsi le choix d'insérer le message dans la plage des fréquences moyennes s'avère très pratique, comme effectué dans la littérature des techniques de tatouage par étalement de spectre [5].
Néanmoins, pour le domaine généré par une transformation ACI, cette information ne peut pas être estimée à l'avance. On peut affirmer une seule chose : les coefficients significatifs du signal doivent être réservés, après compression. Pour chaque bloc codé, l'ACI associe à chaque fonction du dictionnaire de codage ACI, un coefficient correspondant. L'importance d'une fonction ou d'une autre, dans la représentation du bloc, est pondérée par le module de ce coefficient. Or, l'algorithme de tatouage prend toujours une plage fixe de fonctions de base, dans laquelle, il insère le tatouage, sans tenir compte de l'importance de ces fonctions dans la représentation. On peut ainsi tatouer des coefficients, non significatifs, qui s'annulent suite à la compression.
Nous avons essayé de choisir les coefficients à tatouer en fonction de leur module (la plage de coefficients tatoués n'est pas la même pour tous les blocs). Nous étions ainsi amené à établir un ordre donné des coefficients tatoués, selon leur module. Cependant, suite à la compression, il était difficile d'établir le même ordre : les coefficients choisis appartiennent à la même plage de module, ainsi une petite perturbation est capable de changer leur ordre relatif. Ce qui provoque une désynchronisation totale dans le processus de détection.

6 Conclusion

Nous avons présenté deux schémas de tatouage informé à base de treillis, appliqués en ACI et en TCD, testés sur les images d'erreur de prédiction de mouvement d'une séquence vidéo. Les schémas se montrent robustes face à la compression de type MPEG-2. La robustesse de la méthode en TCD dépasse celle utilisant l'ACI. Ceci est dû en particulier au choix des coefficients tatoués en TCD : ces coefficients constituent des composantes de fréquence moyenne, non perturbée par la compression.
Les résultats obtenus pour l'ACI sont prometteurs. Nous travaillons actuellement pour choisir des coefficients plus robustes, en fonction de leur fréquence et leur module. Nous étudions également une amélioration du schéma opérant dans le domaine transformé par ACI, qui consiste à ne pas tatouer les blocs homogènes dans une image d'erreur de prédiction de mouvement : ces blocs contiennent généralement peu d'information utile et risquent d'être complètement altérés suite à une compression.

Références

[1] M. Barni, F. Bartolini, R. Caldelli, A. De Rosa, and A. Piva. A robust watermarking approach for raw video. In *Proceedings of 10th International Packet Video Workshop*, Cagliari, Italy, March 2000.

[2] P. Bas, N. V. Boulgouris, F. D. Kovaros, J. M. Chassery, M. G. Strintzis, and B. Macq. Robust watermarking of video object for MPEG-4 applications. In *SPIE Annual Meetting*, San-Diego, USA, 2001.

[3] S. Bounkong, B. Toch, D. Saad, and D. Lowe. ICA for watermarking digital images. *Journal of Machine Learning Research*, 4 :1471–1498, 2003.

[4] J. J. Chae and B. S. Manjunath. Data hiding in video video. In *IEEE International Conference on Image Processing*, Kobe, Japan, 1999.

[5] I. J. Cox, T. Leighton, and T. Shamoon. Secure spread spectrum watermarking for multimedia. *IEEE Transactions on Image Processing*, 6(12) :1673–1687, 1997.

[6] F. Deguillaume, G. Csurka, J. O'Ruanaidh, and T. Pun. Robust 3D DFT video watermarking. In *Proceedings of SPIE 3657, security and watermarking of multimedia content*, pages 113–124, 1999.

[7] F. J. González-Serrano, H. Y. Molina-Bulla, and J. J. Murillo-Fuentes. Independent component analysis applied to digital watermarking. In *IEEE International Conference on Acoustic, Speech and Signal Processing*, Utah, USA, May 2001.

[8] G. Le Guelvouit and S. Pateux. Wide spread spectrum watermarking with side information and interference cancellation. In *Proceedings of SPIE : Security and Watermarking of Multimedia Contents*, Santa Clara, USA, January 2003.

[9] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley-Interscience, 2001.

[10] C. S. Lu, J. R. Chen, H. Y. M. Liao, and K. C. Fan. Real-time MPEG-2 video watermarking in the VLC domain. In *International Conference on Pattern Recognition*, Quebec, Canada, August 2002.

[11] W. Lu, X. Sun, and K. Okada. A digital watermarking technique based on ICA image features. In *Fifth International Conference on ICA*, Granada, Spain, September 2004.

[12] M. L. Miller, G. J. Döer, and I. J. Cox. Dirty-paper trellis codes for watermarking. In *IEEE International Conference on Image Processing*, New York, USA, September 2002.

[13] MPEG Software Simulation Group (MSSG). *MPEG-2 video codec*. World Wide Web, `http://www.mpeg.org/MPEG/MMSG`.

[14] M. Swanson, B. Zhu, and A. Tewfik. Multiresolution scene-based video watermarking using perceptual models. *IEEE Journal on Selected Areas in Communications*, 16(4) :540–550, 1998.

Mécanisme de synchronisation en tatouage audio pour des perturbations désynchronisantes à forte dérive

Cléo BARAS, Nicolas MOREAU, Bassem ZAYEN

GET - Télécom Paris, Département TSI
46 rue Barrault, 75634 Parix Cédex 13, France
{baras, moreau, zayen}@tsi.enst.fr

Résumé – Pour satisfaire les enjeux de robustesse des systèmes de tatouage, nous proposons dans cet article un mécanisme de synchronisation quasi temps-réel. Ce mécanisme tolère toutes opérations désynchronisantes, à condition qu'elles soient modélisables par un rééchantillonnage du signal audio tatoué. Ces opérations induisent une modification de l'échelle des temps entre l'émission et la réception et peuvent être caractérisées par un facteur de dérive, écart relatif entre les fréquences d'échantillonnage. Des résultats expérimentaux sur des signaux réels montrent que le mécanisme de synchronisation proposé maintient la fiabilité de transmission obtenu à un débit donné à une valeur quasi-constante indépendante de la dérive. Une transmission de fiabilité n'excédant pas 10^{-3} peut donc être obtenue pour des débits inférieurs à 50 bits/s y compris en présence de dérives allant jusqu'à 2%.

Abstract – In this paper, we propose a real-time synchronization mechanism for audio watermarking systems. This mechanism pursues the goal of designing a watermarking system which is robust to all desynchronizing operations, provided that these operations can be modeled as a resampling of the watermarked audio signal. Time is rescaled between the embedder and the receiver with respect to a scaling drift, defined as the relative distance between the sampling rates. Experimental results on real audio signals show that the transmission reliability obtained for a given transmission rate remains steady and does not depend on the scaling factor. Consequently, a real-time transmission with 10^{-3} reliability is feasible for transmission rate lower than 50 bps, even when the scaling drift reaches 2%.

1 Introduction

Si les techniques de tatouage sont apparues pour répondre au problème de protection des droits d'auteur, elles sont aujourd'hui aussi utilisées pour étendre le contenu du signal audio, en y insérant une information supplémentaire à destination d'une application cible. Cette application cible est par exemple l'animation d'un clone virtuel pour le projet RNRT[1] ARTUS[2] [1]. Dans ce contexte applicatif particulier, les performances d'un système de tatouage audio dépendent de quatre critères : la transparence du tatouage, le couple débit / fiabilité de transmission, la robustesse aux perturbations licites subies les signaux audio et le coût en temps de calcul des traitements. Parmi l'ensemble des perturbations envisageables, les plus dégradantes sont sans doute les perturbations dites désynchronisantes, auxquelles nous nous intéressons dans ce papier.

Plusieurs types de désynchronisation peuvent être répertoriés, allant de l'introduction d'un simple retard lors d'un filtrage à la dilation ou la contraction de la durée du signal audio lors de Conversions Analogique-Numérique (CAN) ou d'opération de "time stretching" [2]. Ces perturbations peuvent être modélisées par un rééchantillonnage du signal audio tatoué. Ce rééchantillonnage introduit une dérive entre l'échelle des temps à l'émission et à la réception, modifiant de fait de la localisation des bits transmis et leur durée. Cette dérive, définie comme l'écart relatif entre les fréquences d'échantillonnage, peut prendre des valeurs faibles - typiquement de l'ordre de $0,1\%$ si l'on veut utiliser le canal audio pour transmettre de l'information entre deux PCs par exemple - ou nettement plus élevée - de l'ordre de quelques pour cent dans des applications de radiodiffusion. En effet, une émission radio peut être entrecoupée de séquences sonores, préalablement tatouées et devant être diffusées dans un temps imparti. Le "time stretching" permet alors de modifier pour l'occasion la durée de la séquence de sorte à satisfaire la contrainte de temps de diffusion.

L'état de l'art en matière de synchronisation des systèmes de tatouage propose diverses méthodes que l'on peut regrouper en deux classes. La première exploite des formes d'onde étalées spectralement pour estimer la désynchronisation introduite (différence d'échantillonnage entre l'émission et la réception [3] en audio ou distorsion géométrique en image). Souvent adaptées aux opérations de désynchronisation illicites, elle nécessite un temps de calcul relativement important, qui ne permet pas d'envisager un traitement temps-réel du récepteur. La seconde classe consiste à extraire du signal des points d'intérêt [4, 5], qui définissent une cartographie du signal. L'extraction de cette "carte" à la réception permet là encore d'estimer et d'inverser l'opération désynchronisante pour procéder ensuite à la détection de l'information cachée. Malheureusement, ces points d'intérêt étant souvent choisis en fonction des caractéristiques du signal, les systèmes de tatouage qui en découlent induisent un débit de transmission variable.

Dans cet article, nous proposons un mécanisme de synchronisation quasi-temps réel et induisant un débit de transmission fixe. Ce mécanisme est dédié aux systèmes de tatouage dans le domaine temporel conçu dans un contexte de transmission d'information. Il vise à garantir la robustesse de la transmission en présence de perturbations désynchronisantes dont la dérive peut prendre des valeurs importantes mais ne présente pas de variations. Nous appliquerons plus particulièrement ce

[1] Réseau National de Recherche en Télécommunication
[2] Animation Réaliste par Tatouage à Usage des Sourds

mécanisme à une stratégie de tatouage de type additif par étalement de spectre décrite dans [6], dont nous rappellerons les principes dans la première partie. Le mécanisme de synchronisation fera l'objet de la seconde partie. Nous présenterons pour finir des résultats expérimentaux sur des signaux réels qui permettront de conclure quant aux performances du mécanisme de synchronisation.

2 Système de tatouage

Le système de tatouage considéré est le système de tatouage de type additif par étalement de spectre proposé par Larbi dans [6], dont le schéma de principe est rappelé figure 1.

L'information à émettre est supposée être une séquence de L symboles $\{s_l\}_{l=0..L-1}$ à valeur dans $\{0..M-1\}$. Son insertion requière un dictionnaire d'émission $\mathcal{D} = \{\mathbf{d}_m\}_{m=0..M-1}$, contenant M vecteurs blancs orthogonaux de durée finie N_s, de puissance unité et limités à la bande de fréquence $[0; F_c]$. Sur chaque temps symbole $[lN_s; (l+1)N_s - 1]$, où doit être inséré l'information s_l, le signal modulé $\mathbf{v}$ est choisi égal au vecteur $\mathbf{d}_{s_l}$. Il est ensuite mis en forme spectralement par le filtre $H(f)$ pour obtenir le signal de tatouage $\mathbf{t}$. Ce filtre est conçu de sorte que la Densité Spectrale de Puissance (DSP) de $\mathbf{t}$ coïncide avec un seuil de masquage, limite fréquentielle caractérisant la contrainte d'inaudibilité du signal $\mathbf{t}$ en présence de $\mathbf{x}$. Ce seuil est calculé en appliquant un modèle psychoacoustique (dérivé du modèle MPEG n°1) au signal audio. Le signal audio tatoué $\mathbf{y}$ est finalement obtenu en ajoutant le signal audio $\mathbf{x}$ et le signal de tatouage $\mathbf{t}$.

La réception est basée sur une procédure d'égalisation qui consiste à inverser la mise en forme spectrale de l'émission puis à réestimer le signal modulé à l'aide d'un filtre de Wiener. Le signal audio tatoué reçu $\hat{\mathbf{y}}$ (après d'éventuelles perturbations dans le canal) est donc filtré par $\frac{1}{\hat{H}(f)}$, où $\hat{H}(f)$ est le filtre de mise en forme calculé à partir de $\hat{\mathbf{y}}$. Un filtre de Wiener est ensuite construit de sorte à minimiser l'erreur quadratique entre le signal estimé $\hat{\mathbf{v}}$ et le signal modulé $\mathbf{v}$ relativement aux coefficients du filtre. Ces coefficients $\mathbf{w}$ sont donnés par : $\mathbf{w} = R_{\hat{\mathbf{z}}}^{-1} r_{\mathbf{v}}$, où $R_{\hat{\mathbf{z}}}$ et $r_{\mathbf{v}}$ sont respectivement la matrice et la fonction d'autocovariance du signal $\hat{\mathbf{z}}$ et du signal $\mathbf{v}$ à estimer.

La décision quant à la séquence binaire reçue est effectuée à l'aide d'un démodulateur par corrélation qui sélectionne le vecteur d'un dictionnaire de réception $\hat{\mathcal{D}} = \{\hat{\mathbf{d}}_m\}_{m=0..M-1}$ dont la corrélation avec le signal estimé $\hat{\mathbf{v}}$ est maximum. Ce dictionnaire doit refléter l'ensemble des distorsions introduites par le canal sur les vecteurs du dictionnaire. Le signal n'ayant pour l'instant subi aucune désynchronisation, $\hat{\mathcal{D}}$ est donc identique au dictionnaire d'émission $\mathcal{D}$.

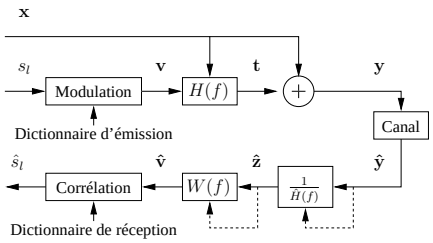


FIG. 1 – Schema de principe du système de tatouage

3 Mécanisme de synchronisation

3.1 Modélisation de la désynchronisation

Les opérations désynchonisantes du canal, telles que les opérations de contraction-dilatation, peuvent être modélisées par l'ajout d'un retard et le réechantillonnage du signal audio tatoué $\mathbf{y}$ (initialement échantillonné à la fréquence F_e) à la fréquence $\hat{F}_e$. En exploitant la relation de conversion d'un signal numérique en signal analogique, le signal audio tatoué réechantillonné $\hat{\mathbf{y}}$ peut être exprimé sous la forme :

$$\hat{y}(\hat{n} + \hat{n}_0) = \sum_{q=-\infty}^{\infty} y(n(\hat{n}) + q) sin_c(q + \delta(\hat{n})),$$

où $\hat{n}_0$ est le retard à l'origine, $n(\hat{n}) = \left[\frac{F_e}{\hat{F}_e}\hat{n}\right]$ ($[x]$ dénotant l'arrondi de x à l'entier le plus proche) modélise la transformation subie par l'échelle des temps entre l'émetteur et le récepteur et $\delta(\hat{n}) = n(\hat{n}) - \left[\frac{F_e}{\hat{F}_e}\hat{n}\right]$ est l'erreur liée à l'arrondi.

Ces opérations modifient donc l'amplitude des échantillons du signal audio tatoué mais surtout la location de chaque symbole et la durée pendant laquelle ils sont insérés. Le mécanisme de synchronisation, pour être efficace, doit donc restituer aussi bien le rythme que les formes prises par l'information tatouée, traduites par un dictionnaire de réception reflétant les perturbations introduites par la désynchronisation. Ces étapes requièrent finalement l'estimation de la fréquence $\hat{F}_e$.

3.2 Principe du mécanisme

Le mécanisme de synchronisation que nous proposons, représenté figure 2, peut être schématisé sous la forme de 3 étapes :

- une *phase d'initialisation* servant à l'estimation de la fréquence $\hat{F}_e$ et du retard $\hat{n}_0$. De cette phase est déduite la dérive $d = \frac{\hat{F}_e - F_e}{F_e}$.
- une *phase d'acquisition*, permettant la synchronisation du récepteur sur des points de référence, cadençant le message binaire émis et l'actualisation de la valeur de $\hat{F}_e$.
- une *phase de poursuite*, réalisant la synchronisation fine de la localisation des symboles et leurs détections avec un dictionnaire de réception adapté à la désynchronisation.

3.2.1 Principe général

Cette procédure de synchronisation nécessite de modifier la structure du signal de tatouage en ajoutant différentes patterns de synchronisation comme schématisé à la figure 3. Une pattern de synchronisation $\mathbf{p}_i$ de durée N_i est insérée M_i fois en amont du signal de tatouage à intervalle de $N_i + N_z$ échantillons. Le signal de tatouage est ensuite découpé sous la forme de *messages* de longueur N_m, contenant une pattern de synchronisation $\mathbf{p}_a$ de durée N_a servant de points de référence et une partie de la séquence binaire à tatouer, de L_m symboles, appelée *sous-séquence*. Les patterns sont choisies étalées spectralement dans la bande de fréquence $[0, F_c]$ et de puissance unité.

La stratégie de tatouage utilisée pour leurs insertions dans le signal audio est identique à celle du système de tatouage décrit section 2 : ces patterns sont filtrées par le filtre de mise en forme psychoacoustique $H(f)$. Elles seront estimées à la

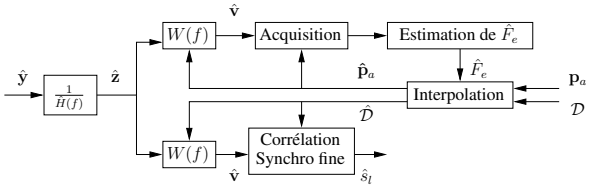


FIG. 2 – Mécanisme de synchronisation pour les phases d'acquisition et de poursuite

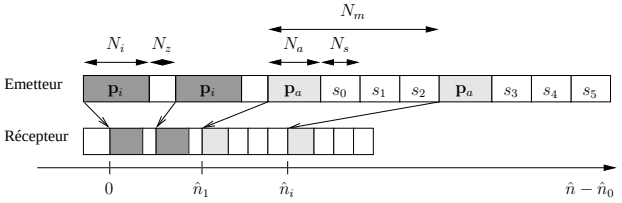


FIG. 3 – Structure du signal modulé (localisant les patterns de synchronisation) à l'émission et à la réception

réception pour les besoins du mécanisme de synchronisation en utilisant la procédure d'égalisation préalablement décrite : le filtre d'inversion du canal $1/\hat{H}(f)$ (dont résulte le signal $\hat{\mathbf{z}}$) puis un filtre de Wiener $W(f)$. Le filtre $\hat{H}(f)$, obtenu suivant un modèle psychoacoustique appliqué au signal audio tatoué $\hat{\mathbf{y}}$, est peu sensible à la désynchronisation (puisqu'il est homogène à une DSP). Par contre, le filtre de Wiener $W(f)$ dépend de la pattern à estimer et donc de la modification qu'elle a pu subir pendant l'opération désynchronisante : il doit donc être calculé spécifiquement en fonction de l'interpolation à la fréquence $\hat{F}_e$ de la pattern à estimer. Plus généralement, on dira par la suite que $W(f)$ est calculé spécifiquement pour estimer le signal $\mathbf{v}$ si ces coefficients choisis de sorte à minimiser l'erreur quadratique entre le signal résultant du filtrage $\hat{\mathbf{v}}$ et le signal à estimer $\mathbf{v}$. Ces coefficients seront : $\mathbf{w} = R_{\hat{\mathbf{z}}}^{-1} r_{\mathbf{v}}$, où $r_{\mathbf{v}}$ est la fonction d'autocovariance de $\mathbf{v}$.

3.2.2 Phase d'initialisation

La phase d'initialisation consiste d'abord à estimer la fréquence de réception $\hat{F}_e$ à l'aide des M_i patterns de synchronisation insérées en amont du signal de tatouage. Pour des raisons d'implémentation, ces patterns sont pour l'instant recherchées sur une durée limitée N_r au début du signal audio.

$\hat{F}_e$ n'ayant pas encore été estimée, $W(f)$ est spécifiquement calculé pour la pattern $\mathbf{p}_i$. P_i versions interpolées de $\mathbf{p}_i$, notées $\hat{\mathbf{p}}_i$, sont évaluées à des fréquences $\hat{F}_i$, couvrant l'intervalle des fréquences de réception tolérées par le système de synchronisation. Pour chaque version interpolée $\hat{\mathbf{p}}_i$, la fonction d'intercovariance entre le signal $\hat{\mathbf{v}}$ (issu du filtrage de Wiener) et $\hat{\mathbf{p}}_i$ est calculée. De l'ensemble de ces P_i fonctions d'intercovariance sont ensuite seulement extraites les M_i valeurs maximales des corrélations. La durée entre ces M_i pics comparée à la durée imposée à l'émission entre les patterns indique la première estimation de $\hat{F}_e$. Cette estimation est ensuite affinée en traitant le signal audio une seconde fois : $W(f)$ est maintenant calculé spécifiquement pour la version interpolée de $\mathbf{p}_i$ à la fréquence $\hat{F}_e$. La fonction d'intercorrélation entre la pattern interpolée et le signal $\hat{\mathbf{v}}$ résultant du filtrage est calculée, mettant en évidence M_i pics de corrélation. Un calcul de la durée entre les pics affine l'estimation de $\hat{F}_e$.

La phase d'initialisation se termine par l'estimation du retard à l'origine $\hat{n}_0$. Ce retard est évalué en détectant la pattern de synchronisation $\mathbf{p}_a$ marquant le début du premier message. Connaissant l'estimation de $\hat{F}_e$, une version interpolée de $\mathbf{p}_a$, notée $\hat{\mathbf{p}}_a$, est calculée et utilisée pour choisir les coefficients spécifiques de $W(f)$. Une méthode de corrélation par fenêtre glissante entre le signal $\hat{\mathbf{v}}$ résultant du filtrage est utilisée pour localiser l'instant de début de la pattern, noté $\hat{n}_1$, sur lequel se synchronise le récepteur.

3.2.3 Phase d'acquisition

La phase d'acquisition consiste à détecter les patterns de synchronisation $\mathbf{p}_a$ insérées régulièrement dans le signal de tatouage pour synchroniser le récepteur. Ces patterns servent en effet de points de référence marquant le début de chaque message. La pattern du premier message ayant été détectée durant la phase d'initialisation, cette étape n'intervient donc qu'à partir de la détection du deuxième message.

Connaissant la position de la pattern du i-ème message, notée $\hat{n}_i$, et une estimation de la fréquence $\hat{F}_e$, la pattern de synchronisation précédant le $i+1$-ème message est recherchée autour de sa position théorique : $\hat{n}_{i+1} = \hat{n}_i \left\lfloor \frac{\hat{F}_e}{F_e} N_m \right\rceil$, évitant ainsi une recherche exhaustive sur toute la longueur du message. Le filtre de Wiener est recalculé pour s'adapter à la version interpolée de $\mathbf{p}_a$ par $\hat{F}_e$. La détection du pic de la fonction d'intercorrélation entre le signal filtré résultant $\hat{\mathbf{v}}$ et la pattern interpolée de $\mathbf{p}_a$ par $\hat{F}_e$ indique l'instant effectif de début du $i+1$-ème message.

La valeur estimée $\hat{F}_e$ de la fréquence d'échantillonnage de réception est ensuite actualisée en évaluant la durée entre la pattern de synchronisation détectée et celle du premier message : $\hat{F}_e = \frac{\hat{n}_{i+1} - \hat{n}_1}{(i+1)N_m} F_e$.

3.2.4 Phase de poursuite

La pattern de synchronisation en amont du i-ème message ayant été localisée à l'instant $\hat{n}_i$, il est alors possible de procéder à la détection de la sous-séquence binaire qui lui fait suite. Connaissant l'estimation de $\hat{F}_e$, nous pouvons déduire les paramètres du processus de détection :

- le dictionnaire de réception adapté $\hat{\mathcal{D}}$ reflétant les distorsions introduites par la désynchronisation : ce dictionnaire contient l'ensemble des vecteurs du dictionnaire d'émission $\mathcal{D}$ interpolés à la fréquence estimée $\hat{F}_e$,
- le filtre de Wiener spécifiquement adapté à l'estimation du signal modulé ayant subi l'interpolation par $\hat{F}_e$,
- la durée d'insertion des symboles à la réception $\hat{N}_s = \frac{\hat{F}_e}{F_e} N_s$ ainsi que leur localisation théorique : le l-ième symbole du i-ème message débute à l'échantillon $\hat{n}_l = \hat{n}_i + \left\lfloor \frac{\hat{F}_e}{F_e}(N_a + lN_s) \right\rceil$.

Une erreur sur l'estimation de $\hat{F}_e$ ou une imprécision liée au calcul de l'arrondi pouvant néanmoins modifier légèrement la localisation du début de chaque symbole, on procède donc à une synchronisation fine de type "early-late gate" [7] pour préciser la localisation de chaque symbole reçu et décider conjointement de sa valeur. Cette synchronisation fine consiste à évaluer la métrique de décision sur un voisinage de la localisation théorique du symbole de N_e échantillons. La valeur maximale

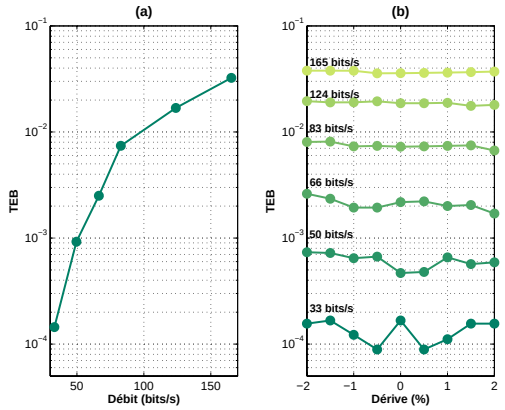


FIG. 4 – TEB obtenu pour différents débits par le système de tatouage avec le mécanisme de synchronisation proposé pour différentes configurations de canal : (a) sans perturbation, (b) opérations d'interpolation avec différentes valeurs de dérive.

de la corrélation indique à la fois la localisation précise du début du symbole reçu et sa valeur.

4 Evaluation des performances

Les performances du mécanisme de synchronisation sont jugées en évaluant la fiabilité de transmission et le coût en temps de calcul du système de tatouage, implémenté en C, face à des opérations de compression-dilatation pour différentes valeurs de la dérive. L'évaluation de la fiabilité consiste à mesurer le taux d'erreur binaire (TEB) obtenu par tatouage d'un ensemble de signaux tests par une séquence binaire de L =100000 bits. Cet échantillon test est constitué de 20 signaux audio de style divers, échantillonnés à $F_e = 44.1$ kHz. La mesure du TEB fournit donc une estimation de la probabilité d'erreur de transmission à une précision de 5.10^{-4} pour un taux de confiance de 70%. Le coût en temps de calcul est quant à lui évalué comme le rapport entre le temps nécessaire à la détection du tatouage et la durée du signal traité (en secondes).

Le dictionnaire utilisé pour le système de tatouage contient $M = 4$ vecteurs, étalés jusqu'à $F_c = 6$ kHz. Il permet d'obtenir un tatouage légèrement audible mais n'induisant pas de gêne auditive lors de l'écoute du signal tatoué (note entre 0 et -1 sur l'échelle perceptuelle définie par la recommandation UIT-R BS 562). Le débit de transmission utile est : $R = \frac{\log_2(M) L_m N_s}{N_m} F_e$.

Les paramètres du mécanisme de synchronisation sont les suivants : les durées de patterns sont $N_i = 12N_{hw}$, $N_z = 2N_{hw}$, $N_a = 15N_{hw}$, où $N_{hw} = 512$ est la taille des fenêtres utilisées pour le calcul des filtres $H(f)$ et $W(f)$, leur taux de répétition est $M_i = 3$, la durée des messages est $N_m = F_e$, la durée de recherche pour la phase d'initialisation est $N_r = F_e$, celle de la localisation de chaque symbole est $N_e = 3$ et les paramètres pour l'interpolation de la pattern de synchronisation pour la phase d'initialisation sont : $P_i = 21$ et $\hat{F}_i = F_e \pm 100i$.

Un rappel des performances du système en terme de débit de transmission et de TEB dans le cas d'un canal sans perturbation est proposé figure 4(a). La figure 4(b) permet d'évaluer les performances du système pour différents débits face à une opération de désynchronisation (de type interpolation) en présentant les TEB en fonction de la dérive. Le mécanisme de synchronisation fait ainsi état de son efficacité puisque les TEB obtenus varient peu en fonction de la dérive et sont d'un ordre de grandeur similaire à ceux obtenus pour une transmission sans désynchronisation.

Le coût lié au temps de calcul est évalué à 0.5 fois le temps réel sans prendre en compte la phase d'initialisation. Cette phase, la plus coûteuse, n'est par contre effectuée qu'une unique fois en début de transmission : elle ne participe pas au "régime courant" de la détection, constitué des phases d'acquisition et de poursuite. Par la suite, elle devra être adaptée au besoin d'une réelle application temps-réel (puisque la recherche du début du tatouage ne doit pas être limitée à quelques secondes).

5 Conclusions et perspectives

Dans cet article, nous avons présenté un mécanisme de synchronisation quasi-temps réel adapté aux systèmes de tatouage dans le domaine temporel dédiés à la transmission d'information. Ce mécanisme est conçu pour garantir la robustesse de la transmission à toutes opérations de désynchronisation, à condition qu'elles puissent être modélisées par un réechantillonnage du signal audio tatoué. Il est basé sur l'ajout de patterns de synchronisation, permettant d'estimer la dérive d'échantillonnage, même forte, pour adapter les paramètres du récepteur. Des résultats expérimentaux sur des signaux réels ont montré que la fiabilité de transmission obtenue pour un canal sans perturbation est correctement maintenue si le canal devient une désynchronisation à forte dérive, attestant ainsi de la performance du mécanisme proposé. Ce mécanisme étant susceptible d'être utilisé avec tous systèmes de tatouage réalisant une insertion dans le domaine temporel, des études sont actuellement menées pour évaluer ces performances pour différents types de récepteur (par filtre blanchissant par exemple) et pour différentes stratégies d'insertion (de type STDM). Les premiers résultats obtenus pour le filtre blanchissant sont très encourageants ; ils sont par contre très mitigés pour la STDM.

Références

[1] RNRT, *ARTUS*, http ://www.telecom.gouv.fr/rnrt/rnrt/projets/res_01_37.htm.

[2] U. Zölzer, *DAFX - Digital audio effects*, Wiley, 2002.

[3] D. Kirovski et H. Malvar, "Spread-spectrum watermarking of audio signals," *IEEE Transactions on Signal Processing*, vol. 51, pp. 1020 –1033, Avril 2003.

[4] A. LoboGuerrero, F. Marques, J. Lienard, et P. Bas, "Enhanced audio data hiding synchronization using non linear filters," *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 885–888, mai 2004.

[5] P. Bas, *Méthodes de tatouage d'images fondées sur le contenu*, Mémoire de Thèse, 2000.

[6] S. Larbi, M. Jaidane, et N. Moreau, "A new wiener filtering based detection scheme for time domain perceptual audio watermarking," *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 949–952, mai 2004.

[7] J. Proakis, *Digital communications*, McGraw-Hill, 2001, 4ème édition.

Etude et extension des *flots de Ricci, Kähler-Ricci et Calabi* dans le cadre du traitement de l'image et de la géométrie de l'information

F. BARBARESCO[1]
[1]THALES AIR DEFENCE, JOINT RADAR SENSORS UNIT, Project & Engineering Department,
7/9, rue des Mathurins, F-92223 Bagneux, Cedex, FRANCE
frederic.barbaresco@fr.thalesgroup.com

Résumé – La théorie des flots géométriques [1] est au cœur de plusieurs disciplines fondamentales des mathématiques (géométrie différentielle et algèbrique, analyse complexe & globale, EDP) et de la physique mathématique (calcul des variations, relativité générale, variété d'Einstein, théorie des cordes) , mais elle repose avant tout sur le socle de la géométrie Riemannienne dont M.Berger dresse le panorama dans un ouvrage récent [5], son extension à des variétés complexes, la géométrie d'Erich Kähler [4,6], dont la vitalité est vantée par J.P. Bourguignon dans [7] et la théorie des surfaces minimales dont la richesse est synthétisée par F. Hélein dans la postface à l'ouvrage [9]. Le présent article souhaite initier une réflexion sur l'utilisation des flots géométriques intrinsèques, tel le flot de Ricci en traitement d'image pour élaborer de nouveaux opérateurs non-linéaires de filtrages anisotropes, et par le biais de la géométrisation de la théorie de l'information par Chentsov, introduire les flots de Kähler-Ricci et de Calabi pour approfondir la notion de noyaux de diffusion de l'information [3].

Abstract – Geometric Flow Theory [1] is cross fertilized by diverse elements coming from Pure Mathematic (geometry, algebra, analyse, PDE) and Mathematical Physic (calculus of variations, General Relativity, Einstein Manifold, String Theory), but its foundation is mainly based on Riemannian Geometry, as explained by M. Berger in a recent panoramic view [5] of this discipline, its extension to complex manifolds, the Erich Kähler's Geometry [4,6], vaunted for its unabated vitality by J.P. Bourguignon in [7], and Minimal Surface Theory recently synthetized by F. Hélein in the book introduction [9]. This paper would like to initiate seminal studies for applying intrinsic flows, such as Ricci flow in Image Processing to conceive new non-linear operator of anisotropic filtering, and thank to information theory geometrization by Chentsov, by using Kähler-Ricci and Calabi flows to deepen the concept of information diffusion kernel [3].

1 La théorie des flots géométriques

La théorie des flots géométriques est une discipline récente [1] dont les ramifications sont nombreuses dans de multiples branches des mathématiques et de la physique mathématique. Contrairement aux flots classiques extrinsèques (flot de courbure moyenne, flot de Wilmore, flot de Laplace-Beltrami, flot de Gauss,…) maintenant largement utilisés en traitement d'image [2,10], la théorie des flots intrinsèques est basée sur des flots qui n'agissent pas explicitement sur la variété elle-même mais implicitement la déforme de proche en proche en modifiant la métrique différentielle g_{ij} associée à cette variété. Dans le flot de Ricci, la métrique est modifiée proportionnellement au tenseur de Ricci R_{ij} associé :

$$\frac{\partial g_{ij}}{\partial t} = 2\left[-R_{ij} + \frac{1}{n} r.g_{ij} \right] \text{ avec } r = \left(\int_{M^n} R d\eta \right) / \left(\int_{M^n} d\eta \right)$$

Le problème peut être formalisé sous forme variationnel en minimisant ,comme le fit D.Hilbert, une fonctionnelle S(g) intégrant la courbure scalaire sur l'ensemble de la variété, à volume constant :

$$S(g) = \int_{M^n} R.\sqrt{\det(g)}.d^n x = \int_{M^n} R.d\eta \text{ avec } V_{M^n} = \int_{M^n} d\eta = cste$$

Après un rappel de la théorie du flot de Ricci, nous appliquons cet outil théorique pour définir un nouvel opérateur non-linéaire de filtrage anisotrope d'image, en lieu et place du flot de Laplace-Beltrami, introduit par N. Sochen [2]. Notre approche consite donc à considérer une image *I(x,y)*, comme une variété ouverte orientée de dimension 2 dans un espace $\mathbf{R}^3$: *S[x,y,I(x,y)]*. La métrique de cette variété s'exprime uniquement à partir du gradient de l'image et elle est déformée suivant un flot de Ricci modifié pour lequel la courbure scalaire moyenne, habituellement intégrée sur la variétée complète, est obtenue par une intégration locale de la courbure scalaire au voisinage du point considéré. En dimension 2, toute métrique est d'Einstein $R_{ij} = \frac{R}{2}.g_{ij}$, aussi est-on ramené au flot suivant : $\frac{\partial g_{ij}}{\partial t} = [r - R] g_{ij}$

Ce flot peut s'identifier en dimension 2 à un flot de Gauss, la métrique étant déformée proportionnellement à la différence entre la courbure de Gauss locale et sa moyenne estimée dans son voisinage : $\frac{\partial g_{ij}}{\partial t} = 2[K - K_m] g_{ij}$. En se plaçant dans un système de coordonnées isothermes introduites par G. Darmois et utilisées par A. Lichnerowicz [8], on montre que l'opérateur de filtrage obtenu possède la propriété de conserver les contours (équation de diffusion de la métrique), caractérisés par une courbure de Gauss nulle (c.a.d. une des courbures principales nulle). En revanche, l'avantage par rapport au flot de Laplace-Beltrami, réside dans le fait que les points isolés de gradient fort sont lissés par le flot de Ricci, alors que ceux-ci ne l'étaient pas et bloquaient le flot introduit par N.Sochen dans [2]. L'image est ensuite reconstruite en définissant une fonction de la variable complexe $z = x + iy$ déduite des gradients de l'image :

$$\frac{\partial I}{\partial t} = \mathrm{Re} \int_{z_0}^{z} \frac{\partial \Theta}{\partial t}(z).dz \text{ avec } \frac{\partial \Theta}{\partial t}(z) = \frac{\partial}{\partial t}\left(\frac{\partial I}{\partial x}\right) - i \frac{\partial}{\partial t}\left(\frac{\partial I}{\partial y}\right)$$

Dans le cadre de la géométrie de l'information et des méthodes à noyaux associées («Information Diffusion Kernels » [3]), les lois statistiques classiques (lois gaussiennes, multivarié-gaussiennes, Gamma, elliptiques,…) conduisent à des métriques complexes, pour lesquelles nous utilisons le flot de Khäler-Ricci (extension du flot de Ricci sur des variétés complexes) afin de moyenner un noyau donné sur les densités a posteriori pour obtenir un noyau sur les données $K_t(x,x') = \iint K_t(\theta,\theta').p(x/\theta).p(x/\theta').d\theta.d\theta'$. Nous établissons les comportements asymptotiques de ces lois soumises au flot de Kähler-Ricci. Dans le cas des modèles autorégressifs complexes de moyenne nulle, on retrouve un potentiel de Kähler Φ, tel que $g_{i\bar{j}} = \frac{\partial^2 \Phi}{\partial z^i \partial \bar{z}^j}$,sur le polydisque unité, de la forme $\Phi = \sum_{k=1}^{n-1} \alpha_k . \ln\left[1 - |z_k|^2\right] = \ln K_D(z,z)$, de noyau de Bergman associé $K_D(z,z) = \prod_{k=1}^{n-1} \left(1 - |z_k|^2\right)^{\alpha_k}$, qui fut le cas traité par E. Kähler à l'origine dans son papier précurseur [6] (cas qu'il a appelé ***hyper-Abélien***). En effet, dans le cadre de la géométrie de l'information, le potentiel de Kähler est fourni par l'entropie, qui donne dans le cas autoregressif la même forme que la cas Hyper-Abélien : $\Phi = \sum_{k=1}^{n-1} (n-k).\ln\left[1 - |\mu_k|^2\right] + n.\ln[\pi.e.P_0]$. Nous verrons que la métrique est d'Einstein en un sens plus large « matriciel » :

$$\boxed{R_{ij} = B^{(n)} g_{ij} \text{ avec } R = Tr\left[B^{(n)}\right]}$$

Nous concluons enfin par l'étude du flot plus récent introduit par Calabi [4] et qui conduit à des métriques dites « extrêmes », en déformant directement le potentiel de Kähler proportionnellement à la valeur de la courbure scalaire: $\frac{\partial \Phi}{\partial t} = R_\Phi - \overline{R}$. Ce flot dit de Calabi se déduit également de la minimisation d'une fonctionnelle dépendant

dans ce cas du carré de la courbure scalaire intégré sur la variété : $T(g) = \int_M R^2 d\eta$.

2 Définition du flot de Ricci

Le flot de Ricci trouve ses racines dans les travaux de Hilbert dans le cadre de la théorie de la Relativité Générale, pour laquelle il formalisa le problème en utilisant un principe d'Action minimale en utilisant les outils du calcul des variations. Ainsi l'Equation d'Einstein fut dérivée par Hilbert à partir de la minimisation d'une fonctionnelle, habituellement nommée de nos jours « Action de Hilbert ». Soit S cette action : $\boxed{S(g) = \int_{M^n} R.\sqrt{\det(g)}.d^n x = \int_{M^n} R.d\eta}$ avec l'élément de volume : $d\eta = \sqrt{\det(g)}.d^n x$ et la courbure scalaire R définie à partir du tenseur de Ricci $R_{\mu\nu}$: $R = \sum_\mu \sum_\nu g^{\mu\nu}.R_{\mu\nu}$ et g le tenseur de la métrique telle que $g = [g_{ij}]$, $ds^2 = \sum_{i,j} g_{ij} dx_i . dx_j$ et $g^{-1} = [g^{ij}]$.

Le théorème fondamental y est donc donné par Hilbert comme suit :
Soit M^n une variété de dimension $n \geq 3$ et supposons que $S(g)$ possède une dérivée variationnelle nulle pour toutes les variations conformes de la métrique g, conservant le volume $V(g)$ constant, alors la courbure moyenne $R(g)$ est constante sur M^n. De plus, si $S(g)$ possède une dérivée variationnelle nulle pour toutes les variations de g, conservant $V(g)$ constant, alors $R(g)$ est constant et g est une métrique de Einstein ; la courbure de Ricci R_{ij}, vérifie $\boxed{R_{ij} = \frac{1}{n} R.g_{ij}}$, et est proportionnelle à la métrique g_{ij}.

Pour des variétés de dimension ≥ 3, en prenant la trace de la dérivée covariante de l'expression précédente et en utilisant l'identité de Bianchi, on montre que pour une métrique d'Einstein, la courbure scalaire est constante.

Une métrique d'Einstein peut donc être obtenue en déformant une métrique suivant le flot modifié suivant (appelé flot de Ricci) : $\boxed{\frac{\partial g_{ij}}{\partial t} = 2\left[-R_{ij} + \frac{1}{n} R g_{ij}\right]}$. Pour résoudre des problèmes d'existence en temps fini, R. Hamilton a proposé en 1982, le flot suivant qui préserve les volumes :

$$\boxed{\frac{\partial g_{ij}}{\partial t} = 2\left[-R_{ij} + \frac{1}{n} r.g_{ij}\right] \quad \text{avec} \quad r = \left(\int_{M^n} R d\eta\right) / \left(\int_{M^n} d\eta\right)}$$

On remarque qu'il s'agit du même flot que précédemment où la courbure scalaire R a été remplacée par la courbure scalaire moyenne r, pour des problèmes d'existence et de convergence du flot. Il est aisé de démontrer que le volume est conservé par ce dernier flot.

3 Lissage NL par le flot de Ricci

3.1 Rappel sur le flot de Laplace-Beltrami

N. Sochen [2] a étudié en traitement d'image de nouveaux opérateurs de filtrage non-linéaire, appelés « Geometric Filter », basés sur le flot extrinsèque, dit de courbure moyenne et l'opérateur de diffusion de Laplace-Beltrami. Le flot de Beltrami obtenu s'interprète comme un flot de courbure moyenne $\frac{\partial S}{\partial t} = \kappa.\vec{N}$ qui agit sur la surface $S = (x, y, I)$. L'image en niveau de gris I est considérée comme une variété plongée dans un espace $\Re^3$, et sur laquelle la métrique différentielle est définie par :

$$ds^2 = \sum_{i,j} g_{i,j} dx_i dx_j = dx^2 + dy^2 + dI^2 = dx^2 + dy^2 + (\partial_x I.dx + \partial_y I.dy)^2$$

où la métrique est donc : $g = [g_{ij}]_{i,j=1,2} = \begin{bmatrix} 1+(\partial_x I)^2 & \partial_x I \partial_y I \\ \partial_y I \partial_x I & 1+(\partial_y I)^2 \end{bmatrix}$

et $g^{-1} = [g^{ij}]_{i,j=1,2} = \frac{1}{1+(\partial_x I)^2+(\partial_y I)^2} \begin{bmatrix} 1+(\partial_y I)^2 & -\partial_y I \partial_x I \\ -\partial_x I \partial_y I & 1+(\partial_x I)^2 \end{bmatrix}$

avec les notations suivantes que nous conserverons dans la suite : $\partial_x I = \frac{\partial I}{\partial x}$, $\partial_y I = \frac{\partial I}{\partial y}$. L'opérateur de Laplace-Beltrami $\frac{\partial I}{\partial t} = \Delta_g I$ donne directement l'expression du nouvel opérateur non-linéaire $\frac{\partial I}{\partial t} = \frac{1}{\sqrt{1+|\nabla I|^2}} div\left(\frac{\nabla I}{\sqrt{1+|\nabla I|^2}}\right)$ qui possède la propriété de lisser le bruit (c'est-à-dire $|\nabla I| << 1 \Rightarrow \frac{\partial I}{\partial t} = div(\nabla I) = \Delta I$) et de conserver les contours (c'est-à-dire $|\nabla I| >> 1 \Rightarrow \frac{\partial I}{\partial t} = 0$).

3.2 Filtrage anisotrope d'image & flot de Ricci

3.2.1 Flot de Ricci agissant sur les gradients de l'image

Nous proposons une approche alternative au flot de Laplace-Beltrami, non plus basée sur le flot de courbure moyenne agissant sur la variété, mais sur un flot de Ricci agissant sur la métrique différentielle de la variété. C'est à dire, partant de la même métrique que précédemment , on va étudier le flot de Ricci ou plus particulièrement la version de Hamilton $\frac{\partial g_{ij}}{\partial t} = 2\left[-R_{ij} + \frac{1}{n} r.g_{ij}\right]$ **avec** $r = \left(\int_{Voisinage} R d\eta\right) / \left(\int_{Voisinage} d\eta\right)$ avec $d\eta = \sqrt{\det(g)}.d^n x$, que nous adaptons au traitement de l'image en calculant la courbure scalaire moyenne non plus sur l'ensemble de la variété, mais sur un voisinage local. L'opérateur ainsi obtenu sur la métrique g_{ij} permettra par identification avec la dérivée partielle : $\frac{\partial g}{\partial t} = \frac{\partial(\nabla I.\nabla I^T)}{\partial t}$ d'en déduire les opérateurs agissant sur le gradient de l'image. Pour nos développements, il nous faut alors calculer les expressions suivantes pour la nouvelle métrique définie sur l'image, afin d'aboutir à l'expression du tenseur de Ricci et de la courbure scalaire. Premièrement, on détermine le symbole de Christoffel donné par : $\Gamma^\lambda_{\mu\nu} = \frac{1}{2} \sum_\tau g^{\lambda\tau} . \left[\frac{\partial g_{\nu\varepsilon}}{\partial x^\mu} + \frac{\partial g_{\tau\mu}}{\partial x^\nu} - \frac{\partial g_{\mu\nu}}{\partial x^\tau}\right]$ et le tenseur de Riemann donné par : $R^\lambda_{\mu\nu\tau} = \frac{\partial \Gamma^\lambda_{\mu\tau}}{\partial x^\nu} - \frac{\partial \Gamma^\lambda_{\mu\nu}}{\partial x^\tau} + \sum_\eta \left[\Gamma^\eta_{\mu\tau} . \Gamma^\lambda_{\eta\nu} - \Gamma^\eta_{\mu\nu} . \Gamma^\lambda_{\eta\tau}\right]$.

Le tenseur de Ricci est obtenu à partir du tenseur de Riemann par la relation $R_{\mu\nu} = \sum_\lambda R^\lambda_{\mu\nu\lambda}$ et la courbure scalaire se déduit du tenseur de Riemann par la relation suivante $R = \sum_\mu \sum_\nu g^{\mu\nu} . R_{\mu\nu}$. On déduit pour notre cas d'analyse l'expression du tenseur de Ricci :

$$\boxed{\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \frac{\left((\partial_{xy} I)^2 - \partial_{xx} I \partial_{yy} I\right)}{\left(1+(\partial_x I)^2+(\partial_y I)^2\right)^2} \begin{bmatrix} \left(1+(\partial_x I)^2\right) & \partial_x I \partial_y I \\ \partial_y I \partial_x I & \left(1+(\partial_y I)^2\right) \end{bmatrix}}$$

En remarquant que la courbure scalaire se réduit à : $\boxed{R = \frac{2\left((\partial_{xy} I)^2 - \partial_{xx} I \partial_{yy} I\right)}{\left(1+(\partial_x I)^2+(\partial_y I)^2\right)^2}}$, on vérifie bien que le tenseur d'Einstein pour n=2 , $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} . R$ est nul $G_{\mu\nu} = 0$. On a alors un flot

de Ricci $\boxed{\frac{\partial g_{ij}}{\partial t} = [r - R]g_{ij}}$, dont on déduit sur les gradients le flot :

$$\frac{\partial}{\partial t}\begin{bmatrix}\partial_x I \\ \partial_y I\end{bmatrix} = \left[r - \frac{2\left((\partial_{xy} I)^2 - \partial_{xx} I \partial_{yy} I\right)}{\left(1 + (\partial_x I)^2 + (\partial_y I)^2\right)^2}\right]\begin{bmatrix}\frac{\left(1 + (\partial_x I)^2\right)}{\partial_x I} \\ \frac{\left(1 + (\partial_y I)^2\right)}{\partial_y I}\end{bmatrix}$$

En dimension 2, toutes les métriques sont des métriques d'Einstein, ou de façon équivalente, une métrique d'Einstein n'a pas besoin d'être de courbure scalaire constante, on a alors à remplacer dans l'équation précédente, pour réduire le défaut de courbure la courbure scalaire par la courbure de Gauss K, avec $R = -2K$: $\boxed{\frac{\partial g}{\partial t} = 2[K - K_m]g}$ ou K_m est la courbure de Gauss moyenne. La courbure de Gauss est égale au produit des courbures des directions principales et est donnée par : $K = \frac{R_{1212}}{\det(g)}$ où R_{1212} est une des composantes du tenseur de courbure de Riemann covariant (ou symbole de Riemann de seconde espèce) : $R_{\tau\alpha\lambda\beta} = \sum_\sigma g_{\sigma\tau} R^\sigma_{\alpha\lambda\beta}$.

3.2.2 Propriétés du flot pour la préservation des contours

Considérons le flot de Ricci non normalisé $\frac{\partial g_{ij}}{\partial t} = -2.R_{ij} \quad \forall i, j$.

Afin d'étudier le comportement du flot suivant les valeurs du gradient de l'image, nous introduisons le système de coordonnées isothermes lié à la métrique, introduit par G. Darmois. Dans un tel système de coordonnées locales $\{x^i\}_{i=1,2}$ isothermes, les quantités suivantes sont nulles : $F^k = \Delta_g x^i = -\sum_{\lambda,\mu} g^{\lambda\mu}\Gamma^k_{\lambda\mu}$ pour $k = 1,2$, avec :

$\Delta_g f = \sum_{\lambda,\mu} g^{\lambda\mu}\left(\partial^2_{\lambda\mu} f - \sum_k \Gamma^k_{\lambda\mu}\partial_k f\right)$ l'opérateur de Laplace-Beltrami.

Etant donnée une variété orientée dans l'espace, elle peut toujours être envisagée comme variété coordonnée dans un système de coordonnées isothermes. Or, ces quantités F^k interviennent d'une manière simple dans les expressions du tenseur de Ricci. Un calcul local permet d'établir [8] : $R_{ij} = -G_{ij} - L_{ij}$ donné par les expressions suivantes :

$$G_{ij} = \frac{1}{2}\sum_{\lambda\mu} g_{\lambda\mu}\partial^2_{\lambda\mu} g_{ij} + H_{ij} \quad \text{et} \quad L_{ij} = \frac{1}{2}\left(\sum_\mu g_{i\mu}\partial_j F^\mu + \sum_\mu g_{j\mu}\partial_i F^\mu\right)$$

Les H_{ij} désignent des polynômes par rapport aux $g_{\lambda\mu}$ et $g^{\lambda\mu}$ et à leur dérivées premières. Ainsi en coordonnées isothermes :

$$-2.R_{ij} = \sum_k\sum_l g^{kl}.\frac{\partial^2 g_{ij}}{\partial x^k \partial x^l} + Q_{ij}\left(g^{-1}, \partial g\right)$$

Dans le cas limite $|\nabla I| << 1$ alors $g^{-1} = \left[g^{ij}\right]_{i,j=1,2} \approx \begin{bmatrix}1 & 0 \\ 0 & 1\end{bmatrix}$ et la métrique est Euclidienne : $ds^2 \approx dx^2 + dy^2$, et la métrique est diffusée isotropiquement $\frac{\partial g_{ij}}{\partial t} = \Delta_g g_{ij}$. Le flot de Ricci diffuse alors la métrique $\left[g_{ij}\right]_{i,j=1,2} = Id + \nabla I.\nabla I^T$ sur la variété et donc induit un lissage des gradients : $\frac{\partial\left(\nabla I.\nabla I^T\right)}{\partial t} = \Delta\left(\nabla I.\nabla I^T\right)$. Le flot de Ricci est stoppé $\frac{\partial g_{ij}}{\partial t} = 0$ uniquement aux endroits où $\boxed{\left(\partial^2_{xy} I\right)^2 = \partial^2_{xx} I.\partial^2_{yy} I}$ (équation de Monge-Ampère) pour lesquels le tenseur de Ricci, la courbure scalaire et de Gauss sont nulles. La courbure de Gauss étant égale au produit des courbures principales, cela signifie que l'EDP de diffusion de la métrique est stoppée aux endroits ou l'une des deux courbures principales est nulles (cas des contours de l'image de type : bord d'objet ou coin d'objet). On voit ici une différence avec le Flot de Laplace-Beltrami qui stoppait le flot aux endroits où le gradient était élevé. Dans le cas du flot de Ricci, des points de bruit forts seront lissés alors qu'ils ne le sont pas avec le flot de Laplace-Beltrami.

3.2.3 Reconstruction de l'image à partir des gradients

Considérons la surface $S = (x, y, I)$ et le repère du plan tangent local : $\partial_x S = \begin{bmatrix}1 & 0 & \partial_x I\end{bmatrix}^T$, $\partial_y S = \begin{bmatrix}0 & 1 & \partial_y I\end{bmatrix}^T$, on définie la fonction :

$\Phi(z) = \frac{\partial S}{\partial x} - i\frac{\partial S}{\partial y} = \begin{bmatrix}1 & -i & \left(\partial_x I - i\partial_y I\right)\end{bmatrix}^T$ avec $z = x + iy$

La surface S est alors reconstruite à une translation prêt, à chaque déformation de la métrique par le flot de Ricci en intégrant dS le long d'un chemin de z_0 à z : $S(x, y) = S(x_0, y_0) + \mathrm{Re}\int_{z_0}^{z}\Phi(z).dz$,

en effet $dS = \frac{\partial S}{\partial x}dx + \frac{\partial S}{\partial y}dy = \mathrm{Re}[\Phi(z)dz]$, et l'intensité de l'image se déduit , en ne conservant que la troisième composante :

$I(x, y) = I(x_0, y_0) + \mathrm{Re}\int_{z_0}^{z}\Theta(z).dz$ avec $\Theta(z) = \frac{\partial I}{\partial x} - i\frac{\partial I}{\partial y}$

le vecteur normal étant donné par $\vec{N} = \Phi(z) \wedge \Phi(z)^* / \left[i.\Phi(z).\Phi(z)^*\right]$

On en déduit aussi l'EDP agissant sur I :

$$\frac{\partial I}{\partial t} = \mathrm{Re}\int_{z_0}^{z}\frac{\partial\Theta}{\partial t}(z).dz \quad \text{avec} \quad \frac{\partial\Theta}{\partial t}(z) = \frac{\partial}{\partial t}\left(\frac{\partial I}{\partial x}\right) - i\frac{\partial}{\partial t}\left(\frac{\partial I}{\partial y}\right)$$

4 Flots de Kähler-Ricci et de Calabi en géométrie de l'information

4.1 Notion de noyaux de diffusion de l'information

Dans le cadre de la géométrie de l'information introduite par Rao et axiomatisé par Chentsov, la métrique différentielle est définie dans l'espace des paramètres θ des densités de probabilités $p(x/\theta)$ par : $ds^2 = \sum_{i,j} g_{ij}(\theta).d\theta_i.d\theta_j^* = d\theta^+.I(\theta).d\theta = Var\left[d\ln p(x/\theta)\right]$ avec la métrique $I(\theta) = \left|g_{ij}\right|_{i,j}$ donnée par la matrice de Fisher :

$$\mathbf{g_{ij}}(\theta) = -\int p(x/\theta).\frac{\partial^2 \ln p(x/\theta)}{\partial\theta_i.\partial\theta_j^*}.dx = -E\left[\frac{\partial^2 \ln p(x/\theta)}{\partial\theta_i.\partial\theta_j^*}\right]$$

Cette métrique est également donnée par le Hessien de l'Entropie : $g_{ij}(\theta) = -\frac{\partial^2 H(p)}{\partial\theta_i\partial\theta_j^*}$ avec $H(\theta) = \int p(x/\theta).\ln p(x/\theta).dx$

Lafferty [3] moyenne un noyau donné sur les densités a posteriori $\frac{\partial\theta}{\partial t} = \Delta_g\theta$, pour obtenir un noyau sur les données :

$K_t(x, x') = \iint K_t(\theta, \theta').p(x/\theta).p(x/\theta').d\theta.d\theta'$

4.2 Potentiel de Kähler et flot de Kähler-Ricci

Nous généralisons l'approche [3] en appliquant le flot de Ricci sur une variété « entropie » dont la métrique est donnée par la matrice de Fisher. Dans notre généralisation le flot de Ricci agit sur des variables complexes, évoluant sur des variétés complexes. Pour ce faire, on utilise l'extension du flot de Ricci à des variétés complexes, dites Kählériennes. Soit M^n une variété complexe à n dimensions, compacte ou non, munie d'une métrique Kählérienne, que l'on peut représenter localement par sa forme Riemannienne définie positive : $ds^2 = 2\sum_{i,j=1}^{n} g_{i\bar{j}}.dz^i dz^{\bar{j}}$ avec $\left|g_{i\bar{j}}\right|_{i,j}$ matrice Hermitienne définie

positive. La condition de Kähler s'exprime par l'existence locale d'une fonction potentielle de Kähler Φ, unique à une fonction pluriharmonique près, telle que $\boxed{g_{i\bar{j}} = \frac{\partial^2 \Phi}{\partial z^i \partial \bar{z}^j}}$. On définit de même les expressions suivantes des tenseurs associés :

$\Gamma_{ij}^k = \sum_{l=1}^{n} g^{k\bar{l}} \frac{\partial g_{i\bar{l}}}{\partial z^j}$ et $\Gamma_{\bar{i}\bar{j}}^{\bar{k}} = \sum_{l=1}^{n} g^{\bar{k}l} \frac{\partial g_{l\bar{i}}}{\partial \bar{z}^j}$, $R_{ijk\bar{l}} = -\frac{\partial^2 g_{ij}}{\partial z^k \partial \bar{z}^l} + \sum_{p,q=1}^{n} g^{p\bar{q}} \frac{\partial g_{i\bar{q}}}{\partial z^k} \frac{\partial g_{p\bar{j}}}{\partial \bar{z}^l}$

Dans la géométrie de Kähler , le tenseur de Ricci est donnée par l'expression condensée : $\boxed{R_{i\bar{j}} = -\frac{\partial^2 \log(\det g_{k\bar{l}})}{\partial z_i \partial \bar{z}_j}}$ dont on déduit la courbure scalaire : $R = \sum_{k,l=1}^{n} g^{k\bar{l}} . R_{k\bar{l}}$. Le flot de Kähler-Ricci associé y est donné par : $\boxed{\frac{\partial g_{ij}}{\partial t} = -R_{ij} + \frac{1}{n} R g_{ij}}$

4.3 Potentiel de Kähler associée à l'entropie d'un processus autorégressif complexe

Si l'on considère un vecteur de données complexe X_n de densité de probabilité de Whishart complexe de moyenne nulle :

$P(X_n / R_n) = (\pi)^{-n} . |R_n|^{-1} . e^{-Tr\left[\hat{R}_n . R_n^{-1}\right]}$ avec $E\left[\hat{R}_n\right] = E\left[X_n X_n^+\right] = R_n$.

Sous l'hypothèse d'un signal suivant un modèle autoregressive complexe, l'expression finale de l'Entropie nous est alors donnée en fonction des coefficients de réflexion $\left\{\mu_k \in C / |\mu_k| < 1\right\}$:

$$\boxed{H_n = \ln|R_n| + n.\ln(\pi e) = \sum_{k=1}^{n-1} (k-n).\ln\left[1-|\mu_k|^2\right] + n.\ln\left[\pi . e . P_0\right]}$$

Si nous revenons à la définition de la métrique de l'information, celle-ci, nous est donnée par le Hessien de l'Entropie :

$g_{ij}(\theta) = -\frac{\partial^2 H(p)}{\partial \theta_i \partial \theta_j^*}$ avec $H(p) = \int p(x/\theta).\ln p(x/\theta).dx$

En considérant, le vecteur de paramètre suivant :

$\theta^{(n)} = \left[P_0 \quad \mu_1 \quad \cdots \quad \mu_{n-1}\right]^T = \left[\theta_1^{(n)} \quad \cdots \quad \theta_n^{(n)}\right]^T$

et en développant le calcul du Hessien, on trouve : $g_{11} = n P_0^{-1}$,

$g_{ij} = \frac{(n-i).\delta_{ij}}{\left(1-|\mu_i|^2\right)^2}$ et $\boxed{ds_n^2 = n.\left(\frac{dP_0}{P_0}\right)^2 + \sum_{i=1}^{n-1} (n-i) \frac{|d\mu_i|^2}{\left(1-|\mu_i|^2\right)^2}}$

On en déduit alors une distance au sens de la géométrie de l'information entre deux modèles autoregressifs complexes de moyenne nulle :

$D\left(\theta^{(1)}, \theta^{(2)}\right) = \int_{\theta^{(1)}}^{\theta^{(2)}} \sqrt{n\left(d \ln P_0\right)^2 + \sum_{i=1}^{n-1} \frac{(n-i)}{4} \left(d \ln \frac{1+r_i}{1-r_i}\right)^2}$ et $r_i = |\mu_i|$

On intègre via un automorphisme holomorphique (transf. de Möbius). Si on ne s'intéresse qu'au terme dépendant des μ_i, on est ramené à :

$$\boxed{D\left(\theta^{(1)}, \theta^{(2)}\right) = \sqrt{\sum_{i=1}^{n-1} \frac{(n-i)}{2} \left(\ln\left[\frac{1+\delta_{i(1)}^{(2)}}{1-\delta_{i(1)}^{(2)}}\right]\right)^2}} \text{ avec } \boxed{\delta_{i(1)}^{(2)} = \left|\frac{\mu_i^{(1)} - \mu_i^{(2)}}{1 - \mu_i^{(1)} \mu_i^{(2)*}}\right|}$$

Nous en déduisons la matrice de Ricci : $\left[R_{k\bar{l}}\right]_{k,l} = \left[-\frac{\partial^2 \log\det\left(g_{i\bar{j}}\right)}{\partial z_k \partial \bar{z}_l}\right]_{k,l}$

$R_{k\bar{l}} = -2\delta_{kl}\left(1-|\mu_k|^2\right)^{-2}$ pour $k = 2,...,n-1$ et $R_{11} = P_0^{-2}$

La courbure scalaire étant définie par ailleurs :

$R = \sum_{k,l} g^{k\bar{l}} . R_{k\bar{l}}$ soit $R = -2.\left[n^{-1} + \sum_{j=1}^{n-1} (n-j)^{-1}\right] = -2.\left[\sum_{j=0}^{n-1} (n-j)^{-1}\right]$

La courbure scalaire diverge lorsque n tend vers l'infini. La métrique est d'Einstein en un sens plus large « matriciel » :

$\boxed{R_{ij} = B^{(n)} g_{ij} \text{ avec } R = Tr\left[B^{(n)}\right]}$ avec $B^{(n)} = diag\left\{\cdots, (n-i)^{-1}, \cdots\right\}$

Si nous examinons le flot de Kähler-Ricci dans le cas non normalisé : $\frac{\partial g_{ij}}{\partial t} = -R_{ij}$, celà donne $\frac{\partial \ln\left(1-|\mu_i|^2\right)}{\partial t} = -\frac{1}{(n-i)}$ et $\frac{\partial \ln P_0}{\partial t} = \frac{1}{n}$

$|\mu_i(t)|^2 = 1 - \left(1-|\mu_i(0)|^2\right) e^{-\frac{t}{(n-i)}} \xrightarrow[t\to\infty]{} |\mu_i(t)|^2 = 1$.

Les coefficients de reflexion tendent asymptotiquement vers le cercle unité avec une vitesse dépendante de leur indice.

4.4 Flot de Calabi & géométrie de l'information

La notion de métriques « extrêmes » de Kähler a été introduite par E. calabi dans [4], en introduisant pour montrer leur existence, un nouveau flot, appelé « flot de Calabi ». Considérons la fonctionnelle suivante : $\boxed{\Theta(g) = \int_M R^2 d\eta}$. Les points critiques de $\Theta(g)$ sont appelés les métriques extrêmes. Toute métrique de Kähler-Einstein est également une métrique extrême qui minimise $\Theta(g)$. La solution est donnée par l'EDP parabolique suivante : $\boxed{\frac{\partial \psi}{\partial t} = R_\psi - \bar{R}}$ avec ψ le potentiel de Kähler associé à la métrique de Kähler $g_{i\bar{j}} = \frac{\partial^2 \psi}{\partial z^i \partial \bar{z}^j}$, R_ψ la courbure scalaire associée et $\bar{R}$ sa valeur moyenne sur la variété.

Considérons le chemin $\psi(t)$ $(0 \le t < 1)$ et supposons l'existence des potentiels de distorsion de Kähler $\left\{\phi(s,t) : 0 \le t < 1\right\}$ soumis à :

$\frac{\partial \phi}{\partial t} = R(\phi) - \bar{R}$ et $\phi(0,t) = \psi(t)$. La longueur du chemin $L(s)$ est alors donnée par : $L(s) = \int_0^1 \left(\int_V \left(\frac{\partial \phi(s,t)}{\partial t}\right)^2 . d\eta_{\phi(s,t)} \right)^{\frac{1}{2}} . dt$

Dans le cadre de la géométrie de l'information, ce flot agit sur l'entropie $-H(p)$

$-\sum_{k=1}^{n-1} (k-n).\frac{\partial \ln\left[1-|\mu_k|^2\right]}{\partial t} - n.\frac{\partial \ln\left[\pi . e . P_0\right]}{\partial t} = -2.\left[\sum_{k=1}^{n} \frac{1}{n-k} + \frac{1}{n}\right]$

Ainsi, on obtient : $\frac{\partial \ln\left(1-|\mu_k|^2\right)}{\partial t} = -\frac{2}{(n-k)^2}$ et $\frac{\partial \ln P_0}{\partial t} = \frac{2}{n^2}$

Nous obtenons un comportement comparable au flot de Kähler-Ricci à un facteur prêt. L'entropie croit en proportion de la courbure scalaire.

Références

[1] International Conference "Geometric Flows : Theory & Computation", IPAM, UCLA, USA, February 3-4 , 2004

[2] A. Spira, N. Sochen & R. Kimmel, "Geometric Filter, Diffusion Flows and Kernels in Image Processing", Handbook of Geometric Computing, Springer Verlag, Heidelberg, pp.203-230, February 2005

[3] Lafferty J. & Lebanon G., "Diffusion Kernels on Statistical Manifolds", CMU-CS-04-101, work supported by NSF, January 2004

[4] E. Calabi, « Extremal Kähler metrics », Seminar Ann. of Math. Stud., n°102, Princeton University Press, pp.259-290, 1982

[5] M. Berger, « Panoramic View of Riemannian Geometry », Springer, 2004

[6] E. Kähler, « Über eine bemerkenswerte Hermitesche Metrik », Abh. Math. Sem. Hamburg Univ., n°9, pp.173-186, 1932

[7] J.P. Bourguignon, « The Unabated Vitality of Kählerian Geometry », in *Kähler Erich Mathematical Works*, Edited by R. Berndt, Berlin, Walter de Gruyter, ix, 2003

[8] A. Lichnerowicz, « Sur les équations relativistes de la gravitation », bulletin de la SMF, t.80, pp.237-251, 1952

[9] F. Hélein, Postface à "Analyse et géométrie", Coll. Géométrie au XXème siècle : histoire et horizons, J. Konneiher & al Ed.

[10] F. Barbaresco & al ,"Lissage anisotropique de paramètres statistiques par extension de l'opérateur diffusif de Beltrami dans le cadre de la géométrie de l'information », GRETSI'03, Sept. 2003

Contrainte de forme invariante aux transformations affines pour les contours actifs orientés région

Alban Foulonneau[1], Pierre Charbonnier[1], Fabrice Heitz[2]

[1]Laboratoire des Ponts et Chaussées de Strasbourg
11 rue Jean Mentelin, BP 9, F-67035 STRASBOURG Cedex, France

[2]LSIIT UMR 7005 CNRS Université de Strasbourg I
Boulevard Sébastien Brandt, F-67400 ILLKIRCH, France
alban.foulonneau@equipement.gouv.fr,pierre.charbonnier@equipement.gouv.fr
fabrice.heitz@ensps.u-strasbg.fr

Résumé – Nous présentons une nouvelle façon de contraindre l'évolution d'un contour actif orienté région par rapport à une forme de référence. En minimisant un *a priori* de forme défini comme une distance entre des *descripteurs* constitués des moments de Legendre de la forme, on obtient une équation d'évolution géométrique qui peut être mise à profit dans des applications de segmentation à deux classes. Le modèle proposé permet de gérer des topologies de formes arbitrairement complexes et est intrinsèquement invariant aux transformations affines. Cela résout directement le problème d'alignement et permet la prise en compte de variabilités d'aspect d'origine géométrique.

Abstract – We present a new way of constraining the evolution of a region-based active contour with respect to a reference shape. Minimizing a *shape prior*, defined as a distance between *descriptors* based on the Legendre moments of the shape, leads to a geometric flow that can be used with benefits in a two-class segmentation application. The proposed model handles arbitrary shape topologies and is intrinsically invariant w.r.t. affine transformations, which saves from performing explicit pose estimation and allows accounting for geometrical aspect variabilities.

1 Introduction

L'introduction de contraintes globales de forme dans les modèles de contours actifs orientés région est un thème de recherche d'actualité (voir [1], chap. 2, pour un état de l'art). L'approche usuelle consiste à ajouter à la fonctionnelle optimisée un terme d'*a priori*, visant à contraindre le contour en évolution à ressembler à une forme de référence.

Fig. 1 – Exemples de formes de référence.

Naturellement, deux problèmes se posent : celui de l'alignement des formes et celui de la gestion de variabilités d'aspect. Le premier est généralement traité en prenant en compte les paramètres de pose (translation, rotation, échelle) de manière explicite, ce qui augmente le nombre de degrés de liberté du problème et oblige à une estimation supplémentaire. Les variabilités autorisées sont, la plupart du temps, limitées à celles issues d'un apprentissage statistique. Nous proposons ici une version étendue de la méthode présentée dans [2], où la contrainte que nous définissons est d'ordre géométrique et s'appuie sur les moments de Legendre. Nous mettons à profit l'information géométrique contenue dans les moments pour rendre l'*a priori* invariant aux transformations affines. On obtient ainsi une équation d'évolution unique, intrinsèquement *affine-invariante*, ce qui résout directement le problème d'alignement et permet la prise en compte de variabilités d'aspect significatives, sans apprentissage.

2 Définition d'un *a priori* de forme affine invariant

Notre modèle repose sur une représentation paramétrique des formes à partir des moments de leur fonction caractéristique, jusqu'à un certain ordre, rassemblés dans un vecteur appelé *descripteur*. Cette approche permet de représenter des objets à topologie non triviale de manière hiérarchique [1].

Pour introduire l'invariance aux variations géométriques d'aspect, nous définissons ce que nous appelons une représentation *canonique*. Il s'agit en fait d'un changement de variables tel que deux formes, différant seulement d'un certain type de transformation, soient représentées par le même ensemble de moments. Ainsi, en utilisant cette représentation à la fois pour la forme de référence et pour la forme en évolution, on obtient directement l'invariance à la transformation considérée. Par exemple, dans le cas de l'invariance par translation et changement d'échelle, la représentation canonique d'une forme est obtenue en

alignant son centre de gravité, $(\overline{x},\overline{y})$, sur le centre du domaine et en normalisant son aire, $|\Omega_{in}|$, à une valeur constante, $1/\beta$. Cela revient à utiliser des moments géométriques centrés normés :

$$\eta_{p,q} = \iint_{\Omega_{in}} \frac{(x-\overline{x})^p (y-\overline{y})^q}{(\beta|\Omega_{in}|)^{(p+q+2)/2}} dxdy, \tag{1}$$

$$\text{avec} \quad \overline{x} = \frac{M_{1,0}}{M_{0,0}},\ \overline{y} = \frac{M_{0,1}}{M_{0,0}} \text{ et } |\Omega_{in}| = M_{0,0} \tag{2}$$

comme nous l'avons proposé dans [2] (où les $M_{p,q}$ sont les moments classiques ou *réguliers* : $M_{p,q} = \iint x^p y^q dxdy$).

Dans le cas, plus général, de l'invariance affine, considérons tout d'abord l'algorithme dit de *compactage*. Celui-ci consiste à aligner l'ellipse d'inertie de la forme sur les axes du repère cartésien, puis à appliquer un changement d'échelle non-isotrope de façon à transformer cette ellipse en cercle. On montre facilement [3] que deux formes identiques à une transformation affine près ont la même forme *compacte*, à une rotation près. En compensant cette rotation, on obtient la forme *normalisée* [3]. Ainsi, après leur *normalisation*, toutes les formes identiques à des transformations affines près auront la même représentation *canonique*. La *normalisation* correspond à la composition d'une rotation d'angle θ, de l'application de facteurs d'échelle l_1 en x et l_2 en y et d'une rotation d'angle γ. L'intérêt est que ces paramètres se calculent à l'aide d'expressions analytiques formées à partir des moments géométriques [3, 4] :

$$\theta = \begin{cases} \frac{1}{2}\tan^{-1}\left(\frac{2\nu_{1,1}}{\nu_{2,0}-\nu_{0,2}}\right) + ... \\ +(1-\text{sign}(\nu_{2,0}-\nu_{0,2}))\,.\text{sign}(\nu_{1,1}).\frac{\pi}{4}, \nu_{2,0}-\nu_{0,2} \neq 0 \\ \frac{\pi}{4}, \nu_{2,0}-\nu_{0,2}=0 \quad \nu_{1,1}>0 \\ -\frac{\pi}{4}, \nu_{2,0}-\nu_{0,2}=0 \quad \nu_{1,1}<0 \end{cases} \tag{3}$$

et

$$l_{1/2} = \frac{(\nu_{2,0}+\nu_{0,2}) \pm \sqrt{(\nu_{2,0}-\nu_{0,2})^2 + 4\nu_{1,1}^2}}{2} \tag{4}$$

où les $\nu_{u,v}$ sont les moments centrés d'ordre 2 de la forme originale et

$$\gamma = \tan^{-1}\left(-\frac{M^c_{1,2}+M^c_{3,0}}{M^c_{0,3}+M^c_{2,1}}\right) + (1-\text{sign}(M^c_{0,3}+M^c_{2,1}))\frac{\pi}{2} \tag{5}$$

où les $M^c_{u,v}$ sont les moments centrés de la forme *compacte*. La *normalisation* d'image, cependant, ne gère pas les réflexions. Celles-ci affectant seulement le signe des moments d'ordre impair, nous choisissons en toute généralité d'imposer le signe des moments d'ordre 3. Au final, notre représentation *canonique* est obtenue à partir de moments géométriques que nous qualifierons d'*affine-invariants* :

$$\eta^A_{u,v} = \left(\text{sign}\left(\widehat{\eta}^A_{3,0}\right)\right)^u . \left(\text{sign}\left(\widehat{\eta}^A_{0,3}\right)\right)^v . \widehat{\eta}^A_{u,v}. \tag{6}$$

où l'expression de $\eta^A_{u,v}$ est donnée ci-dessous (7). Notons qu'un modèle simplifié, gérant seulement les paramètres de pose et les réflexions[1] peut s'obtenir à partir de (7) en posant $\gamma = 0$ et $l_1 = l_2 = 1$.

[1] prendre en compte les réflexions demeure nécessaire dans ce cas pour éviter toute ambiguïté dans la détermination de θ.

De façon à bénéficier des bonnes propriétés des bases orthogonales, nous définissons notre descripteur de forme, $\boldsymbol{\lambda^A}$, à partir des moments de Legendre jusqu'à l'ordre N : $\boldsymbol{\lambda^A} = \{\lambda^A_{p,q}, p+q \leq N\}$, plutôt qu'à partir des moments géométriques, sachant qu'il existe une relation linéaire entre les deux types de moments :

$$\lambda^A_{p,q} = C_{pq} \sum_{u=0}^{p}\sum_{v=0}^{q} a_{pu} a_{qv} \eta^A_{u,v}. \tag{8}$$

Finalement, notre *a priori* de forme est défini comme une distance entre la courbe en évolution, Γ, et la référence, en termes de descripteurs invariants :

$$J_{forme}(\Omega_{in}) = d(\boldsymbol{\lambda^A}(\Omega_{in}), \boldsymbol{\lambda}^A_{ref}), \tag{9}$$

où Ω_{in} est la région intérieure de Γ et $\boldsymbol{\lambda}^A_{ref}$ est le descripteur associé à la forme de référence. Notre critère étant basé sur une représentation canonique des formes, il est intrinsèquement invariant aux transformations affines. Aucune estimation supplémentaire de paramètres n'est nécessaire.

3 Dérivation de l'équation d'évolution du contour contraint

L'équation d'évolution correspondant à la minimisation de J_{forme} se calcule par la méthode, désormais classique, de dérivation Eulérienne [5, 6]. L'expression de la vitesse obtenue est donnée dans le tableau Tab. 1 et dans [1], où sont détaillés les calculs. Notons qu'on obtient une équation d'évolution unique, dont toutes les composantes se calculent à partir de moments. Nous implantons celle-ci par la méthode des courbes de niveaux [7], qui prend en compte de manière naturelle les changements de topologie du contour au cours de son évolution.

4 Application à un problème de segmentation

Afin d'illustrer le comportement de l'*a priori* invariant, nous utilisons une fonctionnelle classique pour la segmentation à deux classes [8] :

$$\begin{aligned} J_{donnees}(\Omega_{in},\Omega_{out}) = &\iint_{\Omega_{in}} (I(x,y)-\mu_{in})^2 dxdy \\ &+ \iint_{\Omega_{out}} (I(x,y)-\mu_{out})^2 dxdy, \end{aligned} \tag{10}$$

où μ_{in} (resp. μ_{out}) est l'intensité moyenne du domaine intérieur (resp. extérieur), Ω_{in} (resp. Ω_{out}), et $I(x,y)$ est l'intensité du pixel. Ces paramètres sont recalculés après chaque étape d'évolution, à partir de la partition générée et de l'image I. Notre stratégie algorithmique consiste à optimiser dans un premier temps le terme $J_{donnees}$ seul (ou, éventuellement, avec un terme additionnel de courbure). On introduit ensuite le critère de forme, ce qui conduit à l'équation d'évolution complète :

$$\frac{\partial\Gamma(t)}{\partial t} = (\underbrace{(I-\mu_{in})^2 - (I-\mu_{out})^2}_{V_{donnees}} + \alpha V_{forme})\boldsymbol{\mathcal{N}}, \tag{11}$$

$$\widehat{\eta}^A_{u,v} = \frac{(l_1.l_2)^{\frac{u+v}{4}}}{(\beta|\Omega_{in}|)^{(u+v+2)/2}} \iint_{\Omega_{in}} \left(\frac{((x-\overline{x})\cos\theta + (y-\overline{y})\sin\theta)}{\sqrt{l_1}} \cos\gamma + \frac{((y-\overline{y})\cos\theta - (x-\overline{x})\sin\theta)}{\sqrt{l_2}} \sin\gamma \right)^u . \left(\frac{((y-\overline{y})\cos\theta - (x-\overline{x})\sin\theta)}{\sqrt{l_2}} \cos\gamma - \frac{((x-\overline{x})\cos\theta + (y-\overline{y})\sin\theta)}{\sqrt{l_1}} \sin\gamma \right)^v dxdy. \quad (7)$$

$$\frac{\partial\Gamma}{\partial t} = \underbrace{\sum_{u,v}^{u+v\leq N} A^A_{uv}.\left(\text{sign}\left(\widehat{\eta}^A_{3,0}\right)\right)^u.\left(\text{sign}\left(\widehat{\eta}^A_{0,3}\right)\right)^v.\left(H^A_{uv} + \sum_{i=0}^{9} \widehat{B}^A_{uvi}.L_i\right)}_{V_{forme}} \boldsymbol{\mathcal{N}},$$

$$A^A_{uv} = 2\sum_{p,q}^{p+q\leq N} (\lambda^A_{p,q} - \lambda^{Aref}_{p,q}).C_{pq}a_{pu}a_{qv}, \qquad \text{les expressions des } \widehat{B}^A_{u,v,i} \text{ sont données dans [1]}$$

$$H^A_{uv} = \frac{(l_1.l_2)^{\frac{u+v}{4}}}{(|\beta\Omega_{in}|)^{(u+v+2)/2}} \left(\frac{((x-\overline{x})\cos\theta + (y-\overline{y})\sin\theta}{\sqrt{l_1}} \cos\gamma + \frac{((y-\overline{y})\cos\theta - (x-\overline{x})\sin\theta)}{\sqrt{l_2}} \sin\gamma \right)^u . \left(\frac{((y-\overline{y})\cos\theta - (x-\overline{x})\sin\theta)}{\sqrt{l_2}} \cos\gamma - \frac{((x-\overline{x})\cos\theta + (y-\overline{y})\sin\theta)}{\sqrt{l_1}} \sin\gamma \right)^v$$

$$L_0 = 1,\ L_1 = x,\ L_2 = y,\ L_3 = x^2,\ L_4 = y^2,\ L_5 = xy,\ L_6 = x^3,\ L_7 = y^3,\ L_8 = x^2y,\ L_9 = xy^3.$$

TAB. 1 – Détails de l'équation d'évolution associée à la minimisation de J_{forme} dans le cas de l'invariance affine.

où V_{forme} est définie dans Tab. (1). Le paramètre α est fixé par l'utilisateur. L'ordre maximum des moments utilisé, N, est choisi pour minimiser l'erreur de représentation de la forme de référence (erreur quadratique moyenne entre la forme de référence et sa reconstruction à partir des moments à l'ordre considéré, voir [1] pour plus de détails). Les images de référence employées sont montrées fig. 1. La première expérience, fig. 2, illustre la segmentation d'objets à topologie non triviale, à l'aide d'un modèle simplifié, invariant à la translation et au changement d'échelle. Cette expérience montre également qu'il est possible d'envisager plusieurs stratégies d'initialisation : courbe encadrant grossièrement ou intersectant partiellement l'objet ou initialisation systématique de type « seed snakes » [9]. Il est, d'autre part, souvent possible de s'affranchir complètement du terme classique de courbure. La seconde expérience, fig. 3, illustre la segmentation d'objets ayant subi des déformations affines significatives.

5 Conclusion

En utilisant les moments de Legendre pour définir des descripteurs de forme invariants aux transformations affines, nous avons proposé un *a priori* capable de contraindre l'évolution d'un contour actif vers une forme de référence, augmentant la robustesse de la segmentation en présence de bruit, de fonds texturés et d'occultations. L'approche gère naturellement l'alignement des formes, les changements de topologie et les déformations affines. Une extension à des déformations projectives est envisagée.

Références

[1] A. Foulonneau, "Une contribution à l'introduction de contraintes géométriques dans les contours actifs orientés région," Thèse de Doctorat, ULP (Université Louis Pasteur) - Strasbourg I, France, décembre 2004.

[2] A. Foulonneau, P. Charbonnier, et F. Heitz, "Contraintes géométriques de formes pour les contours actifs orientés région : une approche basée sur les moments de legendre," *Traitement du Signal*, vol. 21, no. 2, pp. 109–128, septembre 2004, (article invité, numéro spécial GRETSI).

[3] S. Pei et C. Lin, "Image normalization for pattern recognition," *Image and Vision Computing*, vol. 13, no. 10, pp. 711–723, décembre 1995.

[4] M. Teague, "Image analysis via the general theory of moments," *Journal of the Optical Society of America*, vol. 70, no. 8, pp. 920–930, août 1980.

[5] M. Delfour et J. Zolesio, *Shape and geometries : analysis, differential calculus and optimization*, ser. Advances in design and control. SIAM, 2001.

[6] S. Jehan-Besson, "Modèles de contours actifs basés région pour la segmentation d'images et de vidéos," Thèse de Doctorat, Université de Nice-Sophia Antipolis, France, janvier 2003.

[7] S. Osher et J. Sethian, "Fronts propagating with curvature-dependant speed : algorithms based on Hamilton-Jacobi formulations," *Journal of Computational Physics*, vol. 79, no. 1, pp. 12–49, novembre 1988.

[8] T. Chan et L. Vese, "Active contours without edges," *IEEE Transactions on Image processing*, vol. 10, no. 2, pp. 266–277, février 2001.

[9] T. McInerney et D. Terzopoulos, "Topologically adaptable snakes," in *Proc. of IEEE International Conference on Computer Vision*, Cambridge, Massachusetts, USA, 20-23 juin, 1995, pp. 840–845.

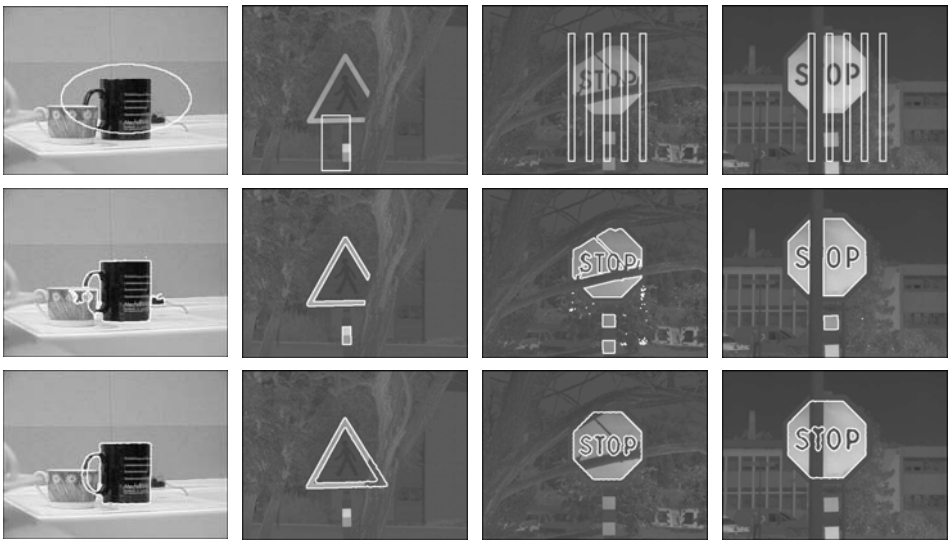


FIG. 2 – Segmentation de formes à topologie complexe (pour les panneaux, on utilise des images de coefficient chromatique rouge). Première ligne : contours initiaux. Seconde ligne : résultats sans contrainte de forme (une contrainte classique de courbure est utilisée seulement dans le cas de la tasse). Troisième ligne : résultats finaux, en ajoutant notre *a priori* jusqu'à l'ordre 40 dans le cas de la tasse et du panneau *stop* et jusqu'à l'ordre 43 pour le panneau triangulaire.

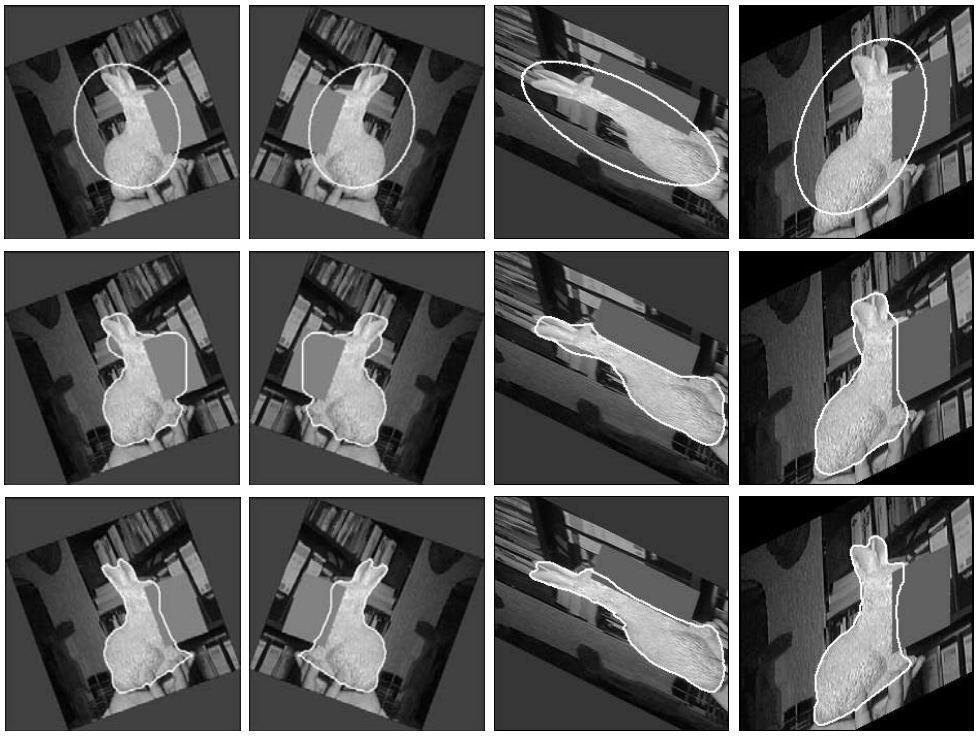


FIG. 3 – Segmentation d'un objet ayant subi des déformations affines. Première ligne : contours initiaux . Seconde ligne : résultats sans contrainte de forme (seule, une contrainte classique de courbure est mise en œuvre). Troisième ligne : résultats finaux, en ajoutant notre *a priori* jusqu'à l'ordre 40. Dans les expériences (a) et (b), l'*a priori* est invariant à la translation, la rotation, le changement d'échelle et les réflexions. Dans les expériences (c) et (d), le modèle affine complet est utilisé.

Formalisme et mesure de la perception : Intérêt de la dynamique sensori-motrice chez un agent robotique

Mickael MAILLARD, Lauence HAFEMEISTER, Philippe GAUSSIER

Laboratoire ETIS UMR CNRS 8051, Groupe Neurocybernétique
E.N.S.E.A, U.C.P., 6 avenue du Ponceau, 95014 Cergy Cedex, France
maillard@ensea.fr, hafemeister@ensea.fr
Gaussier@ensea.fr

Résumé – Notre objectif est de proposer une définition de la perception et de montrer comment un agent autonome peut apprendre à construire sa propre perception de son environnement. Nous présentons cette formalisation dans un cas concret de robotique : un robot Koala muni d'une caméra a pour tâche de se diriger vers un objet précédemment appris. Des outils de visualisation d'une mesure de la perception sont mis en oeuvre. Ceux-ci sont basés sur l'exploitation des sensations et des actions de l'agent enregistrées au cours d'expériences. L'apprentissage sensori-moteur réalisé par l'agent permet à ce dernier de déformer sa perception : du point de vue comportemental, un invariant est créé. Enfin l'observation de l'évolution temporelle de la perception est révélatrice de l'importance des dynamiques des boucles sensori-motrices.

Abstract – We propose a definition of perception and we show how an autonomous agent can learn its own perception of its environment. Next we present this formalism in the case of a robotic application : a Koala robot with a CCD-camera must reach an object learned before. We propose some visualisation tools of a measure of the agent's perception which is based on the record of the couple sensation/action during the experiment. The sensori-motor learning allows to shape the perception of the agent and creates an invariant considering a behavioral point of view. Observation of the temporal evolution of perception emphasises the role of dynamic in the sensori-motor loops.

1 Contexte

En traitement d'images, la perception est souvent associée au seul traitement des données sensorielles. Cependant, de nombreuses expériences en psychologie tendent à montrer que la constitution d'un percept nécessite d'interagir avec l'environnement [8]. Ceci est conforté par des expériences robotiques montrant par exemple qu'un robot peut apprendre à rejoindre un lieu sans être capable de le reconnaître précisément d'un point de vue statique (apprentissage de la constitution d'un bassin d'attraction comportemental autour du lieu but [5]. Avoir un comportement globalement cohérent face à un objet suffit dans bien des cas : on parle alors d'affordances [6]. Dans ce cas, la perception apparaît comme étant la résultante d'associations entre les actions et les sensations. Du point de vue des sytèmes dynamiques [2], il est aussi intéressant de noter que les actions dérivent d'un champ de potentiels [14, 13]. Nous proposons dès lors de modéliser la perception comme un champ de potentiels [4] : les actions d'un agent dérivent de son état de perception. En utilisant le formalisme des systèmes cognitifs développé par Gaussier et al. [3], l'action $|Ac\rangle$ est un vecteur qui se déduit de la perception notée Per par la relation :

$$|Ac\rangle = -m \overrightarrow{grad}\, Per(\vec{p}) \quad (1)$$

La perception dépend de l'état $\overrightarrow{p}$ de l'agent en proie à ses sensations. $\overrightarrow{p}$ est constitué de variables de position spatiale mais aussi de variables déterminant l'état interne de l'agent. m peut être associée à une masse réelle ou virtuelle permettant de prendre en compte l'embodiment de l'agent [11]. Par la suite nous supposerons qu'elle est constante au cours des mesures effectuées.

2 Matériel et méthodes

Afin de comprendre les conséquences de la définition posée, nous avons mis en oeuvre cette approche de la perception dans le cas simple d'un agent robotique. Cet agent est un robot de type Koala muni d'une caméra. Il a pour tâche de se diriger vers un objet précédemment appris. Une reconnaissance explicite du type "nommer" l'objet n'est pas requise ici. On se place au contraire dans un contexte de réalisation d'affordances. Une affordance peut résulter d'une boucle sensori-motrice simple dans laquelle l'agent associe ses sensations à des actions à effectuer. Afin que l'apprentissage de ces quelques associations restent simples, il faut que les caractéristiques visuelles extraites possèdent un certain nombre d'invariants correspondant à la tâche à effectuer. Ainsi la caractérisation par l'agent des vues, image après image est choisie invariante aux changements de contraste et de luminosité. En outre, comme l'agent se déplace, cette caractérisation doit être aussi insensible que possible aux changements d'échelle et aux déformations géométriques de l'image. Finalement une image capturée est caractérisée par des points d'intérêts extraits à l'aide d'un mécanisme multi-échelles [7, 9]. L'algorithme d'extraction des points d'intérêt est inspiré des travaux de Lowe [10]. Les points d'intérêts extraits correspondent aux extrema locaux dans l'espace des échelles des images filtrées par des DOG (Difference Of Gaussians) (fig. 1). Cet algorithme biologiquement plausible présente en outre l'intérêt d'être massivement parallèle. Une parallélisation des calculs sur différentes stations de travail, permet un déroulement de l'algorithme compatible avec du temps réel.

Par la suite, la caractérisation des points d'intérêt que nous avons effectuée est rudimentaire mais son but est d'être suf-

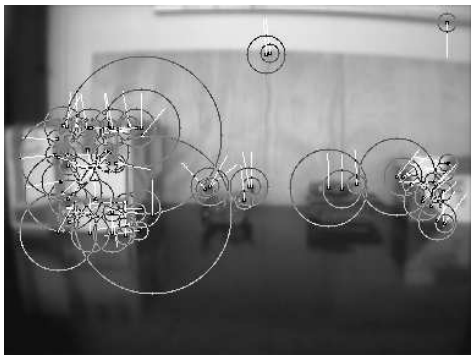


FIG. 1 – Représentation des points caractéristiques extraits en fonction de l'échelle (taille du cercle entourant le point)

fisante pour la tâche considérée. Cette caractérisation consiste à extraire à l'échelle de chaque point d'intérêt, les deux premiers moments des orientations des quatre voisinages du point d'intérêt considéré (figure 2). L'information de l'ensemble des échelles est alors réunie et normalisée sur 8 cartes de neurones de petites tailles (la discrimination est alors moins précise spatialement mais plus robuste à la disparitions de quelques points d'intérêt). Ces huit cartes constituent l'information sensorielle. Afin de construire sa perception de l'environnement, l'agent les associe à des actions. On se place ici dans le cadre d'un apprentissage sensori-moteur de type supervisé : l'agent associe les caractéristiques visuelles multi-échelles d'une image avec la direction permettant de rejoindre l'objet à apprendre. Cette direction est indiquée par l'opérateur à l'aide d'un joystick (fig. 2). Le groupe de neurones noté LMS (Least Mean Square) réalise une minimisation de l'erreur quadratique ([16]) entre sa sortie calculée à partir des cartes de caractéristiques et la commande fournie par le manipulateur. Enfin la prise de décision de l'action à réaliser est effectuée par un groupe de neurones compétitifs noté WTA (Winner Takes All). Il est à noter qu'étant donné les capacités limitées de séparation linéaire d'un tel mécanisme, l'extension à plusieurs objets s'effectue en multipliant les groupes de types LMS : un groupe par objet apprenant suivant le contexte fourni par l'utilisateur. On constate que quelques apprentissages sensori-moteurs permettent alors à l'agent de se diriger vers l'objet appris.

3 Interprétation comportementale et perceptive

Les résultats du système mis en place peuvent être analysés de deux façons différentes. On propose dans un premier temps de visualiser le comportement global du robot par l'intermédiaire de ses trajectoires. Du point de vue d'un observateur, on peut dire que l'agent a accompli sa tâche lorsque son comportement à été globalement cohérent face à l'objet appris, c'est à dire lorsqu'il a réalisé une affordance. La visualisation des trajectoires du robot (fig. 3) montre qu'en présence de l'objet appris, celles-ci convergent nettement vers l'objet. Au contraire si l'objet est absent du champ de vision du robot, les trajectoires sont clairement plus perturbées. Il convient de noter que l'agent n'a pas une reconnaissance explicite et qu'il ne possède pas de système de recherche de l'objet : il se dirige vers ce qui

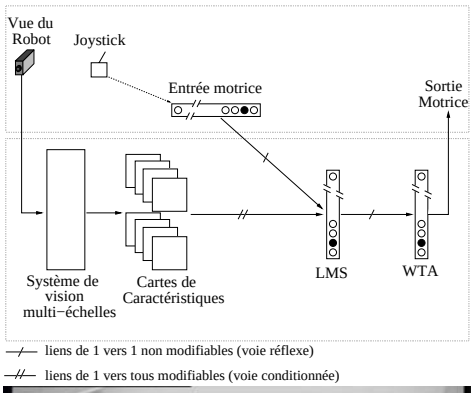


FIG. 2 – *En haut* : Architecture associant les sensations multi-échelles aux actions *En bas* : Exemple d'apprentissage de l'objet "carton" entouré en noir

se trouve dans son champ de vision et qui ressemble le plus à l'objet appris au sens de sa perception. Mais son champ de vision évolue suivant ses actions.

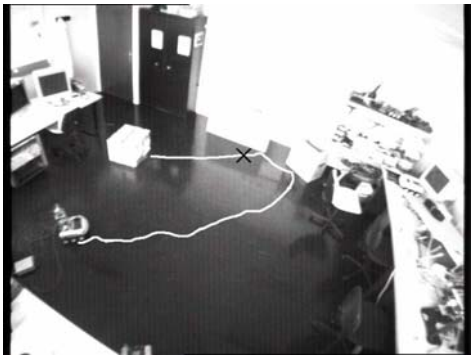


FIG. 3 – Visualisation de la trajectoire du robot. La croix symbolise le lieu où l'objet "carton" rentre dans le champs de vision de l'agent. La trajectoire cesse d'être perturbée à partir de ce moment

Une autre manière d'analyser les résultats consiste à calculer a posteriori la perception de l'agent. Pour un apprentissage donné, l'état du robot se définit par sa position spatiale dans la pièce ainsi que par l'orientation de son corps (et donc de sa caméra) par rapport à l'objet appris. Un moyen de calculer la perception du robot est de lui imposer de parcourir l'espace en mesurant en chaque point l'action qu'il aurait effectuée au vue de ses sensations (fig. 4). D'après l'équation (1), la per-

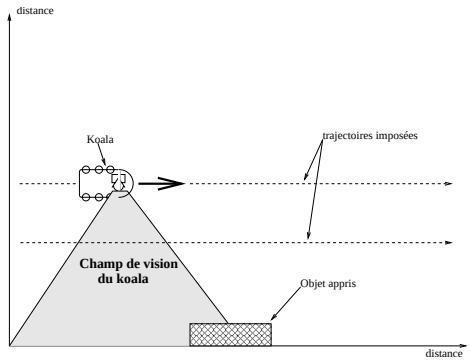


FIG. 4 – : Méthode d'acquisition des couples sensation/action permettant de calculer a posteriori la perception de l'agent.

ception s'obtient alors en intégrant les actions. Différentes mesures ont été effectuées consécutivement au cours d'une phase d'apprentissage avec une orientation de la caméra fixe. L'apprentissage peut être considéré comme la formation d'un bassin d'attraction dynamique [12]. Ce dernier est créé à l'aide de quelques associations sensori-motrices. La forme des bassins obtenus (fig. 5-*a*) explique le comportement globalement cohérent de l'agent face à l'objet : l'agent va "tomber" dans le bassin perceptif et par conséquent se diriger vers l'objet appris. Ces mesures confirment quantitativement les résultats observés sur les trajectoires. En outre la capacité de généralisation de l'agent est facilement visualisable en dehors de ces zones d'apprentissage.

Le caractère frustre des primitives visuelles extraites et leur faible nombre facilite l'apprentissage des associations sensori-motrices. Mais ce caractère contribue aussi à une faible discrimination entre deux objets différents. Pour une orientation du corps donnée, les réponses des neurones codant l'action de l'agent pour deux objets différents placés côte à côte sont très similaires. Même s'il apparaît que l'apprentissage permet de creuser plus le bassin correspondant à l'objet appris, la différence constatée ne permet pas d'expliquer le comportement global de l'agent qui reste pourtant cohérent. En effet les mesures précédentes ne tiennent pas compte de l'aspect dynamique du système : lorsque l'agent se déplace, il modifie l'orientation de son corps par rapport à l'objet pour se diriger vers ce dernier. Il modifie par conséquent ses sensations. C'est ce déplacement qui va permettre de lever les ambiguïtés existantes (fig. 5-*b*). La dynamique de la boucle sensori-motrice permet d'avoir un comportement correct qui serait impossible à obtenir d'un point de vue statique uniquement.

4 Conclusion et perspectives

Pour résumer, nous avons montré par l'intermédiaire d'une formalisation et d'une mesure de la perception, que la prise en compte des dynamiques sensori-motrices dans les problèmes d'affordance est nécessaire et facilite la tâche de l'agent. Ainsi dans le cadre applicatif définit précédemment, il est important de prendre en compte les 3 dimensions du problème, à savoir la position spatiale de l'agent (deux dimensions) et son orientation par rapport à l'objet. Cette dernière évolue au cours du temps puisqu'elle résulte des actions de l'agent. La prise en

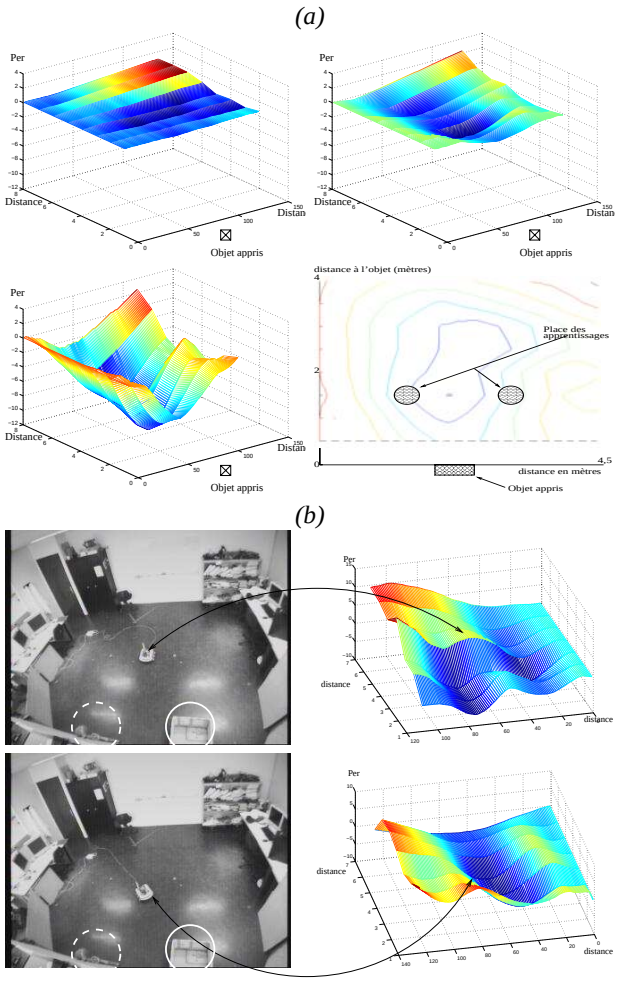


FIG. 5 – *(a)* : Visualisation de la perception à différents stades de l'apprentissage. Les courbes d'iso-perception et la représentation des lieux d'apprentissage montrent la capacité du système à généraliser spatialement. *(b)* : Visualisation de la perception suivant l'orientation du robot. En haut : existence de 2 bassins alors que seul l'objet "carton" (trait plein) a été appris. L'objet lampe (trait en pointillé) est à l'origine du deuxième bassin. En bas : Après quelques actions, l'orientation du robot s'est modifié et un nouvel état perceptif sans ambiguïté apparaît.

compte des dynamiques sensori-motrices permet d'élargir le contexte de la vision active [1] puisque l'agent devient acteur de l'acquisition de ses sensations pour construire sa perception. Nous avons ainsi créé et visualisé un invariant sensori-moteur définissant un comportement simple. Les travaux futurs concerneront l'apprentissage autonome de comportements complexes construits à partir d'interactions entre les différentes boucles sensori-motrices ([15]).

Références

[1] J. Aloimonos, I. Weiss, and A. Bandyopadhay. Active vision. *International Journal of Computer Vision*, pages 333–356, 1998.

[2] R.D. Beer. The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, 11(4) :209–243, 2003.

[3] P. Gaussier. Toward a cognitive system algebra. pages 88–100. European Workshop On Learning Robots, Septembre 2001.

[4] P. Gaussier, J.C. Baccon, K. Prepin, L. Hafemeister, and J. Nadel. Formalization of recognition, affordances and learning or interacting animats. *From animals to animats : Simulation of Adaptive Behavior*, pages 57–64, 2004. MIT Press.

[5] P. Gaussier, C. Joulain, J.P. Banquet, S. Leprêtre, and A. Revel. The visual homing problem : an example of robotics/biology cross fertilization. *Robotics and Autonomous Systems*, 30 :155–180, 2000.

[6] J. Gibson. *The Ecological Approach to Visual Perception.* Houghton Mifflin, Boston, 1979.

[7] A.J. van Doorn J.J. Koenderink. Representation of local geometry in the visual system. *Biological Cybernetics*, pages 55 :367–375, 1987.

[8] C. Lenay, O. Gapenne, and J. Stewart. *Espace du Geste et de la Posture*, chapter Perception spatiale prothétisée : outils d'étude de la localisation et de la reconnaissance de formes. Presses universitaires de Caen, 2001.

[9] T. Lindeberg and H. Romeny. *Geometry-Driven Diffusion*, chapter Linear Scale-Space, pages 1–77. Kluwer Academic Publishers, Dordrecht, Netherlands, 1994.

[10] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, pages 91–110, 2004.

[11] R. Pfeifer and C. Scheier. Sensory-motor coordination : the metaphor and beyond robotics and autonomous systems. *Robotics and autonomous Systems*, 1997.

[12] G. Schöner, M. Dose, and C. Engels. Dynamics of behavior : theory and applications for autonomous robot architectures. *Robotics and Autonomous Systems*, 1995.

[13] A. Steinhage and G. Schöner. The dynamic approach to autonomous robot navigation. page ss7. ISIE, 1997.

[14] E. Thelen and L.B. Smith. *A dynamic system approach to the development of cognition and action*. MIT Press, 1994.

[15] F. Varela, E. Thompson, and E. Rosch. *The Embodied Mind*. MIT press, 1993.

[16] B. Widrow and M.E. Hoff. Adaptive swithing circuits. 4 :96–104, 1960. Institute of Electrical Engineers (IRE) Western Electronic Show and Convention (WESCON) convention records.

Modèle statistique et description locale d'apparence pour la détection des contours des lèvres.

P. GACON[1], P.-Y. COULON[1], G. BAILLY[2]
[1]LIS, Laboratoire des Images et des Signaux, [2]ICP, Institut de la Communication Parlée, 47 Avenue Félix Viallet, 38031 Grenoble
pierre.gacon@lis.inpg.fr , pierre-yves.coulon@lis.inpg.fr , bailly@icp.inpg.fr

Résumé – La segmentation des lèvres est une tâche importante qui a de nombreuses applications dans divers domaines de recherches tels que la reconnaissance de la parole ou la synthèse de visage réaliste. Dans ce travail, nous portons une intention particulière sur la détection du contour intérieur des lèvres, opération difficile à effectuer avec robestesse du fait de non linéarités. Nous proposons une méthode basée sur un modèle actif de la forme et de l'apparence des lèvres et mettant en oeuvre une prédiction non-linéaire de descripteurs locaux d'apparence. Les caractéristiques d'apparence sont également séparés en une composante statique et dynamique afin d'optimiser le traitement d'un séquence d'image en adaptant le modèle au locuteur.

Abstract – Lips segmentation is an important task with many applications in various research area such as speech recognition or realistic face synthesis. In this work we have a special focus on the detection of the inner mouth contour which is quite difficult to achieve with robustness because of nonlinearities . We propose a method based on an active model for shape and appearance with a nonlinear prediction of local appearance descriptors. The appearance is divided in two components, static and dynamic, in order to adapt the model to the speaker in a video sequence.

1. Introduction

La segmentation des lèvres est une tâche qui trouve des applications dans divers domaines de recherches tel la reconnaissance automatique de la parole (dans le cadre d'interface homme/machine par exemple), l'identification de personnes ou l'augmentation de l'intelligibilité de la parole dans des situations bruitées, les télécommunications à très bas débit ou l'animation d'un avatar d'une personne dans un environnement virtuel.

Si les applications sont nombreuses, cette tâche n'en demeure pas moins ardue. La variabilité des formes de lèvre ou de teinte de peau d'un locuteur à un autre, les changements d'éclairage, les mouvements de lèvres peu habituels ou extrêmes, sont autant d'éceuils qui réduisent la précision et la robustesse. Les méthodologies utilisées sont ainsi aussi nombreuses que variées mais l'on peut tout de même discerner trois grandes catégories : les méthodes sans modèle global, les méthodes avec modèle paramétrique et les méthodes avec modèle statistique, cette dernière approche ayant été souvent favorisée ces dernières années.

Dans le cadre des méthodes sans modèle, Zhang [1] avait par exemple proposé d'utiliser la teinte et une détection de contour pour segmenter la bouche mais sans imposer la moindre contrainte de forme, ce qui conduisait à des résultats grossiers. Delmas [2] utilisa les mêmes informations mais avec une optique de contours actifs (ou snakes) afin d'imposer une certaine régularité à la forme. Cette famille de méthode peut donner des résultats parfaitement satisfaisants si les conditions sont bonnes : éclairage contrôlé, bon contraste entre la couleur de la peau et celle des lèvres. Néanmoins, si les conditions sont plus difficiles, la détection devient moins précise et souffre donc d'un manque de robustesse qui en rend l'utilisation difficile pour des applications ou les contours doivent être connus avec finesse. Surtout, rien n'assure que la forme des lèvres détectées sera cohérente

Dans l'objectif d'avoir un contour détecté plus réaliste, plusieurs auteurs ont proposé d'utiliser des modèles paramètriques. La difficulté de ces modèles est de trouver le bon dosage de flexibilité : un modèle peu flexible donnera toujours une forme de lèvre plausible mais échouera à segmenter des formes de bouches inhabituelles alors qu'un modèle trop flexible donnera des formes non-réalistes dans certains cas.

A titre d'exemple, Hennecke et al. [3] ont introduit les patrons déformables (deformable template). Le patron est un modèle de lèvres contrôlé par un jeu de paramètres dont les valeurs sont choisies en minimisant une fonction de coût. Dans un souci d'avoir un modèle suffisamment flexible afin de pouvoir modéliser n'importe quelle configuration de lèvre, Eveno [4] a proposé d'utiliser des courbes paramétriques cubiques pour décrire les contours en imposant conditions et limites aux dérivées des courbes au niveaux des points saillants.

Au cours des dernières années, l'approche par modélisation statistique des lèvres a été de plus en plus plébiscitée. Cootes et al. [5] ont ainsi proposé d'utiliser les modèles de formes actifs (active shape model). La forme d'un objet est apprise sur une base de donnée contenant de nombreux exemples: on procède alors à une analyse en composantes principales (ACP) de la forme afin de diminuer la taille de l'espace de représentation et de commander le modèle en réglant un nombre limité de paramètres. L'intérêt principal est que l'on générera toujours une forme plausible, la seule limitation à la flexibilité du modèle étant la taille de la base d'apprentissage (une configuration absente des exemples pouvant être mal modélisée). Par la suite Cootes et al. [7] ont proposé d'étendre le principe à la description de l'apparence avec les modèles d'apparence actifs (active appearance model) dans lesquelles non seulement la forme de la bouche mais aussi les valeurs des niveaux de gris de la région d'intérêt sont modélisées.

Dans notre travail, cette approche a été retenue et nous avons plus particulièrement dirigé nos recherches sur deux axes principaux. Le premier est la précision de la détection du contour intérieur des lèvres qui présente un comportement non linéaire (ouverture de la bouche, apparition/disparition des dents) et qui est donc moins bien modélisé par un modèle fondé sur une analyse en composantes principales que le contour extérieur. Le second point est une volonté de personnaliser

au fil d'une séquence d'image le modèle au locuteur traité, afin d'augmenter la précision et la rapidité de convergence du modèle.

Pour ce faire nous avons séparé l'apparence globale en deux quantités: une apparence "statique" qui est caractéristique d'un locuteur donné (et qui sera initialisée sur la première image d'une séquence grâce à une classification sommaire des pixels) et une apparence "dynamique" qui correspond aux fluctuations induites par le mouvement (et donc plus particulièrement la parole). En outre, les critères classiques de maximisation du flux du gradient à travers une courbe s'appliquant mal aux contours intérieurs de la bouche, on utilisera des descripteurs locaux gaussiens dont les réponses seront prédites non-linéairement.

2. Base d'apprentissage et données

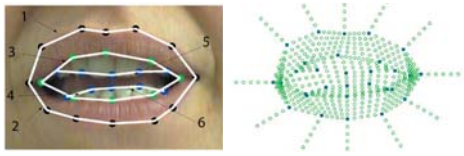


Figure 1 : Exemple d'image annotée de la base d'apprentissage et grille utilisée pour l'apprentissage de l'apparence.

La base d'apprentissage est constituée de séquences vidéo de 12 locuteurs différents. Afin de construire le modèle, N=450 images ont été annotées manuellement en 30 points de contrôle décrivant la forme: 12 points sur le contour externe des lèvres et 8 points sur le contour interne ainsi que 10 points sur les dents (lorsque les dents ne sont pas visibles, les points correspondant existent toujours mais sont confondus avec le contour intérieur de la bouche). Les coordonnés des points pour chaque image sont sauvegardées dans les vecteurs de forme $\boldsymbol{s}_i$ ($1 \le i \le N$). L'opérateur assigne également à chaque image un Etat Général de la Bouche (EGB). L'EGB est une variable d'état décrivant sommairement 4 états de la bouche (fermée, ouverte, grande ouverte, sourire).

Comme les notions de couleur et de luminosité sont mélangées dans l'espace couleur RVB, nous utilisons l'espace couleur YCbCr où la chrominance et la luminance sont décorellées. L'apparence échantillonnée (c'est à dire les 3 valeurs YCbCr associées à un pixel) est extraite de l'image en 728 points d'une grille d'apprentissage déduite des 30 points de contrôle (cf figure 1). L'utilisation de la grille définit précisément si un pixel correspond à de la peau, aux lèvres, aux dent ou au fond de la bouche. Cela est particulièrement utile pour décrire l'intérieur de la bouche et éviter un effet de flou.

Dans notre méthode, nous faisons la distinction entre l'apparence statique et l'apparence dynamique. La première correspond aux caractéristiques spécifiques d'un locuteur donné (couleurs des lèvres et de la peau et contraste entre les deux) et donc à la variabilité extrinsèque de l'apparence. On la définit comme l'apparence moyenne d'un locuteur donné et elle est stockée dans les vecteurs $\boldsymbol{a}_{m,i}$ ($1 \le i \le N$). A l'inverse, l'apparence dynamique correspond aux variations intrinsèques et est la différence entre l'apparence totale et l'apparence statique. Elle est stockée dans les vecteurs $\boldsymbol{a}_{d,i}$ ($1 \le i \le N$).

Enfin, notons que notre modèle n'étant qu'en deux dimensions il ne tolère donc que de petites rotations du visage.

3. Modélisation

3.1 Modèle actif de forme et d'apparence

On procède ici à l'analyse en composante principale des données collectées dans la partie 2. Par exemple dans le cas de la forme (contenue dans les vecteurs $\mathbf{s}_i$) on calcule le vecteur moyen $\overline{\mathbf{s}}$ puis la matrice de covariance $\mathbf{S}$. Les vecteurs propres de $\mathbf{S}$ correspondent aux différents modes propres de variations, ceux associés à de grandes valeurs propres modélisant une part plus importante de la variance des données.

$$\bar{\boldsymbol{s}} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{s}_i \;;\; \mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \left(\boldsymbol{s}_i - \bar{\boldsymbol{s}}\right)\left(\boldsymbol{s}_i - \bar{\boldsymbol{s}}\right)^T$$

On choisit de garder 95% de la variance totale, les vecteurs propres correspondant étant sauvés dans la matrice $\mathbf{P}_S$. Finalement, on peut générer n'importe quelle forme s en ajustant simplement le vecteur $\boldsymbol{\sigma}$ (contenant les poids affectés à chaque mode propre) dans l'équation suivante: $\boldsymbol{s} = \overline{\mathbf{s}} + \mathbf{P}_s\boldsymbol{\sigma}$. On procède de même avec les deux apparences puis on procède à une ACP de deuxième niveau similairement à [6] afin d'avoir un modèle combiné modélisant de façon cohérente la forme et l'apparence dynamique. On obtient le jeu d'équations suivant:

$$(1)\ \mathbf{c} = \begin{bmatrix} \mathrm{W}.\boldsymbol{\sigma} \\ \alpha_d \end{bmatrix} = \bar{\boldsymbol{c}} + \mathbf{P}_c\chi \;\Rightarrow \begin{cases} (2)\ \boldsymbol{s} = \bar{\boldsymbol{s}} + \mathbf{P}_s\boldsymbol{\sigma} \\ (3)\ \boldsymbol{a}_d = \overline{\boldsymbol{a}_d} + \mathbf{P}_d\alpha_d \end{cases}$$

$$(4)\ \boldsymbol{a}_m = \overline{\boldsymbol{a}_m} + \mathbf{P}_m\alpha_m$$

L'équation (2) contrôle donc le modèle de forme, l'équation (3) contrôle l'apparence dynamique et l'équation (4) l'apparence statique. Enfin l'équation (1) correspond au modèle combiné forme/apparence dynamique avec W un coefficient normalisant les unités. Segmenter les lèvres sur une image revient finalement à optimiser le jeu de paramètres correspondant aux modes principaux des ACP retenus, soit 18 dans notre cas : 9 dans le vecteur χ et 9 dans le vecteur $\boldsymbol{\alpha}_\mathbf{m}$.

Les valeurs moyennes de χ sont également calculées pour chaque EGB et seront utilisées par la suite.

3.2 Descripteurs locaux, principe et validation

La maximisation d'un flux de vecteur gradient à travers les courbes des lèvres est un critère souvent retenu pour en accomplir la segmentation. Néanmoins il s'applique mal aux contours intérieurs du fait du comportement non linéaire de cette zone (fermeture/ouverture, apparition du contour des dents ce qui peut entre autre entraîner la disparition ou l'inversion de la direction des vecteurs gradients).

Pour résoudre ce problème, nous avons utilisé des descripteurs locaux d'apparence basés sur les réponses de filtres dérivés gaussiens (limités à la première dérivée).

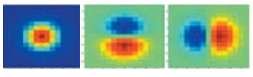


Figure 2 : Réponses impulsionelle des filtres gaussiens G (moyenne), Gx et Gy (gradients horizontal et vertical)

Le critère calculé est la différence entre la réponse de ces descripteurs le long des contours et une prédiction de leur réponse faite pour une configuration donnée de la forme.

Cette prédiction est effectuée par un réseau de neurones non linéaire dont l'entrée est la forme et dont la dimension de l'espace de sortie (la réponse des filtres) est diminuée par ACP afin de réduire la taille du réseau et les temps de calcul. Ainsi

forme et apparence sont décrites par des modèles linéaires mais leur couplage est non-linéaire.

Nous avons testé validé la pertinence de cette idée en comparant ses performances de segmentation à celles d'autres méthodes proches. Ces comparaisons se feront dans un cas mono-locuteur (avec une base d'apprentissage de 50 images). La variabilité inter-locuteur n'est donc pas prise en compte, l'apparence statique correspondant simplement à l'apparence moyenne du locuteur. La réponse des filtres ou les valeurs des pixels dépendent ainsi uniquement du mouvement des lèvres et des différences d'illumination, effet qui peut être atténué en filtrant la luminance par le filtre rétine [7] (opérateur permettant de diminuer les variations d'illuminations sur une image en utilisant un modèle s'inspirant de l'oeil humain).

Dans tous les modèles, la forme sera décrite par un modèle actif obtenu par une ACP. En revanche l'apparence sera décrite par A) les descripteurs gaussiens locaux ou B) les valeurs des pixels obtenus grâce à la grille d'échantillonnage. Pour chacune de ces modélisations de l'apparence, nous allons tester trois types de méthode: 1) un modèle d'apparence classique type [6] (ACP sur l'apparence puis ACP de second niveau pour avoir un modèle combiné forme/apparence), 2) l'apparence est prédite à partir des paramètres du modèle de forme grâce à une régression linéaire (entraînée sur la base d'apprentissage), 3) l'apparence est prédite à partir des paramètres du modèle grâce à un réseau de neurones non-linéaire. Il est à noter que pour les méthodes 2) et 3) le nombre de paramètres à optimiser est inférieur à la méthode 1) puisque seule la forme est optimisée.

Nous testons ensuite les 6 cas sur 50 images non présentes dans la base d'apprentissage. Les erreurs sont données en pourcentage de la largeur de la bouche.

méthode	contour extérieur	contour intérieur	dents	tous points	nombre d'itérations
A) 1)	1.6 / 1.1	2.4 / 1.7	2.6 / 1.8	2.1 / 1.5	29.8
2)	1.7 / 1.2	2.9 / 1.8	3.1 / 2.2	2.4 / 1.7	17.2
3)	1.4 / 0.8	1.8 / 0.9	1.8 / 0.9	1.5 / 0.8	9.8
B) 1)	1.4 / 0.9	2 / 1.1	2.2 / 1.2	1.8 / 1	31.1
2)	1.5 / 1	2.2 / 1.3	2.5 / 1.4	2 / 1.2	15.1
3)	1.5 / 0.9	2.1 / 1.3	2.4 / 1.3	2 / 1.1	15.5

Tableau 1: Résultats donnés en pourcentage: erreur/ écart type. Méthodes: A) apparence décrite par descripteurs locaux, B) apparence décrite par pixels, 1) ACP sur l'apparence, 2) apparence prédite par régression linéaire sur la forme, 3) apparence prédite par réseau de neurones.

Les résultats montrent que les meilleurs résultats sont obtenus pour la méthode proposée dans notre contribution.

On constate que les 3 méthodes où l'apparence est décrite par la grille d'échantillonnage donnent des résultats très proches les unes des autres, l'utilisation de la grille permettant de réduire les problèmes dus aux non-linéaritées grâce à sa description précise de la zone intérieure de la bouche. La méthode B)1) donne alors les meilleurs résultats, au prix d'un nombre supérieur d'itérations dû au plus grand nombre de paramètres à régler.

L'utilisation des descripteurs locaux ne donne de résultats véritablement intéressants que dans le cas de la prédiction non-linéaire. Dans les cas 1) et 2), la linéarité de la relation forme/ apparence résulte en effet en une mauvaise segmentation de l'intérieur de la bouche, ce qui entraîne également des erreurs sur le contour extérieur. En revanche, avec la prédiction non linéaire, les descripteurs locaux permettent d'obtenir de meilleurs résultats du fait qu'ils sont moins sensibles au bruit qu'une mesure directement liée aux pixels et que leur information de type gradient leur permet de coller au plus près des contours.

La validation ayant été probante, ce principe a été étendu au cas multi-locuteur. Pour tenir compte de la variabilité extrinsèque rajoutée, l'apparence statique (qui est justement caractéristique de cette variabilité) a alors été rajoutée en entrée du réseau de neurone.

Pour effectuer la prédiction nous avons décidé d'utiliser un réseau à deux couches construit par rétropropagation. Nous avons donc les 18 entrées correspondant aux paramètres du modèle, 15 unités cachées et 15 sorties correspondant à l'espace réduit obtenu grâce à l'ACP sur la réponse des filtres.

3.3 Classification initiale par modèle couleur

Un modèle colorimétrique est utilisé pour discriminer grossièrement les pixels de lèvres et de peau, ceci afin d'initialiser l'apparence statique sur la première image d'une séquence d'un locuteur donné. A partir de la base d'apprentissage annotée, les valeurs CbCr de chacune des deux classes peau et lèvre sont extraites puis leur distribution statistique est modélisée par une mixture de gaussienne. A partir d'un simple calcul de probabilité, on peut alors déterminer si un pixel a plus de chance d'être de la peau ou des lèvres (on fixe également un seuil assez restrictif sur la valeur de la probabilité pour ne pas assigner les pixels incertains). On obtient alors une segmentation grossière mais suffisante pour l'utilisation qui en est faite.

Figure 3 : Exemple de classification de pixels (blanc: lèvre, gris: peau, noir: non affecté)

3.4 Modèle de commissures des lèvres

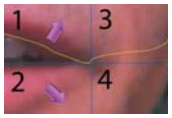


Figure 4 : Zone des commissures des lèvres et ligne des minima de luminance (en jaune).

Les commissures des lèvres sont considérées comme des points-clés dont la connaissance détermine la position et l'échelle de la bouche. Ces points sont délicats à détecter car ne se trouvant généralement pas sur des contours nets, mais au contraire dans des zones floues ou ombrées. Ainsi, ils seront vus comme les points de jonction de quatre quadrants de caractéristiques différentes (repérés sur la figure 4). Les régions 1 et 2 sont non homogènes car présentant un contour entre les lèvres et la peau au contraire des régions 3 et 4. Chacune de ces régions sera décrite par un jeu de filtres dérivés gaussiens dont les réponses seront modélisées statistiquement par une mixture de gaussiennes.

Si l'on peut considérer, comme cela a déjà été proposé par [4], que les points de commissures se trouvent sur le ligne des minima de luminance par colonne (la ligne jaune sur la figure

4), il reste seulement à trouver l'indice des colonnes. On teste alors les points de la ligne d'intérêt et l'on sélectionne finalement le couple de points le plus probable comme commissures des lèvres. Des valeurs initiales pour les paramètres S et T, correspondant à l'échelle et la position des lèvres, sont ainsi obtenus. La méthode est robuste dans plus de 95% des cas pour une précision correspondant à 5% de la taille réelle des lèvres.

4. Segmentation

Lors du traitement d'une image inconnue, nous voulons segmenter les lèvres en trouvant le meilleur jeu de paramètres χ, α_m, S et T pour notre modèle actif. Cette tâche est effectuée en minimisant une fonction de coût grâce à l'algorithme classique du simplex. Cette fonction est la somme pondérée de deux erreurs quadratiques moyenne (EQM): l'EQM entre la réponse observée sur l'image des descripteurs locaux et leur réponse prédite et l'EQM entre l'apparence modélisée et l'apparence observée sur l'image. Les paramètres sont considérés avoir convergé dès que la différence entre les valeurs maximum et minimum du simplex passe en dessous d'un seuil.

Pour la première image d'une séquence, les paramètres du simplex sont initialisés en testant les jeux de paramètres moyens correspondant aux différents EGB (pour χ), en utilisant la classification de pixel (pour α_m) et en utilisant le modèle de commissures (pour S et T). Par la suite, les valeurs finales des paramètres des images précédentes sont utilisées pour obtenir les valeurs initiales.

Au fil de la séquence, α_m va converger (l'intervalle de recherche des paramètres étant diminué progressivement) ce qui correspond au fait que les caractéristiques du locuteur sont de mieux en mieux connues. Lorsque la convergence a eu lieu (la variation d'α_m passe en-dessous d'un seuil), la fonction de coût sera calculée comme étant uniquement l'EQM sur les descripteurs, seul le vecteur de paramètre χ sera dès lors optimisé à chaque nouvelle image. Le principe est résumé par la figure 5.

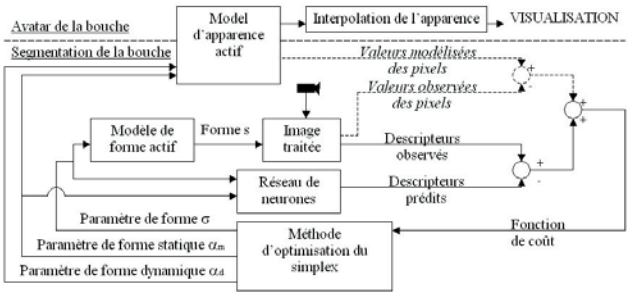


Figure 5 : Principe de la segmentation.
En pointillé: calcul effectué tant que α_m n'a pas convergé.

5. Résultats

La table 1 présente quelques résultats de segmentation chiffrés, tandis que les figures 6 et 7 montrent des exemples de segmentation et de génération d'avatar par interpolation de l'apparence échantillonnée. Au final, notre méthode produit des résultats fiables et robustes et semble particulièrement adaptée pour modéliser avec fidélité les contours intérieurs de la bouche.

Position de l'erreur	Contour extérieur	Contour intérieur	Dents	Total
Test sur la base d'apprentissage	2.7 / 1.3	2.8 / 1.4	3 / 1.4	2.8 / 1.3
Locuteur dans la base, image inconnue	2.7 / 1.3	2.8 / 1.5	3.1 / 1.4	2.8 / 1.3
Leave-one-out	3 / 1.4	3.1 / 1.4	3.3 / 1.5	3.1 / 1.3

Tableau 2 :
Résultats donnés en pourcentage: erreur/écart type, leave-one-out: locuteur testé enlevé de l'apprentissage

Figure 6 : Exemples de segmentation

Figure 7 : Exemples d'interpolation d'avatars

Une évaluation subjective de la méthode est en cours d'expérimentation en quantifiant l'intéllégibilité du locuteur synthétisé (prononçant des numéros de téléphone pour divers niveaux de bruit sur le canal audio).

6. Références

[1] X. Zhang, R. M. Mersereau, M. A. Clements and C. C. Broun, "Visual Speech Feature Extraction for Improved Speech Recognition", In Proc. ICASSP'02, 2002, pp. 1993-1996

[2] P. Delmas, N.Eveno, and M. Lievin, "Towards Robust Lip Tracking", *International Conference on Pattern Recognition (ICPR'02),* Québec City, Canada, Août 2002

[3] M. Hennecke, V. Prasad, and D. Stork. "Using deformable templates to infer visual speech dynamics", *28 h Annual Asimolar Conference on Signals, Systems, and Computer*, *volume 2*, IEEE Computer, Pacific Grove, pages 576-582, 1994.

[4] N.Eveno, A. Caplier, and P-Y Coulon, "Automatic and Accurate Lip Tracking", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no.5, pp. 706-715, Mai 2004

[5] T.F. Cootes, C.J. Taylor, D.H. Cooper, "Active Shpae Models - Their Training and Application", Computer Vision and Image Understanding, Vol. 61, No. 1, Janvier, pp. 38-59, 1995

[6] T. F. Cootes. "Statistical models of appearance for computer vision", Rapport technique disponible en ligne sur http://www.isbe.man.ac.uk/bim/refs.html, 2001.

[9] W.H.A. Beaudot, "The neural information processing in the vertebrate retina: A melting pot of ideas for artificial vision", PhD Thesis in Computer Science, INPG (France) décembre 1994

MGS2 : Optimisation multicritères de contours actifs par algorithmes génétiques

Nicolas CLADEL, Renaud SÉGUIER

Équipe SCEE (Supelec - IETR)
Supelec, avenue de la boulaie, BP 81127, 35511 Cesson-Sévigné Cedex, France
nicolas.cladel@supelec.fr, renaud.seguier@supelec.fr

Résumé – Dans cet article nous proposons une évolution de notre précédent travail sur l'optimisation multicritères de contours actifs. Cette approche permet une exploration globale de l'image et une gestion efficace de plusieurs énergies par la représentation de Pareto. Notre nouvel algorithme, le MultiObjective Genetic Snakes 2 (MGS2), associe l'algorithme multicritères NSGA2 [4] au codage des contours actifs des MGS [3]. Les MGS2 exploitent de nouvelles énergies d'attache à l'image afin d'assurer la convergence des contours vers un objet creux en environnement bruité à partir d'une initialisation aléatoire. Nous procédons ici à l'étude de ces énergies et à l'analyse de la convergence de l'algorithme à travers une série de tests sur une base d'objets synthétiques. Nous présenterons également des résultats de l'application de MGS2 à la lecture labiale.

Abstract – We propose in this paper an evolution of our preceding work on the optimization multicriterion of active contours. This approach allows a global analysis of the image and an efficient use of several energy. Our new algorithm, MultiObjective Genetic Snakes 2 (MGS2) associates the multiobjective algorithm NSGA2 [4] the coding of active contours of the MGS [3]. The MGS2 exploit new energies of fastener to the image in order to ensure the convergence of contours towards a hollow object in disturbed environment starting from a random initialization. We carry out here the study of these energies and the analysis of the convergence of the algorithm through a series of tests on a basis of synthetic objects. We will also have results of the application of MGS2 at the labial reading.

1 Introduction

Notre équipe s'intéresse à l'analyse d'images en environnement réel pour des applications telles que la visiophonie et les interactions homme-machine. Nous nous focalisons dans cette étude sur la segmentation de lèvres en environnement réel. Nous souhaitons être peu dépendant des informations à priori sur la position des contours ou sur la teinte de l'objet. Compte tenu de la variabilité de la forme des bouches inter et intra locuteurs, nous avons fait le choix de ne pas prendre en compte un modèle à priori des contours recherchés. Dans un soucis de généralisation de l'application de notre algorithme, nous allons étendre le domaine d'application du MGS2 à d'autres objets. l'objectif de cette étude est donc l'extraction par contours actifs d'objets similaires aux lèvres, c'est à dire creux et peu convexes, dans un environnement bruité.

Le principe adopté est l'utilisation conjointe des énergies de type région et gradient afin d'obtenir une segmentation efficace en présence de bruit important. Nous avions proposé une telle approche de coopération énergétique [3] appliquée aux contours actifs dans le cadre de la lecture labiale. Cette méthode donnait de bons résultats mais était lente et complexe. Nous en présentons ici une amélioration en intégrant le formalisme des contours des MGS dans une structure génétique multicritères simple et rapide, le NSGA2 [4]. Ces améliorations nous permettent de gérer plus efficacement un plus grand nombre d'énergies. Nous proposons également de nouveaux potentiels d'attache à l'image que nous étudions à travers des tests sur une base d'objets synthétiques.

Cet article est découpé en trois parties. La première concerne l'adaptation des contours actifs aux algorithmes génétiques, la seconde les énergies proposées et leur étude et la dernière décrit l'application des MGS2 à la lecture labiale.

2 Algorithme Multiobjective Genetic Snakes 2 (MGS2)

Cet algorithme associe le formalisme des contours actifs des MGS [3] à la structure génétique multicritères NSGA2. Les MGS permettent une exploration globale de l'image et une gestion efficace de plusieurs potentiels. Nous allons rappeler ici le codage et la déformation des contours actifs que nous effectuons dans MGS et MGS2.

Les double snakes sont représentés par un chromosome codé sous la forme d'un tableau de dimension $2 \times N$, N étant le nombre de nœuds d'un contour (figure 1). Un gêne représente donc un nœud d'un contour et ne pourra prendre qu'un ensemble défini de valeurs dans l'image [1]. Les double contours découpent l'image en trois régions, la région centre à l'intérieur du contour interne, la région objet entre les contours interne et externe, et la région externe à l'extérieur du contour externe. Les contours ainsi modélisés sont initialisés aléatoirement puis

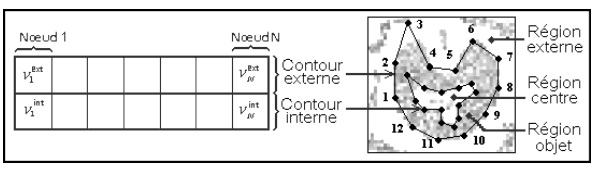


FIG. 1 – Codage des contours

vont être modifiés par les opérateurs génétiques de croisement et de mutation. La sélection d'une suite de gênes revient à découper une portion du contour (figure 2-a). Le croisement de deux contours P1 et P2 se traduit par l'échange de portions de contours (figure 2-b). Cette opération est appliquée simultanément aux contours interne et externe. La mutation d'un gêne implique le déplacement d'un nœud (figure 2-c).

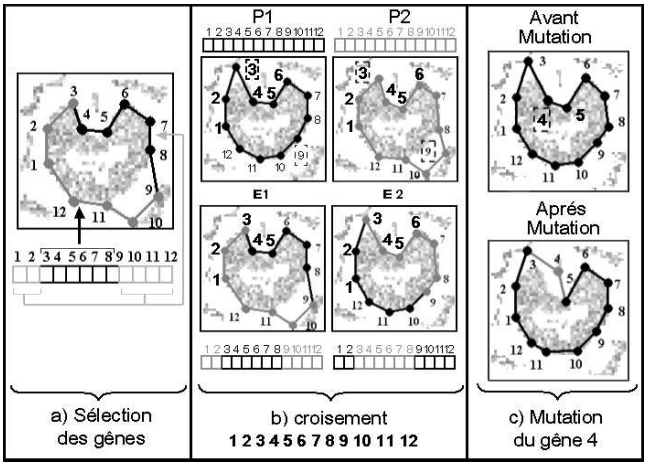


FIG. 2 – Déformation des contours par opérateurs génétiques

La représentation de Pareto attribue un rang [4] à chaque individu de la population en fonction de ses valeurs énergétiques. Les MGS fournissent ainsi en fin de traitement un ensemble de doubles contours potentiellement optimaux parmi lesquels un individu optimal doit être sélectionné en fonction des préférences de l'utilisateur.

3 Énergies

3.1 Définition des descripteurs

La segmentation doit prendre en compte de manière coopérative les énergies de type contour et de type région. Nous travaillons sur quatre images pour caractériser l'objet à extraire : l'image originale de l'objet (figure 3-a) et deux images issues des pré-traitements effectués sur cette image originale. La première est binaire (figure 3-b) et la seconde en niveaux de gris représente l'amplitude du gradient (figure 3-c). La quatrième image représente les contours orientés vis à vis du centre de l'image : les nœuds du contour externe sont attachés aux pixels noirs et ceux du contour interne aux pixels gris (figure 3-d). Nous proposons trois types de descripteur :

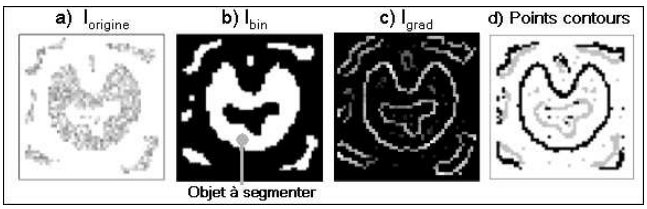


FIG. 3 – Images sources.

– **Un descripteur gradient (**$D_{gradient}$**)** proportionnel à l'amplitude du gradient le long des contours interne et externe et au nombre de points contours appartenant aux courbes. Ce descripteur tend vers 0 lorsque les doubles snakes sont situées sur des pixels de forte amplitude gradient identifiés comme points contours :

$$D_{gradient}(C) = e^{-\frac{NbPtsContour(C)}{NbPtsTotal(C)} \sum\limits_{\{x,y\}\in C} |\nabla I_{origine}(x,y)|}$$

avec :
  - $C = C_{interne} \bigcup C_{externe}$
  - NbPtsContour(C) : nombre de points contours appartenant aux contours interne et externe.
  - NbPtsTotal(C) : nombre de pixels appartenant au contour.
– **Un descripteur variance (**$D_{variance}$ relatif à la variance des régions décrites par les contours interne et externe (figure 1) dans l'image originale. Ce descripteur tend ainsi vers 0 lorsque la région est homogène.

$$D_{variance} = \sigma^2(\Omega) = \frac{1}{|\Omega|} \int\int_{\Omega} (I_{origine}(x,y) - \mu(\Omega))^2 dxdy)$$

avec :
  - $|\Omega|$: taille de la région Ω courante (objet ou centre).
  - $\mu(\Omega)$: moyenne des niveaux de gris des pixels appartenant à la régionΩ.
– **Un descripteur d'agrégation (**$D_{agregation}$**)** calculé à partir de l'image binaire et proportionnel au nombre de pixels à 0 ou à 1 présents dans les régions décrites par les contours. Des coefficients pondérateurs permettent de privilégier l'ajout de pixels ayant le label de la région voulue par rapport à la suppression des autres pixels. Ce descripteur tend ainsi vers 0 pour une région constituée de nombreux pixels de la région recherchée :

$$D_{agregation}(\Omega) = NbPixelTotal - \left[\alpha_{agreg} \cdot NbPixels_{\Omega}^{label(\Omega)} - \alpha_{sup} \cdot NbPixels_{\Omega}^{label(\overline{\Omega})}\right]$$

avec :
  - *NbPixelTotal* est le nombre total de pixels de l'image : *NbPixelTotal*=hauteur.largeur .
  - $NbPixels_{\Omega}^{label(\Omega)}$ est le nombre de pixels ayant le label *label*(Ω) dans la région Ω.
  - label(Ω) est le label associé aux régions d'intérêt :
$$\begin{pmatrix} \Omega \\ objet \\ centre \end{pmatrix} \rightarrow \begin{pmatrix} label \\ 1 \\ 0 \end{pmatrix}$$
  - $\overline{\Omega}$ est le complémentaire Ω :
$$\begin{pmatrix} \Omega \\ objet \\ centre \end{pmatrix} \rightarrow \begin{pmatrix} \overline{\Omega} \\ centre \\ objet \end{pmatrix}$$
  - α_{agreg} et α_{sup} sont des coefficients pondérateurs. Si α_{agreg} est supérieur à α_{sup}, l'agrégation de pixels région sera privilégiée à la surpression des pixels non-région. Ces coefficients sont normés : $\alpha_{agreg} + \alpha_{sup} = 1$.

Les descripteurs de variance et d'agrégation induisent des comportements complémentaires : la variance permet de tendre vers une région compacte homogène alors que l'agrégation va autoriser la segmentation d'une région intégrant du bruit.

3.2 Étude des configurations énergétiques

Nous allons ici étudier les différentes configurations d'énergies possibles exploitées par les MGS2. Pour déterminer la configuration optimale, nous testons notre algorithme sur une base d'objets synthétiques. Cette base est constituée de différents objets et sources de bruits. Afin d'évaluer la qualité de la convergence induite par les configurations, nous calculons le nombre de pixels différents (N_{diff}) entre la surface de l'objet segmenté et celle de l'objet réel. La figure 4 montre l'évolution de cette valeur au cours du temps. Nous étudions ainsi la meilleure valeur (pointillés), la valeur moyenne (semi-pointillés) et la valeur correspondant à l'individu optimal (continu). Les résultats de la segmentation sont présentés dans la figure 5.

Les configurations énergétiques possibles peuvent prendre en compte cinq énergies : l'énergie gradient et les énergies région (variance ou agrégation) appliquées aux régions définies par le contour externe ou interne. Nous utilisons deux contours de 28 nœuds chacun. L'objet à segmenter est creux, situé au centre de l'image(figure 3).

Nous avons testé les cinq énergies ensembles (figure 4-E1). Les doubles snakes convergent vers une mauvaise segmentation car l'algorithme gère difficilement autant d'énergies. Pour analyser l'influence des deux descripteurs de type région, nous avons choisi deux configurations énergétiques : la première est composée du gradient et de l'agrégation appliqués aux régions centre et objet (figure 4-E2), la seconde est définie à partir de E2 en remplaçant l'agrégation par la variance (figure 4-E3). Nous observons que les valeurs optimales vers lesquelles tendent ces descripteurs au cours du temps ne correspondent pas à la segmentation cherchée. Ainsi E2 guide les doubles contours vers une région homogène restreinte (figure 5-R2) et E3 vers une région étendu au bruit voisin de l'objet (figure 5-R3). Afin de pallier à ces problèmes nous rajoutons à la configuration E3, le descripteur de l'agrégation appliqué au contour externe (figures 4-E4 et 5-R4). Les double snakes convergent bien vers l'objet et plus rapidement que que dans les autres cas. La configuration énergétique que nous retiendrons est donc composée de quatre énergies : le gradient, la variance appliquée au contour interne, la variance sur le contour externe et l'agrégation sur ce même contour.

4 Application à la lecture labiale

Nous allons maintenant présenter l'application des MGS2 à la lecture labiale [6]. Les MGS2 sont ici associés à un opérateur de contours actifs variationnels développé dans [1] et implémentés en mode vidéo. Ces tests sont réalisés sur des séquences de prononciation de chiffres de la base M2VTS [5].

Le procédé de lecture labiale mis en œuvre recherche un mot unique dans une séquence. La base d'images M2VTS est composée de séquences de chiffres prononcées par dix personnes, avec quatre séquences par personne. Chaque séquence est prise à une semaine d'intervalle. L'apprentissage est réalisé sur 3 séquences et la reconnaissance est effectuée sur la quatrième.

Chacun des chiffres prononcé est associé à une signature composée d'informations sur la forme de la bouche, le nombre de pixels labellisés lèvres lors de pré-traitements région [3] et les dérivées de chacun de ces paramètres. Nous obtenons ainsi un vecteur de huit coordonnées pour chaque image. La classification du mot est réalisée par la méthode du Dynamic Time Warping.

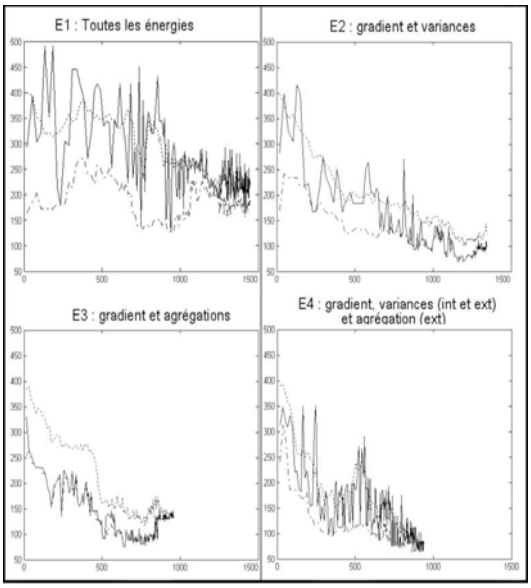


FIG. 4 – Courbes des tests énergétiques.

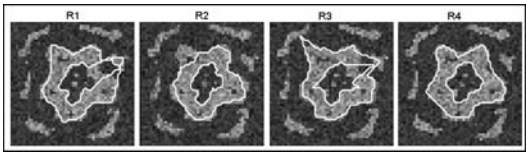


FIG. 5 – Contours des individus optimaux des tests énergétiques.

Les informations sur la forme de la bouche (hauteur et largeur des lèvres) sont extraits des contours segmentés en mode vidéo. Ce processus de segmentation exploite le MGS2 et un opérateur de contours actifs [1] au cours de deux étapes. La première étape est une phase d'initialisation (figure 6) : MGS2 est appliqué sur la première image de chaque séquence et le contour optimal obtenu est affiné par l'algorithme de contours actifs. La seconde étape est de type vidéo et utilise le résultat de chaque image pour initialiser la segmentation par contours actifs de l'image suivante (figure 7).

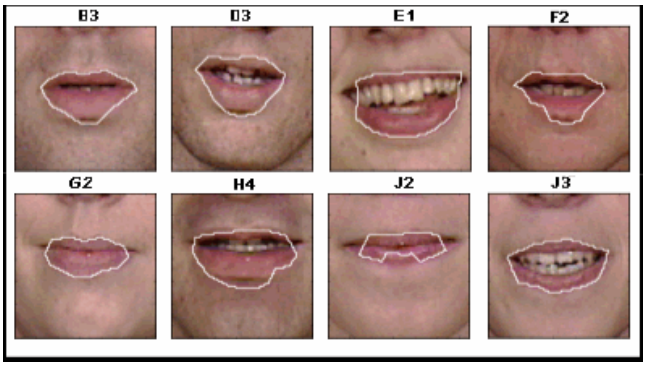


FIG. 6 – Exemples de segmentation d'images initiales de séquences de locuteurs

Nous obtenons ainsi 75% de bonne segmentation des images

initiales. Nous observons des imperfections au niveau des commissures des lèvres (E1-J2, figure 6). En effet les pixels appartenant à ces régions ne sont pas labellisés lèvre en raison de leur trop faible luminosié et sont donc rejetés par l'algorithme.

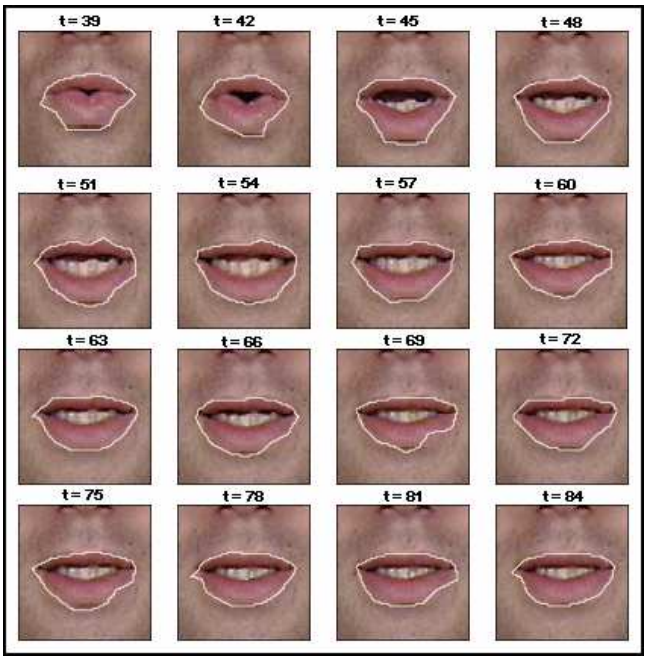


FIG. 7 – Extrait d'une séquence segmentée en mode vidéo

Le temps moyen de segmentation de l'étape d'initialisation est de 150 secondes sous matlab et celui de la seconde étape est de 25 secondes par image.

Nous obtenons alors une moyenne de 69% de bonne classification sur les dix personnes de M2VTS. Nous avons ainsi améliorés nettement nos précédents résultats qui étaient de 58% sur la première personne. En effet, notre ancienne méthode n'était pas assez robuste pour être appliquée sur les dix personnes.

5 Conclusion

Dans cet article nous avons présenté une nouvelle version de l'algorithme MGS [3] effectuant une optimisation multicritères de contours actifs par algorithmes génétiques. Les MGS2 permettent ainsi de segmenter plus rapidement un objet dans une image bruitée. Cet algorithme est également plus simple à paramétrer car il ne gère qu'une seule population de contours. Ces améliorations nous permettent aussi de manipuler des contours plus importants, 28 nœuds contre 8 pour les MGS.

Au cours de tests sur des objets synthétiques, nous avons pu observer l'influence des énergies d'attache à l'image sur la convergence des contours actifs et déterminer les paramètres énergétiques optimaux des MGS2. L'efficacité des MGS2 a été vérifiée à travers une application de lecture labiale.

Dans le cadre de travaux futurs, nous chercherons à améliorer la définition des énergies et à appliquer les MGS2 à l'extraction des contours cardiaques. Nous développerons également la spécialisation des MGS2 aux contours labiaux afin d'accélérer la segmentation des lèvres pour la lecture labiale, notamment par l'ajout de modèles. Ces modèles devront être suffisamment flexibles pour s'adapter à la variabilité de forme des lèvres.

Remerciements

Cette étude a été soutenue par la région Bretagne.

Références

[1] N. Cladel et R. Seguier. *Active contours multiobjective optimisation by hybrid algorithms*. Irish machine vision and image processing (IMVIP), 2004.

[2] C.A. Coello Coello et G.T. Pulido. *Multiobjective optimization with a micro genetic algorithm*. Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'2001), Morgan Kaufmann Publishers,San Francisco, California, pp 274–28, 2001.

[3] Renaud Seguier et Nicolas Cladel. *Multiobjectives Genetic Snakes : Application on Audio-Visual Speech Recognition*. 4th EURASIP Conference focused on Video, Image Processing and Multimedia Communications,Zagreb, Croatia, 2003.

[4] K. Deb. *A fast elitist non dominated sorting genetic algorithm for multiobjective optimization : NSGA II*. Parallel problem solving form nature - PPSN VI, Springer lecture notes in computer science, pp. 849-858, 2000.

[5] S. Pigeon and L. Vandendrope. *The M2VTS Multimodal Face Database (Release 1.00)*. J. Bigun, G. Chollet, and G. Borgefors eds., Audio- and Video-Based Biometric Person Authentication, Springer, pp 403-409, 1997.

[6] P. DAUBIAS and P. DELEGLISE. *Evaluation of an automatically obtained shape and appearance model for automatic audio visual speech*. Eurospeech, 2001.

Séparation voix / musique à partir d'enregistrements mono : quelques remarques sur le choix et l'adaptation des modèles

Alexey OZEROV[1], Rémi GRIBONVAL[2], Pierrick PHILIPPE[1], Frédéric BIMBOT[2]

[1]France Télécom R&D
4, rue du Clos Courtel, BP 91226, 35512 Cesson Sévigné cedex, France

[2]IRISA (CNRS & INRIA) - projet METISS
Campus de Beaulieu, 35042 Rennes Cedex, France

alexey.ozerov@francetelecom.com, remi.gribonval@irisa.fr,
pierrick.philippe@francetelecom.com, frederic.bimbot@irisa.fr

Résumé – Le problème de l'extraction de voix chant´ee dans des enregistrements musicaux monophoniques, c'est-à-dire la s´eparation voix / musique avec un seul capteur, est ´etudi´e. Les approches utilis´ees sont bas´ees sur les modèles statistiques *a priori* de deux sources. Une ´etude comparative des diff´erentes modèles et estimateurs est effectu´ee, ainsi qu'une ´etude d'impact de l'h´et´erog´en´eit´e entre les donn´ees d'apprentissage et les donn´ees à s´eparer. On montre que l'adaptation du modèle de la musique sur les parties non vocales permet d'obtenir des bonnes performances dans un cadre relativement r´ealiste.

Abstract – The problem of singing voice extraction from mono audio recordings, i.e., one microphone separation of voice and music, is studied. The approaches are based on *a priori* probabilistic models for two sources. A comparative study of different models and estimators is done together with a study of the impact of heterogeneity between training data and data to be separated. We show that the adaptation of music model from the non vocal parts allows to obtain the good results in realistic conditions.

1 Introduction

Nous nous intéressons `a l'extraction de voix chantée dans des enregistrements musicaux, c'est-`a-dire `a la séparation de la voix par rapport `a l'accompagnement musical. Nous considérons des enregistrements mono (par opposition `a stéréo). Il s'agit donc de séparation de sources avec un seul capteur. Nous supposons que chaque enregistrement est une simple somme $x(k) = v(k) + m(k)$ du signal de voix $v(k)$ et du signal de musique $m(k)$. Etant donné le signal observé $x(k)$, notre probl`eme consiste `a estimer la contribution de la voix $\hat{v}(k)$ pour l'indexer, la reconnaître, etc.

Notre approche est basée sur des mod`eles statistiques *a priori* des deux sources et un apprentissage sur des données d'entraînement. Notre premi`ere contribution est une étude de l'effet d'hétérogénéité des données d'apprentissage sur la performance de séparation. Nous montrons qu'en adaptant le mod`ele de musique `a partir des parties non vocales de la chanson, il est possible d'obtenir de bonnes performances tout en restant dans un contexte réaliste. La seconde contribution de cet article est une étude des interactions entre le crit`ere de performance, le domaine de modélisation des algorithmes et la fonction de coût optimisée.

Le reste de cet article est organisé comme suit. Les méthodes de séparation et les mesures de performance sont décrites dans les sections 2 et 3. Les questions d'adaptation des mod`eles et du choix de la méthode de séparation sont abordées dans la section 4. La section 5 contient la description des données expérimentales et la définition de l'estimateur oracle. Les résultats des expérimentations sont résumés dans la section 6.

2 Méthodes de séparation

Dans cette section, trois méthodes de séparation de sources avec un seul capteur [1, 2, 3] sont présentées. Les deux derni`eres méthodes [2, 3] ont été utilisées `a l'origine pour le débruitage de la parole. Cette tâche peut être présentée sous la forme du probl`eme de séparation de deux sources (la parole et le bruit). Le bruit est souvent considéré stationnaire et donc modélisé par un mod`ele simple. Ici, nous considérons les extensions de ces deux méthodes pour un probl`eme plus complexe, c'est-`a-dire la séparation de deux sources non stationnaires (la voix et la musique).

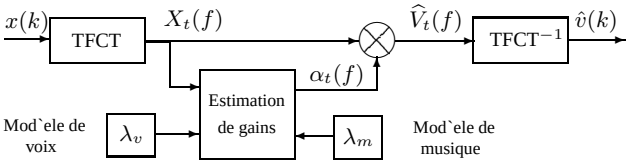


FIG. 1 – Le principe des méthodes de séparation de sources basées sur les mod`eles MMG

La figure 1 illustre le principe de ces trois méthodes de séparation de sources [1, 2, 3]. Soit $X_t(f)$ la Transformée de Fourier `a Court Terme (TFCT) du mélange $x(k)$ pour la trame numéro $t$ et la fréquence $f$. A partir de $X_t(f)$ et de param`etres de mod`eles de voix $\lambda_v$ et de musique $\lambda_m$ (estimés lors d'une phase d'apprentissage), un gain $\alpha_t(f) \geq 0$ est calculé en minimisant un crit`ere d'erreur donné. Ensuite, l'estimation de la TFCT de la voix $\widehat{V}_t(f)$ est obtenue en multipliant $X_t(f)$ par ce gain et la voix estimée $\hat{v}(k)$ est obtenue en appliquant la TFCT

inverse `a l'estimation $\widehat{V}_t(f)$.

Comme crit`ere d'erreur, on consid`ere souvent l'Erreur Quadratique Moyenne (EQM) spectrale :

$$\|\widehat{V} - V\|^2 = \sum_{t,f} |\widehat{V}_t(f) - V_t(f)|^2, \quad (1)$$

On peut aussi être amené `a utiliser l'EQM log spectrale : $\|\log|\widehat{V}| - \log|V|\|^2$.

Benaroya [1] consid`ere les Mod`eles de Mélange de Gaussiennes (MMG) spectraux λ_v^{spec} et λ_m^{spec}. Les spectres `a court terme de la voix $V_t$ sont modélisés comme des variables aléatoires de densité MMG avec des vecteurs moyens nuls et des matrices de covariance $\Sigma_{vi} = \text{diag}\left[\{\sigma_{vi}^2(f)\}_f\right]$, c'est-`a-dire

$$p(V_t|\lambda_v^{\text{spec}}) = \sum_i \omega_{vi} N(V_t; 0, \Sigma_{vi}), \quad (2)$$

o`u $N(V_t; \mu, \Sigma)$ est la densité d'un vecteur gaussien V_t avec le vecteur moyen μ et la matrice de covariance Σ. La diagonale de chaque matrice de covariance $\{\sigma_{vi}^2(f)\}_f$ représente une Densité Spectrale de Puissance (DSP) locale. Ce MMG de voix est noté $\lambda_v^{\text{spec}} = \{\omega_{vi}, \Sigma_{vi}\}_i$. De la même façon, la musique est modélisée par le mod`ele MMG $\lambda_m^{\text{spec}} = \{\omega_{mj}, \Sigma_{mj}\}_j$. Le gain $\alpha_t(f)$ minimisant le crit`ere d'EQM spectrale est calculé comme suit :

$$\alpha_t(f) = \sum_{i,j} \gamma_{i,j}(t) \frac{\sigma_{vi}^2(f)}{\sigma_{vi}^2(f) + \sigma_{mj}^2(f)}, \quad (3)$$

o`u $\gamma_{i,j}(t)$ est la probabilité de choisir la paire d'états $(i, j)$ pour l'observation $X_t$ [1]. Ce gain satisfait $\alpha_t(f) \in [0, 1]$ et l'estimation des sources revient `a effectuer un *filtrage de Wiener pondéré*.

Pour les mêmes mod`eles, Ephraim et Malah [2] développent l'estimateur minimisant le crit`ere d'EQM log spectrale avec le gain

$$\alpha_t(f) = \sum_{i,j} \gamma_{i,j}(t) \frac{\sigma_{vi}^2(f)}{\sigma_{vi}^2(f) + \sigma_{mj}^2(f)} \exp\left[\frac{E_1(\theta_{ij})}{2}\right], \quad (4)$$

o`u $\theta_j = \frac{\sigma_{vi}^2(f)|X_t(f)|^2}{[\sigma_{vi}^2(f)+\sigma_{mj}^2(f)]\sigma_{mj}^2(f)}$ et $E_1(\theta) = \int_\theta^\infty \frac{e^{-t}}{t} dt$ est connue comme l'intégrale exponentielle.

Burshtein et Gannot [3] proposent d'utiliser les mod`eles MMG log spectraux $\lambda_v^{\log}$ et $\lambda_m^{\log}$. Les logarithmes des spectres de la voix $\log|V_t|$ (ici l'opération $\log|\cdot|$ pour un vecteur s'applique élément par élément) sont modélisés par un MMG avec des vecteurs moyens μ_{vi} et des matrices de covariance diagonales Σ_{vi} :

$$p\left(\log|V_t|| \lambda_v^{\log}\right) = \sum_i \omega_{vi} N(\log|V_t|; \mu_{vi}, \Sigma_{vi}), \quad (5)$$

Ce MMG est paramétrisé comme $\lambda_v^{\log} = \{\omega_{vi}, \mu_{vi}, \Sigma_{vi}\}_i$. La musique est modélisée par le MMG $\lambda_m^{\log}$. Avec cette modélisation, les DSP locales sont plutôt représentées par les vecteurs moyens μ_{vi}. Les sources sont estimées en utilisant le crit`ere d'EQM log spectrale et en faisant l'approximation MIXMAX (Mixture Maximum) [4] (voir [3] pour les détails).

L'estimation du mod`ele de chaque source `a partir de données d'apprentissage est basée sur le crit`ere du Maximum de Vraisemblance (MV). En pratique, l'apprentissage utilise l'algorithme EM (Expectation-Maximisation) [5].

3 Mesures de performance

De nombreuses mesures de performance peuvent être utilisées pour évaluer la qualité de séparation. Nous en considérons ici deux : le RSDN et la DLSN.

Pour un enregistrement (une chanson) donné, le RSD Normalisé (RSDN) mesure l'amélioration du Rapport Source `a Distorsion (RSD) [6]

$$\text{RSD}(\hat{v}, v) = 10 \log_{10}\left[\frac{\langle \hat{v}, v\rangle^2}{\|\hat{v}\|^2\|v\|^2 - \langle \hat{v}, v\rangle^2}\right] \quad (6)$$

entre le signal non traité x et la voix estimée $\hat{v}$: $\text{RSDN} = \text{RSD}(\hat{v}, v) - \text{RSD}(x, v)$.

La DLS Normalisée (DLSN) est l'amélioration de la Distorsion du Log Spectre (DLS) [7]

$$\text{DLS}(\hat{v}, v) = \frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{1}{F}\sum_{f=0}^{F-1}\left(10\log_{10}\frac{|V_t(f)|^2 + \epsilon}{|\widehat{\mathcal{V}}_t(f)|^2 + \epsilon}\right)^2\right]^{\frac{1}{2}} \quad (7)$$

entre x et $\hat{v}$: $\text{DLSN} = \text{DLS}(x, v) - \text{DLS}(\hat{v}, v)$, o`u $\widehat{\mathcal{V}}_t(f)$ dans l'équation (7) est la TFCT de la voix estimée $\hat{v}(k)$. Remarquons que, puisque la TFCT est une transformée redondante, *la TFCT d'estimation* $\widehat{\mathcal{V}}_t(f)$ et *l'estimation de la TFCT* $\widehat{V}_t(f)$ ne coïncident pas en général : $\widehat{\mathcal{V}}_t(f) \neq \widehat{V}_t(f)$.

Le choix de la mesure de performance dépend fortement de l'application pour laquelle la séparation est effectuée. Par exemple, si la séparation est faite pour la Reconnaissance Automatique de la Parole (RAP) effectuée sur le signal de la parole séparé, il semble plus judicieux de choisir la DLSN. En effet, les coefficients cepstraux utilisés par la plupart des syst`emes du RAP étant obtenus `a partir des log spectres par une transformation linéaire, il faut vraisemblablement mieux minimiser la distorsion log spectrale.

4 Adaptation des modèles et choix de la méthode de séparation

Les méthodes de séparation de sources existantes donnent des performances satisfaisantes quand les données d'apprentissage sont homog`enes aux données `a séparer, par exemple dans le cas o`u les sources d'apprentissage sont extraites de l'enregistrement musical que l'on souhaite séparer [1]. Les mod`eles appris dans ces conditions sont appelés *modèles adaptés* aux sources. Dans ce travail, nous considérons un cas plus réaliste o`u les données d'apprentissage et les données `a séparer sont hétérog`enes. Nous y étudions d'abord les performances de *modèles généraux* de voix et de musique, c'est-`a-dire de mod`eles appris sur des extraits issus d'autres chansons que celles `a séparer. Enfin, nous nous intéressons `a un cas intermédiaire (*modèles semi-adaptés*) qui peut correspondre `a un cadre d'utilisation réaliste : le mod`ele général de voix est combiné `a un mod`ele de musique adapté par apprentissage sur des segments non vocaux (sans voix chantée) de l'enregistrement `a séparer. Pour cela chaque enregistrement est au préalable segmenté `a la main en parties vocales / non vocales.

Comme précisé section 3, le choix de la mesure de performance dépend de l'application visée. Par ailleurs, la mesure influence le choix d'une méthode particuli`ere. Pour comprendre

cette dépendance, nous testons les méthodes présentées dans la section 2 avec les mesures RSDN et DLSN. Ces méthodes sont caractérisées par le domaine de modélisation (MMG spectral / log spectral) et le crit`ere d'erreur minimisé (EQM spectrale / log spectrale). Nous nous attendons *a priori* `a ce qu'en basculant progressivement du domaine spectral dans le domaine log spectral, le RSDN se dégrade et la DLSN s'améliore.

5 Cadre expérimental

5.1 Description des données.

La base d'apprentissage du mod`ele général de voix contient 34 extraits de voix chantée masculine issus de chansons populaires. Chaque extrait dure approximativement une minute. Le mod`ele général de musique est appris sur 30 extraits de musique populaire sans voix. Chaque extrait dure également environ une minute et tous les extraits proviennent d'auteurs différents. La base d'évaluation contient cinq chansons du même genre pour lesquelles les pistes de voix et de musique sont disponibles séparément, ce qui permet d'évaluer la performance de la séparation en comparant l'estimation `a l'original. Tous les enregistrements sont en mono et échantillonnés `a 11025 Hz.

5.2 Estimateur oracle et limites de performance

Remarquons que l'estimation de la voix $\hat{v}$ ne dépends que du mélange x et de l'ensemble des gains $A = \{\alpha_t(f)\}_{t,f}$ (fig. 1). Il est donc possible d'exprimer cette estimation comme $\hat{v} = g(x, A)$. L'ensemble des gains A appartient `a un ensemble des gains admissibles $\mathcal{A}$ dépendant de la méthode de séparation. Ici, on consid`ere $\mathcal{A}_{[0,1]} = \{A|\alpha_t(f) \in [0, 1]\}$ pour le filtrage de Wiener pondéré [1] et $\mathcal{A}_+ = \{A|\alpha_t(f) \geq 0\}$ pour les autres méthodes [2, 3]. Etant donné une mesure de performance $h(\hat{v}, v, x)$, *l'estimateur oracle* consiste `a trouver l'ensemble des gains $A \in \mathcal{A}$ qui donne la meilleur performance [8] :

$$\tilde{v} = g(x, \tilde{A}), \quad \tilde{A} = \arg\max_{A\in\mathcal{A}} h(g(x, A), v, x) \qquad (8)$$

Cet estimateur nous permet donc de calculer pour un jeu de données la limite de performance qui ne peut pas être dépassée avec la méthode correspondante.

Comme, il est difficile de calculer l'oracle (8) pour les mesures de performance RSDN et DLSN, on calcule `a la place les estimateurs oracles pour le Rapport Signal `a Bruit (RSB) spectral défini comme

$$\text{RSB spec.} = 10\log_{10}\left[\|V\|^2/\|\widehat{V} - V\|^2\right] \qquad (9)$$

Les valeurs de la RSDN et du DLSN (voir (6), (7)) calculés pour ces estimateurs oracles nous indiquent des performances que l'on pourrait atteindre en améliorant l'estimation des gains $\alpha_t(f)$.

6 Expérimentations et résultats

Les probl`emes qui se posent, et que nous allons étudier dans cette section, sont les suivants :

1. choix des param`etres de la TFCT (taille et type de la fenêtre d'analyse)
2. effet du corpus d'apprentissage (hétérogénéité ou homogénéité) et dimensionnement des mod`eles MMG (nombre de Gaussiennes)
3. choix du domaine de modélisation et du crit`ere d'erreur (spectre / log spectre)

6.1 Choix de la fenêtre d'analyse

En utilisant l'estimateur oracle calculé pour $A \in \mathcal{A}_{[0,1]}$, nous avons testé quelle taille et quel type de fenêtre d'analyse dans la TFCT donnent la meilleur performance (fig. 2). Le meilleur résultat est obtenu avec une fenêtre de Hamming de taille 1024 échantillons (soit 93 ms), qui a été utilisée pour le reste des expériences.

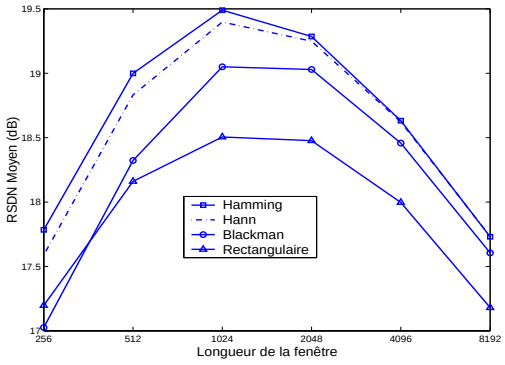


FIG. 2 – Le RSDN pour l'estimateur oracle en fonction de la taille et du type de fenêtre d'analyse.

6.2 Effet de l'hétérogénéité des données

Avec les MMG spectraux et l'estimateur minimisant l'EQM spectrale (3), nous avons testé l'effet sur le RSDN du nombre de Gaussiennes des MMG de voix n_v et de musique n_m dans les trois configurations suivantes :

1. mod`eles généraux de voix et de musique,
2. mod`ele général de voix et mod`ele adapté de musique (appris sur les parties non vocales), c'est-`a-dire mod`eles semi-adaptés,
3. mod`ele adapté de voix (appris sur la voix séparée) et mod`ele adapté de musique.

Les résultats sont résumés sur la figure 3. Dans le cas du filtrage de Wiener ($n_v = n_m = 1$), l'adaptation du mod`ele de musique augmente le RSDN de 2 dB par rapport aux mod`eles généraux. L'adaptation supplémentaire du mod`ele de voix augmente encore la performance de 1.5 dB. Avec deux mod`eles généraux, l'augmentation du nombre total de Gaussiennes $n_v n_m$ n'améliore pas sensiblement la performance par rapport `a $n_v = n_m = 1$, voire la fait lég`erement décroître. D`es que l'on adapte le mod`ele de musique sur les parties non vocales, l'augmentation du nombre $n_v n_m$ de Gaussiennes permet d'améliorer la performance. Avec $n_v = n_m = 128$, le RSDN est de 2.5 dB plus grand que pour le filtrage de Wiener semi-adapté et de 4 dB plus grand que le meilleur résultat avec les deux mod`eles généraux. Les résultats obtenus avec deux mod`eles adaptés (sans doute irréalistes en pratique) nous montrent qu'il reste

encore une marge de 3 `a 4 dB pour améliorer l'apprentissage du mod`ele de voix.

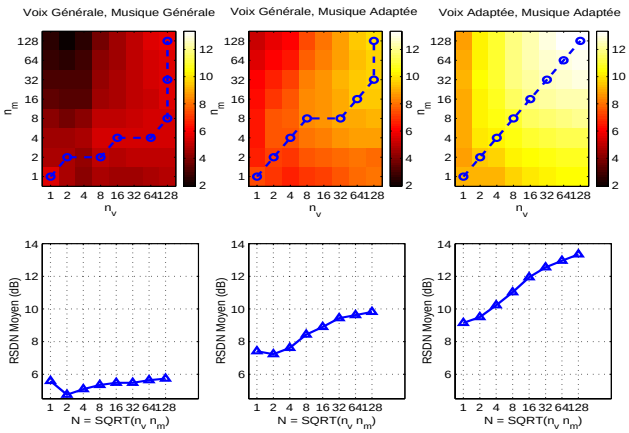


FIG. 3 – En haut : Les RSDN en fonction de (n_v, n_m). La ligne pointillée représente pour un N donné ($N = 1, 2, 4, \ldots, 128$) la paire (n_v^*, n_m^*) qui donne la meilleur performance parmi toutes les paires (n_v, n_m), telles que $n_v n_m = N^2$. En bas : Le RSDN le long de la ligne pointillée. Une référence de performance obtenue avec l'estimateur oracle est de l'ordre 19.5 dB (fig. 2).

6.3 Effets du critère optimisé et du domaine de modélisation

Enfin, nous avons comparé les performances des algorithmes en fonction du domaine de modélisation (MMG spectral / log spectral) et du crit`ere d'erreur minimisé (EQM spectrale / log spectrale), avec des mod`eles semi-adaptés et $n_v = n_m = 64$. Pour chaque paire (mod`ele, crit`ere d'erreur) les deux mesures de performance (le RSDN et le DLSN) sont calculées. Les résultats, accompagnés par des références de performance obtenues avec des estimateurs oracles, sont résumés dans le tableau 1. Comme on s'y attendait (sec. 4), en passant progressivement du domaine spectral dans le domaine log spectral, le RSDN se dégrade. Par contre, la DLSN ne s'améliore pas de façon monotone. En effet, la DLSN est plus mauvaise pour la deuxi`eme méthode (spectral / log spectrale) que pour la premi`ere (spectral / spectrale). Nous avons rajouté dans le tableau 1 les valeurs de la mesure DLSN', calculée en remplaçant la TFCT d'estimation $\widehat{\mathcal{V}}_t(f)$ par l'estimation de la TFCT $\widehat{V}_t(f)$ dans l'equation (7). Cette mesure est ajoutée `a titre informatif, puisque elle ne peut pas être calculée quand la séparation est terminée, car $\widehat{V}_t(f)$ n'est plus accessible (voir la fig. 1). Pour la DLSN', cette amélioration monotone est vérifiée, vraisemblablement parce que la DLSN' est plus cohérente avec le crit`ere d'EQM log spectrale que la DLSN. Le meilleur RSDN est toujours obtenu pour la premi`ere méthode (spectral / spectrale) et la meilleure DLSN pour la troisi`eme (log spec. / log spec.).

7 Conclusions et perspectives

Nous proposons une procédure d'adaptation du mod`ele de musique permettant de rester dans un cadre d'utilisation réaliste,

TAB. 1 – Les performances des méthodes (mod`ele MMG / le crit`ere d'EQM minimisé). Les références de performance obtenues `a l'aide des oracles sont indiquées dans les parenth`eses. La DLSN' est ajoutée `a titre informatif.

MMG / EQM	RSDN	DLSN	DLSN'
spectral / spectrale [1]	9.1 (19.5)	4.5 (13.0)	*1.5*
spectral / log spec [2]	8.4 (20.1)	3.5 (13.0)	*3.3*
log spec / log spec [3]	6.4 (20.1)	5.4 (13.0)	*4.9*

puisque la segmentation manuelle en parties vocales / non vocales peut être effectuée par un utilisateur. Dans le cadre de notre étude, nous avons montré que cette adaptation permet d'améliorer considérablement la performance de séparation par rapport au cas de deux mod`eles généraux. Dans l'avenir, nous comptons remplacer la segmentation manuelle en parties vocales / non vocales par un module de segmentation automatique [9].

Les effets du crit`ere optimisé et du domaine de modélisation ont été étudiés en utilisant deux mesures de performance différentes. Cet étude donne des indication sur le choix d'une méthode de séparation en fonction de la mesure de performance, qui dépend en même temps de la tâche pour laquelle la séparation est effectuée.

Références

[1] L. Benaroya, "Séparation de plusieurs sources sonores avec un seul microphone," Ph.D. dissertation, Université de Rennes 1, 2003.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean square error log-spectral amplitude estimator," in *IEEE Trans. on Acoust., Speech, and Sig. Proc.*, vol. ASSP-33, Apr 1985, pp. 443–445.

[3] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," in *EUROSPEECH*, vol. 6, Budapest, Hungary, Sep 1999, pp. 2591–2594.

[4] A. Nádas, D. Nahamoo and M. A. Picheny, "Speech recognition using noise-adaptive prototype," in *IEEE Trans. on Speech and Audio Proc.*, 1989, pp. 1495–1505.

[5] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, Feb 1989.

[6] R. Gribonval, L. Benaroya, E. Vincent and C. Févotte, "Proposals for performance measurement in source separation," in *ICA*, Apr 2003, pp. 763–768.

[7] Valin J.-M., Rouat J. and Michaud F., "Microphone array post-filter for separation of simultaneous non-stationary sources," in *ICASSP*, 2004.

[8] E. Vincent et R. Gribonval, "Construction d'estimateurs oracles pour la séparation de sources," in *GRETSI*, 2005, (`a paraître).

[9] Wei-Ho Tsai, Dwight Rogers and Hsin-Min Wang, "Blind clustering of popular music recordings based on singer voice characteristics," *Computer Music Journal*, 2004.

Séparation aveugle sous-déterminée de sources audio par la méthode EMD (Empirical Mode Decomposition)

Abdeldjalil AISSA-EL-BEY, Karim ABED-MERAIM, Yves GRENIER

Département Traitement du Signal et des Images,
ENST-Paris, 46 rue Barrault 75634, Paris Cedex 13, France
{elbey, abed, grenier}@tsi.enst.fr

Résumé – Dans le cadre de la séparation aveugle de sources, nous montrons dans cet article comment effectuer la séparation de mélanges instantanés de sources audio en utilisant une méthode basée sur l'algorithme de Décomposition Modale Empirique (ou EMD, pour Empirical Mode Decomposition). Cette approche nous permet, en particulier de traiter le cas sous déterminé (c'est à dire le cas où l'on a moins de capteurs que de sources). L'approche EMD se base sur le fait que les signaux audio (et particulièrement les signaux musicaux) peuvent être bien modélisés localement par une somme de signaux périodiques. Ces signaux seront donc décomposés en utilisant l'algorithme EMD et recombinés par classification suivant leurs directions spatiales regroupant ainsi les composantes de chacune des sources. Nous présenterons quelques résultats de simulation qui permettent d'évaluer les performances de cette nouvelle méthode.

Abstract – This paper introduces new algorithm for the blind separation of audio sources using Empirical Mode Decomposition (EMD). Indeed, audio signals and, in particular, musical signals can be well approximated by a sum of damped sinusoidal (modal) components. Based on this representation, we propose a two steps approach consisting of a signal analysis (extraction of the modal components) using EMD followed by a signal synthesis (pairing of the components belonging to the same source) using vector clustering. For the signal analysis, a major advantage of the proposed method resides in its ability to separate more sources than sensors. Simulation results are given to assess the performance of the proposed algorithm.

1 Introduction

La séparation aveugle de sources est un problème qui consiste à retrouver des signaux statistiquement indépendants à partir de leurs mélanges (observations) et cela sans connaissance *a priori* de la structure des mélanges ou des signaux sources.
La séparation de sources intervient dans des applications diverses [1] telles que la localisation et la poursuite de cibles en radar et sonar, la séparation de locuteurs (problème dit de "cocktail party"), la détection et la séparation dans les systèmes de communication à accès multiple, l'analyse en composantes indépendantes de signaux biomédicaux (e.g., EEG ou ECG), etc. Ce problème a été intensément étudié dans la littérature et beaucoup de solutions efficaces ont déjà été proposées [1].
Néanmoins, le cas sous-déterminé où le nombre de sources est supérieur à celui des capteurs (observations) reste relativement peu traité et sa résolution est l'un des problèmes ouverts de la séparation aveugle de sources. Dans le cas de signaux non-stationnaires (incluant en particulier les signaux audio), certaines solutions utilisant la transformée temps-fréquence des observations existent pour le cas sous-déterminé [4, 6]. Nous proposons ici une approche alternative utilisant la décomposition modale empirique EMD (Empirical Mode Decomposition) des signaux observés [2, 3]. Cette technique, récemment proposée, permet la décomposition des différents modes d'un signal supposé *localement périodique*, sans qu'il soit nécessairement harmonique au sens de Fourier. Les signaux audio et plus particulièrement les signaux musicaux peuvent être modélisés par une somme de sinusoïdes amorties [7]. C'est cette propriété que nous proposons d'exploiter ici pour la séparation de sources audio par le biais de la décomposition EMD.

2 Formalisation du problème

Le modèle de séparation aveugle de sources suppose l'existence de N signaux indépendants $s_1(t), \ldots, s_N(t)$ et M observations $x_1(t), \ldots, x_M(t)$ qui représentent les mélanges. Ces mélanges sont supposés instantanés, i.e.

$$x_i(t) = \sum_{j=1}^{N} a_{ij} s_j(t) \qquad i = 1, \ldots, M \tag{1}$$

Ceci peut être représenté par l'équation de mélange :

$$\mathbf{x}(t) = \mathbf{A}\mathbf{s}(t) \tag{2}$$

où $\mathbf{s}(t) \stackrel{\text{def}}{=} [s_1(t), \ldots, s_N(t)]^T$ est le vecteur colonne de dimension $N \times 1$ qui regroupe les signaux sources, le vecteur $\mathbf{x}(t)$ regroupe de la même manière les M signaux observés, et $\mathbf{A} \stackrel{\text{def}}{=} [\mathbf{a}_1, \ldots, \mathbf{a}_N]$ est la matrice de mélange de taille $M \times N$ où $\mathbf{a}_i = [a_{1i}, \ldots, a_{Mi}]^T$ contient les coefficients du mélange. Nous supposerons que pour tout couple (i, j) les vecteurs $\mathbf{a}_i$ et $\mathbf{a}_j$ sont linéairement indépendants. Les signaux sources sont supposés décomposables en somme de composantes modales $c_i^j(t)$, i.e :

$$s_i(t) = \sum_{j=1}^{l_i} c_i^j(t) \qquad t = 0, \ldots, T-1 \tag{3}$$

L'hypothèse d'indépendance des sources est remplacée ici par l'hypothèse de quasi-orthogonalité des composantes modales, i.e.

$$\frac{\langle c_i^j | c_{i'}^{j'} \rangle}{\|c_i^j\| \|c_{i'}^{j'}\|} \approx 0 \quad \text{pour} \quad (i,j) \neq (i',j') \tag{4}$$

où

$$\langle c_i^j | c_{i'}^{j'} \rangle \stackrel{\text{def}}{=} \sum_{t=0}^{T-1} c_i^j(t) c_{i'}^{j'}(t)^* \quad (5)$$

et

$$\|c_i^j\|^2 = \langle c_i^j | c_i^j \rangle \quad (6)$$

L'hypothèse (4) peut être parfois restrictive. Ainsi nous proposons dans la section 3.4 une méthode permettant de s'affranchir de cette contrainte.

3 La séparation de sources utilisant l'approche EMD

En se basant sur le modèle précédent, on propose une approche en deux étapes qui sont :

– *Étape d'analyse* : Dans cette étape on applique une décomposition modale sur chaque sortie de capteur afin d'en extraire toutes les composantes harmoniques ou pseudo-harmoniques. Cette décomposition consiste en l'algorithme EMD introduit en [2, 3].

– *Étape de synthèse* : Dans cette étape nous regroupons ensemble les composantes modales correspondant au même signal source afin de reconstituer le signal d'origine. Ceci est fait par une méthode de classification basée sur la direction spatiale des composantes que l'on estime par corrélation de celles-ci avec le signal d'antenne observé.

3.1 Analyse des signaux utilisant l'EMD

Une nouvelle technique non-linéaire, appelée EMD a été récemment introduite par N.E. Huang et al. pour représenter les signaux non-stationnaires [2]. Dans l'algorithme EMD le signal non stationnaire est considéré à l'échelle de ses oscillations locales. Plus précisément, la décomposition d'un signal $z(t)$ par l'EMD se résume comme suit :

1. Identification de tous les extrema de $z(t)$.
2. Interpolation entre les minima (resp. maxima) du signal, conduisant à une enveloppe $e_{min}(t)$ (resp. $e_{max}(t)$).
3. Calcul de la moyenne $m(t) = (e_{min}(t) + e_{max}(t))/2$.
4. Extraction du détail $d(t) = z(t) - m(t)$.
5. Itération sur le résidu $m(t)$ jusqu'à l'obtention d'un résidu final d'énergie quasi-nulle.

En appliquant alors l'algorithme EMD sur les x_i qui s'écrivent $x_i(t) = \sum_{j=1}^{N} a_{ij} s_j(t) = \sum_{j=1}^{N} \sum_{k=1}^{l_j} a_{ij} c_j^k(t)$, on obtient des estimées $\widehat{c}_j^k(t)$ des composantes $c_j^k(t)$.

3.2 Synthèse des signaux utilisant la classification vectorielle

Pour la synthèse des signaux sources on observe que l'hypothèse de quasi-orthogonalité nous permet d'avoir :

$$\frac{\langle \mathbf{x} | c_i^j \rangle}{\|c_i^j\|^2} \stackrel{\text{def}}{=} \frac{1}{\|c_i^j\|^2} \begin{bmatrix} \langle x_1 | c_i^j \rangle \\ \vdots \\ \langle x_M | c_i^j \rangle \end{bmatrix} \approx \mathbf{a}_i \quad (7)$$

où $\mathbf{a}_i$ représente la i^{eme} colonne de $\mathbf{A}$. nous pouvons alors associer chaque composante $\widehat{c}_j^k$ à une direction spatiale (vecteur colonne de $\mathbf{A}$) que l'on estime par

$$\widehat{\mathbf{a}}_j^k = \frac{\langle \mathbf{x} | \widehat{c}_j^k \rangle}{\|\widehat{c}_j^k\|^2}$$

Deux composantes du même signal source étant associées au même vecteur colonne de $\mathbf{A}$, nous proposons de regrouper ces composantes par classification sur les vecteurs $\widehat{\mathbf{a}}_j^k$ selon N classes. Pour cela, nous avons utilisé l'algorithme k-means [9] qui se résume comme suit :

1. Initialement, sélectionner N centroïdes arbitrairement dans l'ensemble des vecteurs.
2. Assigner chaque vecteur à la classe dont le centroïde est le plus proche au sens de la distance Euclidienne.
3. Calculer les centroïdes de chaque classe.
4. Répéter l'étape 2 et 3 jusqu'à ce qu'il n'y ait aucun changement de vecteurs entre les classes.

Finalement, on pourra reconstruire les sources initiales à une constante près en additionnant les différentes composantes d'une même classe.

3.3 Association et sélection

Remarquons, qu'en appliquant l'approche décrite précédemment (analyse plus synthèse) sur toutes les sorties d'antenne $x_1(t), \cdots, x_M(t)$, nous obtenons M estimées de chacune des sources. Nous avons observé que la qualité d'estimation des sources varie d'un capteur à un autre, et ceci dépend fortement des coefficients du mélange, en particulier, du rapport signal à interférence (RSI) de la source désirée. En conséquence, nous proposons une méthode aveugle de sélection pour choisir la meilleure des M estimées d'une source donnée. Pour cet effet, nous avons besoin tout d'abord de regrouper les estimées d'une même source ensemble. Ceci est réalisé par corrélation i.e. un signal est associé aux $(M-1)$ signaux (les signaux issus des $(M-1)$ autres capteurs) qui lui sont les plus corrélés. Le facteur de corrélation de deux signaux s_1 et s_2 est calculé par $\frac{\langle s_1 | s_2 \rangle}{\|s_1\| \|s_2\|}$. Une fois l'association des sources effectuée, on propose de sélectionner la source estimée qui a l'énergie maximale, i.e.

$$\widehat{s}_i(t) = \max_j \{E_i^j = \|\widehat{s}_i^j(t)\|^2, \quad j = 1, \cdots, M\} \quad (8)$$

où E_i^j représente l'énergie de la i^{eme} source obtenue à partir du j^{eme} capteur.

3.4 Projection sous-espace

Afin de relâcher la contrainte (4), on suppose qu'une composante $c_j^k(t)$ peut être présente dans plusieurs sources. Ceci est le cas pour certains signaux musicaux tels que ceux traités dans [5]. Pour simplifier, nous supposons ici qu'une composante appartient au plus à deux sources. Supposons donc que la composante $c_j^k(t)$ est présente dans les sources $s_{j_1}(t)$ et $s_{j_2}(t)$ avec les amplitudes α_{j_1} et α_{j_2} respectivement. Il s'en suit que la direction spatiale associée à cette composante estimée par (7) est donnée par :

$$\widehat{\mathbf{a}}_j^k \approx \alpha_{j_1} \mathbf{a}_{j_1} + \alpha_{j_2} \mathbf{a}_{j_2}. \quad (9)$$

Il s'agit maintenant de trouver les indices j_1 et j_2 des deux sources associées à cette composante, ainsi que les amplitudes α_{j_1} et α_{j_2}. Pour ce faire, on propose une approche basée sur la projection en sous-espace. Supposons que la matrice de mélange $\mathbf{A}$ est connue et vérifie les conditions $M > 2$ et tout triplet de vecteurs colonnes de $\mathbf{A}$ sont linéairement indépendants. Observons alors :

$$\mathbf{P}^{\perp}_{\widetilde{\mathbf{A}}}\mathbf{a}^k_j = 0$$

si est seulement si $\widetilde{\mathbf{A}} = [\mathbf{a}_{j_1}\ \mathbf{a}_{j_2}]$ où $\mathbf{P}^{\perp}_{\widetilde{\mathbf{A}}}$ représente la matrice de projection orthogonale sur l'orthogonal du sous-espace image de $\widetilde{\mathbf{A}}$, i.e.

$$\mathbf{P}^{\perp}_{\widetilde{\mathbf{A}}} = \mathbf{I} - \widetilde{\mathbf{A}}\left(\widetilde{\mathbf{A}}^H\widetilde{\mathbf{A}}\right)^{-1}\widetilde{\mathbf{A}}^H. \qquad (10$$

En pratique, en prenant en compte le bruit, on détecte les colonnes j_1 et j_2 en minimisant :

$$(j_1, j_2) = \arg\min_{(l,m)}\left\{\|\mathbf{P}^{\perp}_{\widetilde{\mathbf{A}}}\widehat{\mathbf{a}}^k_j\| \mid \widetilde{\mathbf{A}} = [\mathbf{a}_l\ \mathbf{a}_m]\right\}$$

Une fois $\widetilde{\mathbf{A}}$ trouvée, on estime les pondérations α_{j_1} et α_{j_2} par

$$\begin{bmatrix}\alpha_{j_1}\\ \alpha_{j_2}\end{bmatrix} = \widetilde{\mathbf{A}}^{\#}\widehat{\mathbf{a}}^k_j \qquad (11$$

où $\widetilde{\mathbf{A}}^{\#}$ représente la pseudo-inverse de $\widetilde{\mathbf{A}}$. Dans cette article nous avons traité toutes les composantes comme étant associées à deux signaux sources. Si jamais une composante n'est présente que dans une seule source, un des deux coefficients estimés dans (11) devrait être zéro ou proche de zéro.

Dans ce qui précède la matrice de mélange $\mathbf{A}$ est supposée connue, il faut donc l'estimer avant d'appliquer la projection par sous-espace. Nous nous proposons d'estimer les colonnes de $\mathbf{A}$ comme étant les centroïdes moyennés (i.e. comme chaque source est estimée M fois, et que chaque sources correspond à une classe et à un centroïde, alors, on moyenne les centroïdes des M classes d'une source donnée) des N classes obtenues dans l'étape de synthèse de notre algorithme. Cette approche suppose implicitement que la majorité des composantes n'appartiennent qu'à une seule source et par conséquent que les vecteurs $\widehat{\mathbf{a}}^k_j$ représentent dans leurs majorité une des colonnes de la matrice $\mathbf{A}$.

Remarques :

- *Cas sur-déterminé* : Dans ce cas, on peut effectuer la séparation par une inversion de la matrice de mélange $\mathbf{A}$. Cette dernière est estimée par l'approche décrite précédemment.
- *Estimation du nombre de sources* : C'est une tâche difficile dans le cas sous-déterminé. Il existe certaines solutions utilisant soit des approches tensorielles [8] soit d'autres utilisant des techniques de classifications avec estimation conjointe du nombre de classes [9]. Cependant, ces méthodes sont très sensibles au bruit, à la dynamique des signaux sources et au conditionnement de la matrice de mélange. Dans cet article on suppose que le nombre de sources est connu (ou correctement estimé).

4 Simulations

Nous présentons ici quelques résultats de simulation pour illustrer l'exécution de notre algorithme de séparation de sources aveugle. Pour cela, nous considérons une antenne linéaire uniforme (ALU) de capteurs avec $M = 3$ capteurs recevant $N = 4$ signaux sources audio (à l'exception de la troisième expérience où N varie dans l'interval [2,6]). Selon le modèle ALU, la matrice $\mathbf{A}$ est engendrée avec des angles d'arrivés des sources aléatoires. La taille des observations $T = 10000$ (les signaux sources sont échantillonnés à une fréquence de 44.1 khz). Les signaux observés sont corrompus par un bruit blanc additif de covariance $\sigma^2\mathbf{I}$ (σ^2 étant la puissance de bruit). La qualité de séparation est mesurée par l'erreur quadratique moyenne normalisée (EQMN) des sources estimées pour 100 tirages aléatoires du bruit.

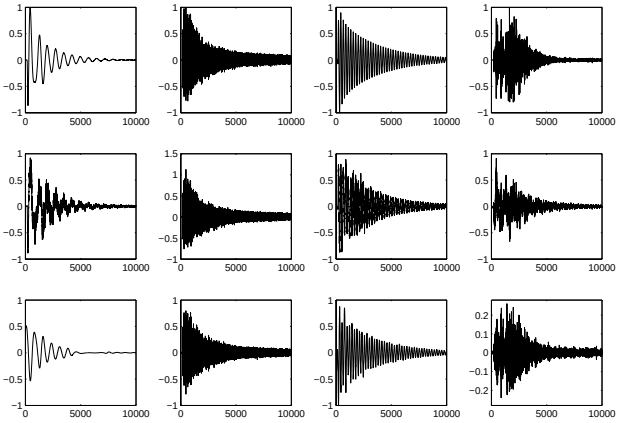


FIG. 1 – Exemple de séparation de 4 signaux audio pour 3 capteurs : signaux originaux (ligne du haut) ; signaux estimés par pseudo-inversion quand la matrice de mélange est connue (ligne du milieu) ; signaux estimés par EMD (ligne du bas).

La figure 1 représente les résultats obtenus en utilisant notre algorithme pour les 4 sources audio représentées par la première ligne de la figure. La deuxième ligne montre le résultat obtenu en utilisant la pseudo inverse de la matrice de mélange $\mathbf{A}$ pour estimer les sources originales en connaissant $\mathbf{A}$ avec exactitude, et la dernière donne les estimées des sources en utilisant notre algorithme. La figure 2 représente la variation de l'erreur quadratique moyenne normalisée (EQMN) des estimées des signaux sources en fonction du rapport signal à bruit (RSB). Sur cette figure nous comparons les résultats obtenus par une sélection basée sur un critère énergétique (qui consiste à sélectionner l'estimée la plus énergétique parmi les M estimées da la même source), une sélection optimale au sens de l'EQMN (celle-ci est donnée seulement à titre comparatif car elle suppose la connaissance des signaux sources) et les résultats obtenus par pseudo-inversion de la matrice $\mathbf{A}$. La figure 3 représente la variation de l'erreur quadratique moyenne normalisée des estimées des signaux sources en fonction du nombre de sources qui varie de 2 à 6. Dans les cas ou $N = 2$ et $N = 3$ (cas sur-déterminé) la séparation est effectuée par pseudo-inversion de l'estimée de la matrice de mélange $\mathbf{A}$. La figure 4 représente la variation de l'EQMN des estimées des signaux sources en fonction du RSB. Sur cette figure nous comparons les résultats obtenus par l'utilisation de la décomposition EMD uniquement et celles obtenues avec l'utilisation de

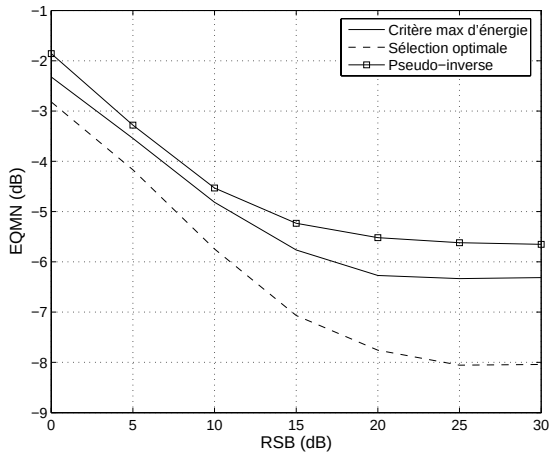


FIG. 2 – Performances de l'algorithme de séparation de 4 sources audio pour 3 capteurs : les courbes représentent la valeur moyenne de l'erreur quadratique de l'estimation des signaux sources en fonction du rapport signal à bruit.

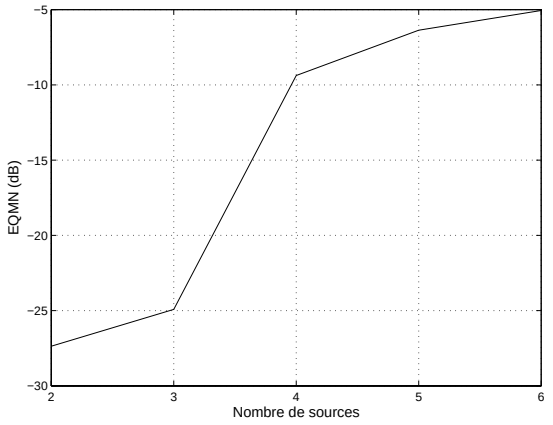


FIG. 3 – Performance de l'algorithme de séparation pour 3 capteurs : la courbe représente la valeur moyenne de l'erreur quadratique de l'estimation des signaux sources en fonction de leur nombre.

la décomposition EMD conjuguée à la projection sous-espace. On peut observer un gain à moyen et fort RSB mais une légère dégradation à faible RSB pour la méthode EMD avec projection sous-espace. Cette dégradation est due au fait que l'on associe systématiquement toute composante à deux sources est que l'estimation des amplitudes par (11) est fortement bruitée à faible RSB. Ainsi, si une composante n'appartient effectivement qu'à une seule source s_{j_1} le coefficient de la deuxième source α_{j_2} devrait être nul. En présence de bruit l'équation (11) donne une valeur non nulle de ce coefficient qui dégrade l'estimation des sources.

5 Conclusion

Dans cet article nous présentons une nouvelle méthode de séparation aveugle de sources audio utilisant l'algorithme EMD.

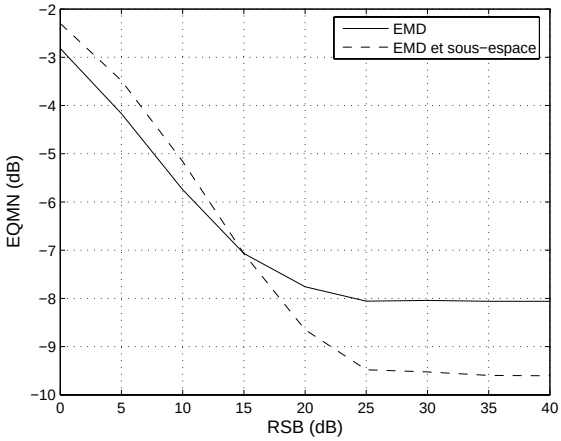


FIG. 4 – Performances des l'algorithmes de séparation de 4 sources audio pour 3 capteurs : les courbes représentent la valeur moyenne de l'erreur quadratique de l'estimation des signaux sources en fonction du rapport signal à bruit pour l'algorithme EMD classique ainsi que celle pour l'EMD avec projection sous-espace.

L'avantage majeur de la méthode proposée est sa capacité à séparer plus de sources que de capteurs, et dans ce cas elle produit une meilleure qualité de séparation que ce que l'on obtient par pseudo-inversion de la matrice de mélange (même si elle est connue).

Références

[1] A.K. Nandi (editor), "Blind estimation using higher-order statistics." *Kluwer Acadimic Publishers*, Boston 1999.

[2] N.E Huang, Z. Shen, S.R. Long, M.L. Wu, H.H. Shih, Q. Zheng, N.C. Yen, C.C. Trung and H.H. Liu, "The empirical mode decomposition and Hilbert spectrum for nonlinear and non-stationary times series analysis", *Proc. Roy. Soc.* London A, Vol. 454, pp. 903-995, 1998.

[3] P. Flandrin, G. Rilling and P. Goncalvès, "Empirical mode decomposition as a filter bank", *IEEE SPL*, 2004.

[4] L. Nguyen, A. Belouchrani, K. Abed-Meraim and B. Boashash, "Separating more sources than sensors using time-frequency distributions." in *Proc. ISSPA*, Malysia, 2001.

[5] J. Rosier, Y. Grenier, "Unsupervised Classification Techniques for Multipitch Estimation", in *116th Convention of the Audio Engineering Society*, Berlin, 2004.

[6] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals : demixing n sources from 2 mixtures," in *ICASSP*, Turkey, 2000.

[7] R. Boyer and K. Abed-Meraim, "Audio modeling based on delayed sinusoids." *IEEE-Tr-SAP*, Mars 2004.

[8] L. De Lathauwer, B. Moor, J. Vandewalle, "ICA techniques for more sources than sensors", *Higher-oder statistic Proc. of the IEEE Sig. Proc. Workshop*, 1999.

[9] I.E. Frank and R. Todeschini, "The data analysis handbook", *Elsevier, Sci. Pub. Co.*, 1994.

Séparation de sources dans le cas sous-déterminé : comparaison de deux approches basées sur des décompositions parcimonieuses

Sylvain LESAGE, Sacha KRSTULOVIC, Rémi GRIBONVAL

Projet METISS, IRISA-INRIA
Campus de Beaulieu, 35042 Rennes Cedex, France
sylvain.lesage@irisa.fr, sacha.krstulovic@irisa.fr, remi.gribonval@irisa.fr

Résumé – Cet article traite de la séparation de sources dans le cas sous-déterminé quand la matrice de mélange est connue. On se place dans le cadre des approches basées sur la décomposition parcimonieuse du mélange. Dans la nouvelle méthode proposée, on décompose le mélange par Matching Pursuit en introduisant une nouvelle classe de dictionnaires multi-canal, où les atomes sont caractérisés par une direction spatiale et une forme d'onde. La connaissance de la matrice de mélange est directement intégrée à la décomposition. Par rapport à la séparation par Matching Pursuit multi-canal suivie d'un clustering, le nouvel algorithme introduit moins d'artéfacts pour un niveau d'interférences résiduelles comparable. Nous comparons ces deux méthodes à l'algorithme de Bofill-Zibulevski et à DUET. Nous étudions également l'apport d'un lissage des décompositions et l'importance de la qualité de l'estimation de la matrice de mélange.

Abstract – This paper focuses on underdetermined source separation with a known mixing matrix. The approach is based on a sparse decomposition of the mixture. In the proposed method, the mixture is decomposed with Matching Pursuit by introducing a new class of multichannel dictionnaries, where the atoms are given by a spatial direction and a waveform. The knowledge of the mixing matrix is directly integrated in the decomposition. Compared to the separation by multichannel Matching Pursuit followed by a clustering, the new algorithm introduces less artifacts whereas the level of residual interferences is about the same. These two methods are compared to Bofill-Zibulevsky's algorithm and DUET. We also study the effect of smoothing the decompositions and the importance of the quality of the estimation of the mixing matrix.

1 Introduction

Le probl`eme de la séparation de sources [1] consiste `a retrouver des signaux inconnus (les sources) `a partir de la seule connaissance de mélanges de ces signaux (les canaux). Chaque canal x_n est la somme des sources fi ltrées :

$$x_n(t) = \sum_{i=1}^{I} (a_{n,i} * s_i)(t)$$

avec $a_{n,i}$ des fi ltres. Dans le cas d'un mélange linéaire instantané, les fi ltres correspondent simplement `a une multiplication par une constante, si bien qu'en notation matricielle le mélange s'écrit $\mathbf{x} = \mathbf{A}\mathbf{s}$, o`u $\mathbf{A}$ est la matrice de mélange et les matrices $\mathbf{x}$ et $\mathbf{s}$ ont pour lignes les signaux x_n et s_i. Dans le cas déterminé (resp. surdéterminé), lorsque le nombre de canaux observés est égal (resp. supérieur) `a celui des sources, il est équivalent d'estimer la matrice de mélange et d'estimer les sources. Dans le cas sous-déterminé, au contraire, la connaissance ou l'estimation de la matrice de mélange n'est pas suffi sante pour retrouver les sources et il faut généralement recourir `a un mod`ele des sources pour les estimer [2]. En général, dans les performances d'un algorithme donné, il n'est pas facile de distinguer l'effet de la qualité de l'estimation de la matrice de mélange de celui de l'inadéquation du mod`ele choisi.

Dans cet article, nous nous intéressons au cas sous-déterminé. Notre approche emploie des mod`eles reposant sur l'existence de représentations parcimonieuses des sources [3], et suppose que la matrice de mélange est parfaitement connue. Nous comparons deux algorithmes de séparation basés sur des variantes de Matching Pursuit (MP) [4]. La premi`ere variante consiste `a décomposer le mélange multi-canal sans connaissance de la matrice de mélange, puis `a utiliser la matrice de mélange pour classifi er les composantes obtenues et les affecter aux sources `a estimer [5, 6]. La seconde variante consiste `a utiliser la matrice de mélange d`es l'étape de décomposition parcimonieuse, sans qu'aucune étape de classifi cation ultérieure ne soit nécessaire. Les performances de ces deux algorithmes sont comparées au meilleur séparateur linéaire (MSL) [7], `a l'algorithme de Bofi ll-Zibulevski (BZ) [8] et `a l'algorithme DUET [6].

Cet article est organisé comme suit : dans la section 2, on rappelle la défi nition générale du Matching Pursuit. Les algorithmes de séparation qui en découlent sont décrits dans la section 3, puis nous détaillons les conditions expérimentales et les résultats obtenus dans la section 4.

2 Matching Pursuit

Un signal x (considéré comme un vecteur de l'espace de Hilbert $\mathcal{H}$ des signaux `a énergie fi nie) admet une décomposition parcimonieuse dans le dictionnaire $\mathcal{D} = \{g_k\}$ d'atomes $g_k$ – ou signaux élémentaires $g_k$ – s'il peut s'écrire comme une combinaison linéaire $x = \sum_k c_k g_k$ o`u peu de coeffi cients $\{c_k\}$ sont non négligeables. L'algorithme Matching Pursuit calcule itérativement des approximations parcimonieuses de la forme $x = \sum_{m=1}^{M} c_{k_m} g_{k_m} + R^M$ o`u $R^M$ est un résidu qui tend vers zéro lorsque le nombre d'itérations $M$ tend vers l'infi ni. Son principe est de sélectionner `a chaque étape l'atome le plus corrélé avec le résidu, puis de mettre `a jour le résidu en ôtant la contribution de cet atome.

Les crit`eres d'arrêt les plus courants sont basés sur le niveau absolu ou relatif de l'énergie du résidu et/ou sur un nombre fi xe d'itérations `a effectuer. Par ailleurs, le dictionnaire de Gabor est

classiquement utilisé pour décomposer de façon parcimonieuse les signaux audio. Il représente une collection d'atomes temps-fréquence de Gabor de la forme :

$$g_{s,u,\xi}(t) = w\left(\frac{t-u}{s}\right) \cdot \exp\left(2j\pi\xi(t-u)\right).$$

Ces atomes sont défi nis par le choix d'une fenêtre w d'énergie unitaire (Hanning, Gaussienne, ...), d'un facteur d'échelle s, d'une localisation temporelle u, et d'une fréquence ξ. Un tel dictionnaire permet un calcul rapide des produits scalaires d'un signal avec les atomes par application de FFT fenêtrées.

3 Séparation par Matching Pursuit

Pour résoudre des probl`emes de séparation de sources dans le cas multi-canal, une méthode `a base de décompositions parcimonieuses sur un dictionnaire a été proposée [5, 9]. Elle utilise le Matching Pursuit multi-canal, suivi d'un clustering. Apr`es avoir rappelé le principe de cette méthode, nous proposons une variante o`u la défi nition du dictionnaire inclut des connaissances sur la matrice de mélange $\mathbf{A}$.

3.1 Matching Pursuit multi-canal

Pour la décomposition parcimonieuse de signaux multi-canal, on utilise un dictionnaire $\boldsymbol{\mathcal{D}}$ comportant des atomes multi-canal $\mathbf{g}$, défi nis par $\mathbf{g} = (c_1 g, c_2 g, \ldots, c_N g)$, o`u $g \in \mathcal{D}$ est un atome mono-canal d'un dictionnaire $\mathcal{D}$ et o`u les coeffi cients $c_1, \ldots, c_N$ satisfont $\sum_{n=1}^{N} c_n^2 = 1$. Le Matching Pursuit multi-canal conduit `a une décomposition du type :

$$(x_1, \ldots, x_N) = \widehat{\mathbf{x}}^M + (R_1^M, \ldots, R_N^M).$$

avec $\widehat{\mathbf{x}}^M := \sum_{m=1}^{M} (c_{1,k_m} g_{k_m}, \ldots, c_{N,k_m} g_{k_m})$. L'algorithme se déroule de la façon suivante :

1. Initialisation : $M = 1, \quad R_n^0 = x_n, \quad c_{n,k} = 0, \quad \forall n, \forall k$;
2. Calcul du produit scalaire entre chaque canal du résidu R_n^{M-1} et chaque atome g_k du dictionnaire mono-canal.
3. Sélection de $k_M = \arg\max_k \sum_{n=1}^{N} |\langle R_n^{M-1}, g_k \rangle|^2$
4. Pour chaque canal n, mise `a jour du résidu : $R_n^M = R_n^{M-1} - \langle R_n^{M-1}, g_{k_M} \rangle g_{k_M}$ et des coeffi cients : $c_{n,k_M}^M = c_{n,k_M}^{M-1} + \langle R_n^{M-1}, g_{k_M} \rangle$
5. Si le crit`ere d'arrêt n'est pas atteint, $M \leftarrow M + 1$, puis retour `a 2

A partir du signal $\widehat{\mathbf{x}}^M$, approché par Matching Pursuit multi-canal, chaque source s_i est estimée en utilisant les atomes de la décomposition qui lui sont attribués de la mani`ere suivante : la matrice de mélange $\mathbf{A}$ étant supposée connue et ses colonnes $\mathbf{a}_i = (a_{n,i})_{1 \le n \le N}$ normées ($\|\mathbf{a}_i\|_2 = 1$), on attribue l'atome $k_M$ `a la source d'indice

$$\hat{i}_M = \arg\max_i |\langle \mathbf{c}_{k_M}, \mathbf{a}_i \rangle|.$$

Cela revient `a partitionner l'espace des coeffi cients $\mathbf{c} = (c_n)_{1 \le n \le N}$ en I sous-ensembles correspondants aux colonnes $\mathbf{a}_i$ de $\mathbf{A}$ (I étant le nombre de sources). La source s_i est alors reconstituée par :

$$\hat{s}_i = \sum_{M \text{ tq } \hat{i}_M = i} \langle \mathbf{c}_{k_M}, \mathbf{a}_i \rangle g_{k_M}. \qquad (1)$$

On nomme MPC1 cet algorithme, et MPC2 une variante qui consiste `a attribuer chaque atome aux deux sources qui l'entourent. Cette seconde sélection, également utilisée dans l'algorithme BZ [8], correspond `a la minimisation de la norme l_1 de la projection des coeffi cients $\mathbf{c}_{k_M}$ sur deux directions de la matrice de mélange :

$$(\hat{i}_M, \hat{j}_M) = \arg\min_{(i,j)} \left\| [\mathbf{a}_i \ \mathbf{a}_j]^{-1} \mathbf{c}_{k_M} \right\|_1.$$

3.2 Dictionnaire multi-canal directionnel

En combinant l'expression des mélanges $x_n = \sum_{i=1}^{I} a_{n,i} \cdot s_i$ et celle d'une possible décomposition parcimonieuse $s_i = \sum_{k=1}^{K} c_{i,k} g_k$ de chaque source s_i sur le dictionnaire mono-canal $\mathcal{D}$, on peut écrire $x_n = \sum_{i,k} a_{n,i}\, c_{i,k}\, g_k$. Cela se traduit matriciellement par $\mathbf{x} = \mathbf{AC}\boldsymbol{\Phi}^T$, avec $\boldsymbol{\Phi}^T$ la matrice dont les lignes sont les atomes mono-canal g_k, et $\mathbf{C} = \{c_{i,k}\}_{i,k}$ une matrice de coeffi cients parcimonieux. Cette décomposition peut aussi s'écrire $\mathbf{x} = \sum_{i,k} c_{i,k}\, \mathbf{a}_i\, g_k$, c'est-`a-dire que $\mathbf{x}$ admet une décomposition parcimonieuse sur le dictionnaire multi-canal "directionnel" constitué des atomes $\mathbf{a}_i\, g_k = (a_{1,i}\, g_k, \ldots, a_{N,i}\, g_k)$. Le produit scalaire se calcule alors par :

$$\langle \mathbf{R}^M, \mathbf{a}_i g_k \rangle = \mathbf{a}_i^T \mathbf{R}^M g_k^T.$$

et la reconstruction de la source s_i se fait par :

$$\hat{s}_i = \sum_k c_{i,k} g_k. \qquad (2)$$

Ce nouvel algorithme est appelé MPD et ses propriétés théoriques ont été étudiées dans [10]. L'utilisation d'un dictionnaire directionnel revient `a appliquer Matching Pursuit multi-canal en contraignant les coeffi cients $\mathbf{c}_{k_M}$ de la section 3.1 `a être proportionnels `a une colonne $\mathbf{a}_i$ de $\mathbf{A}$.

4 Expériences

Nous comparons MPC1, MPC2 et MPD avec les algorithmes de référence MSL, BZ et DUET. Les expériences portent sur le mélange stéréo linéaire instantané de trois sources musicales. La premi`ere est un enregistrement de violoncelle, la deuxi`eme de percussions, et la troisi`eme de piano. La fréquence d'échantillonnage des signaux est de 8000 Hz pour une durée de 2.4s, soit 19200 échantillons. La matrice de mélange est la suivante :

$$\begin{bmatrix} \cos(\pi/8) & \cos(\pi/4) & \cos(3\pi/8) \\ \sin(\pi/8) & \sin(\pi/4) & \sin(3\pi/8) \end{bmatrix}$$

Les sources y sont séparées d'un angle de $\pi/8$. L'énergie des percussions, situées au milieu, est environ deux fois plus faible que les énergies du piano et du violoncelle qui sont du même ordre.

Nous utilisons les mesures de performance de séparation proposées dans [7], qui permettent d'analyser fi nement l'origine des distorsions distinguant la source estimée de la source réelle. Ces mesures, exprimées en décibels, sont le rapport entre le signal et a) la distorsion relative totale (SDR), b) celle due aux interférences (SIR) et c) celle due aux artéfacts de calcul (SAR). Pour ces trois mesures, de la même famille que le classique rapport signal `a bruit, les performances sont d'autant meilleures que le rapport est élevé.

4.1 Algorithmes de référence

Les performances des algorithmes basés sur Matching Pursuit sont comparées `a celles de trois algorithmes : le meilleur séparateur linéaire (MSL) [7], DUET [6] et l'algorithme de Bofi ll-Zibulevski (BZ) [8].

Le premier consiste uniquement en l'application d'une matrice $\mathbf{B}$ au signal. $\mathbf{B}$ est telle que les sources estimées $\hat{\mathbf{s}} = \mathbf{Bx}$ minimisent la distorsion due aux interférences [7]. Si on suppose les sources mutuellement orthogonales, et si on connaît la matrice de mélange $\mathbf{A}$ et la matrice diagonale $\mathbf{D}$ des normes des sources, alors en notant $\hat{\mathbf{A}} = \mathbf{AD}$, la matrice $\mathbf{B}$ est donnée par $\hat{\mathbf{A}} : \mathbf{B} = \mathbf{D}\hat{\mathbf{A}}^H(\hat{\mathbf{A}}\hat{\mathbf{A}}^H)^{-1}$.

L'algorithme DUET [6] applique une transformée de Fourier `a court terme (TFCT) `a chaque canal du signal, puis applique un masquage en supposant une seule source active pour chaque "case" temps-fréquence, et enfi n inverse la TFCT pour construire la source estimée.

L'algorithme de Bofi ll-Zibulevski [8] repose sur le même principe que DUET, `a la différence pr`es que chaque case temps-fréquence est attribuée aux deux sources les plus proches. Cette attribution est déterminée par une minimisation de norme l_1.

Dans toutes les expériences, ces deux algorithmes sont appliqués avec une fenêtre de Hanning, de taille 4096 échantillons, avec un décalage de 2048 échantillons (50% de la taille de la fenêtre). Les performances de DUET et BZ sont fortement dépendantes du dimensionnement de la fenêtre, et nous avons observé qu'une taille plus grande et surtout plus petite les dégrade fortement dans les cas étudiés.

4.2 Variantes des algorithmes MP

Dans cette expérience, on étudie l'influence du nombre d'itérations, de la composition du dictionnaire, de l'exploitation du résidu et d'un post-traitement par lissage dans le cadre de l'algorithme MPD. Deux dictionnaires peuvent être utilisés pour la décomposition :

- un "petit" dictionnaire constitué d'atomes de Gabor de taille $s = 4096$ qui se recouvrent `a moitié ($u = ns/2$, $n \in \mathbb{N}$). Il correspond `a la TFCT utilisée par les algorithmes DUET et BZ.
- un "grand" dictionnaire constitué d'atomes de Gabor de taille $s = 64$ `a 16384 (par puissance de deux). Le décalage entre atomes est également de 50% de la taille d'un atome.

La fi gure 1 représente les rapports SDR, SIR et SAR de la source "piano" estimée par les différents algorithmes, en fonction du nombre d'itérations (les résultats étant similaires pour les deux autres sources).

Premi`erement, on remarque que tout au long des itérations l'utilisation du grand dictionnaire conduit `a une meilleure séparation que le petit dictionnaire. En effet, dans le cas du grand dictionnaire, l'algorithme MP choisit lui-même la taille de fenêtre optimale. Ainsi la nécessité d'optimiser la taille de fenêtre *a priori* est supprimée, contrairement au cas des algorithmes BZ et DUET.

Par ailleurs, on peut remarquer que l'amélioration des performances est monotone lorsque le nombre d'itérations augmente. Dans le détail, les artéfacts, qui dominent la distorsion, sont importants lorsque les sources sont reconstruites avec peu

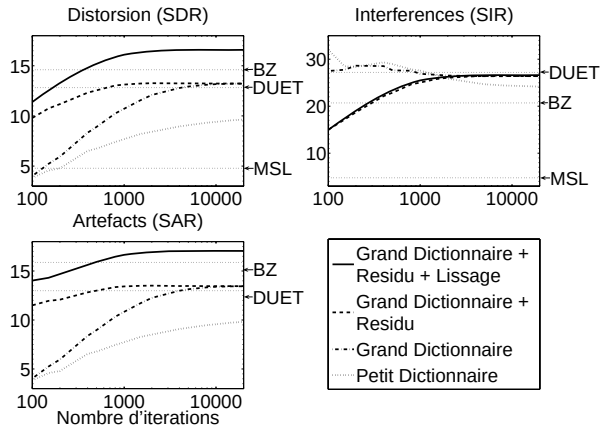


FIG. 1 – Distorsions (dB), source "piano" estimée par MPD

d'atomes, et ils diminuent au fi l des itérations grâce `a l'apport de nouveaux atomes. Apr`es suffi samment d'itérations, MPD devient meilleur que DUET sur le plan des taux d'artéfacts et de distorsion.

Pour compenser la distorsion due au faible nombre d'atomes, on peut séparer le résidu de la décomposition $\mathbf{R}^M$ en utilisant le séparateur linéaire $\mathbf{A}^H(\mathbf{AA}^H)^{-1}\mathbf{R}^M$ et l'ajouter aux sources estimées. Le séparateur utilisé suppose que tous les résidus des sources ont la même énergie. Asymptotiquement, cette hypoth`ese se vérifi e, les sources les plus énergétiques étant celles dont les atomes sont les premiers sélectionnés. L'addition du résidu augmente largement la performance globale (SDR) pour un faible nombre d'itérations. Cet effet se décompose en une forte diminution des artéfacts, au détriment d'une augmentation des interférences. Pour un plus grand nombre d'itérations, l'énergie du résidu approche de zéro et l'amélioration apportée sur le SDR diminue. La réduction des artéfacts permet d'obtenir de meilleures performances globales que DUET `a partir d'un nombre réduit d'itérations (moins de 1000).

Néanmoins, les performances de MPD avec séparation du résidu restent inférieures `a celles de l'algorithme BZ avec une fenêtre optimisée. Partant de l'hypoth`ese que le lissage introduit par le recouvrement des fenêtres de la TFCT dans l'algorithme BZ joue un rôle dans sa bonne performance [11], nous avons appliqué un lissage des sources estimées par MPD avec ajout du résidu. Ce lissage consiste `a effectuer plusieurs estimations des sources `a partir de versions décalées du dictionnaire et `a produire la moyenne de ces estimations. L'amélioration apportée par le lissage, tr`es nette sur le plan des artéfacts ($\sim$ 3dB), mais pas systématique pour les interférences, permet d'obtenir de meilleurs performances en SDR et SAR que DUET et BZ.

Pour MPC1 et MPC2, le changement de dictionnaire, l'ajout du résidu et le lissage produisent des effets du même type que ceux observés pour MPD. Le caract`ere itératif de ces algorithmes implique un fort coût de calcul, compensé par la mise en œuvre d'une implémentation rapide de l'algorithme [12].

4.3 Comparaison des performances

L'expérience suivante est destinée `a évaluer la capacité des différents algorithmes `a maintenir une bonne séparation lorsque la matrice de mélange n'est pas connue, mais seulement approchée. Une imprécision volontaire est introduite par

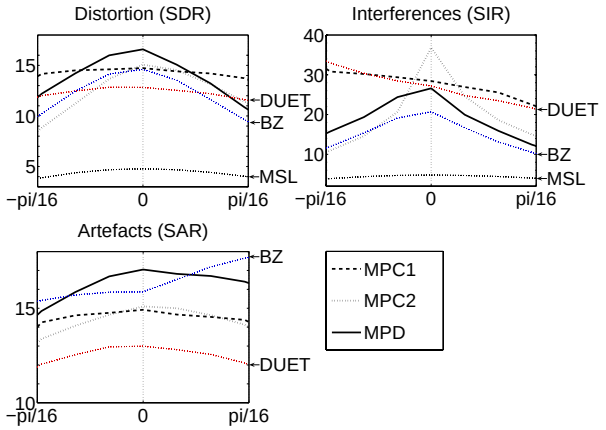


FIG. 2 – Distorsions (dB), source “piano” en fonction de l’angle de perturbation

rotation de la vraie matrice. Les directions des trois sources sont décalées d’un angle égal, qui varie entre $-\pi/16$ et $\pi/16$. Les expériences sont effectuées avec le “grand” dictionnaire. Elles incluent la séparation du résidu et le lissage, et utilisent 5000 itérations. Les performances sont données sur la fi gure 2, en fonction de l’angle de perturbation, pour le piano.

Évolution du SAR – Les méthodes étudiées gardent un niveau d’artéfacts `a peu pr`es constant en fonction de la perturbation. MPD et BZ introduisent le moins d’artéfacts, suivies par MPC1 et MPC2 qui présentent des performances équivalentes, elles-mêmes meilleures que DUET. Ces niveaux d’artéfacts sont intrins`eques aux mod`eles sous-jacents `a chaque méthode.

Évolution du SIR – Pour les méthodes MPC1 et DUET, les atomes temps-fréquence sont attribués `a une seule source. Ainsi, ces méthodes produisent le plus faible niveau d’interférences et restent robustes `a la perturbation de la matrice de mélange. Dans le cas de MPC2, MPD et BZ, le fait d’attribuer les atomes temps-fréquence `a plusieurs sources introduit une plus grande sensibilité `a la perturbation de la matrice de mélange. Pour une matrice de mélange bien estimée, MPC2 produit le moins d’interférences.

Évolution du SDR – Par défi nition, la distorsion globale (SDR) est dominée par le minimum du SAR ou du SIR. Pour une matrice de mélange bien estimée, les méthodes basées sur Matching Pursuit obtiennent de meilleures performances globales que les méthodes de référence. Par ordre décroissant de performance, les méthodes s’échelonnent de la mani`ere suivante : MPD, MPC1, MPC2, BZ, DUET et MSL. En revanche, lorsqu’une perturbation est introduite sur la matrice de mélange, les méthodes MPC1 et DUET (attribution `a une seule direction) s’av`erent plus robustes que les méthodes MPC2, BZ (attribution `a deux directions, ce qui entraîne une plus grande sensibilité en mati`ere d’interférences) et que MPD (sélection des atomes par Matching Pursuit seulement sur les directions estimées des sources).

5 Conclusions

Nous avons comparé plusieurs méthodes de séparation de sources sous-déterminée par décomposition parcimonieuse, en supposant que la matrice de mélange est connue. Dans les algorithmes MPC1 et MPC2, la matrice de mélange est utilisée *a posteriori* pour classer et regrouper les atomes issus de la décomposition par Matching Pursuit. Dans l’algorithme MPD, la connaissance de la matrice de mélange est incluse *a priori* dans la défi nition du dictionnaire. La version de MPD avec séparation du résidu et lissage donne de meilleures performances, en mati`ere de distorsion globale et d’artéfacts, que les méthodes DUET et BZ. Lorsque la matrice de mélange est bien estimée, MPD et MPC2 donnent les meilleurs résultats. En revanche, MPC1 et DUET semblent plus robustes `a une erreur d’estimation de la matrice de mélange.

Le formalisme proposé permet d’envisager l’estimation conjointe de la matrice de mélange et des sources, et par ailleurs l’apprentissage de dictionnaire multi-canal directionnel `a partir du mélange.

Références

[1] J.-F. Cardoso, “Blind signal separation : statistical principles,” *Proc. IEEE, Special issue on blind identification and estimation*, vol. 9, no. 10, pp. 2009–2025, Oct. 1998.

[2] O. Bermond and J.-F. Cardoso, “Méthodes de séparation de sources dans le cas sous-déterminé,” in *Proc. GRETSI, Vannes, France*, 1999, pp. 749–752.

[3] M. Zibulevsky and B. Pearlmutter, “Blind source separation by sparse decomposition in a signal dictionary,” *Neural Computations*, vol. 13, no. 4, pp. 863–882, 2001.

[4] S. Mallat and Z. Zhang, “Matching pursuit with time-frequency dictionaries,” *IEEE Trans. on Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.

[5] R. Gribonval, “Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture,” in *Proc. Int. Conf. Acoust. Speech Signal Process. (ICASSP’02), Orlando, Florida, USA, May 2002*, 2002.

[6] O. Yilmaz and S. Rickard, “Blind separation of speech mixtures via time-frequency masking,” *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[7] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, “Proposals for performance measurement in source separation,” in *Proc. 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, Nara, Japan, Apr. 2003, pp. 763–768.

[8] P. Bofill and M. Zibulevsky, “Blind separation of more sources than mixtures using sparsity of their short-time fourier transform,” in *Proc. ICA2000*, Helsinki, june 2000, pp. 87–92.

[9] R. Gribonval, “Piecewise linear source separation,” in *Proc. SPIE’03 – ”Wavelets : Applications in Signal and Image Processing”, San Diego, California, USA*, vol. 5207, August 2003, pp. 297–310.

[10] R. Gribonval and M. Nielsen, “Beyond sparsity : recovering structured representations by l1-minimization and greedy algorithms. – application to the analysis of sparse underdetermined ICA–,” IRISA, Tech. Rep. 1684, Jan. 2005, http ://www.irisa.fr/metiss/gribonval/.

[11] S. Araki, S. Makino, H. Sawada, and R. Mukai, “Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask,” in *Proceedings in Acoustics, Speech, and Signal Processing (ICASSP’ 05)*, vol. 3, march 2005, pp. 81–84.

[12] R. Gribonval and S. Krstulović,“The Matching Pursuit ToolKit”, voir `http ://www.irisa.fr/metiss/matchp/`.

Séparation d'un mélange sous-déterminé de sources par simulation stochastique

Éric Le Carpentier

Institut de Recherche en Communications et Cybernétique de Nantes
IRCCyN - 1, rue de la Noë - BP 92 101 - 44321 Nantes CEDEX 03
Eric.Le-Carpentier@irccyn.ec-nantes.fr

Résumé – Nous proposons dans cet article un algorithme de Gibbs pour l'identification, la séparation et la déconvolution d'un mélange mono-capteur de sources i.i.d. mutuellement indépendantes régies par une loi de Bernoulli-Gauss, et filtrées linéairement. Il s'agit d'une étude préparatoire, l'objectif à moyen-terme étant l'analyse des signaux électromyographiques pour le diagnostic des pathologies neuro-musculaires. L'extension au cas multi-capteur (mélange convolutif) est immédiate.

Abstract – This paper deals with the derivation of a Gibbs algorithm for mixture separation and deconvolution, when the observed signal is a sum of linear filtering of mutually independant Bernoulli-Gauss i.i.d. sequences. This is a first step towards electromyographic signals analysis, for automatic diagnosis purpose. The extension to multi-sensors case is straightforward.

1 Introduction

Soit un signal scalaire à temps discret $y = (y_n)_{n\in\mathbb{N}}$ modélisable de la façon suivante : pour tout instant n :

$$\begin{aligned} y_n &= \sum_{i=1}^{s} (h^{(i)} * x^{(i)})_n + w_n \\ &= \sum_{i=1}^{s}\sum_{d=0}^{\ell} h_d^{(i)}\, x_{n-d}^{(i)} + w_n \end{aligned}$$

où les $x^{(i)}$, $1 \leq i \leq s$ sont des séquences aléatoires i.i.d. mutuellement indépendantes et suivant une loi de Bernoulli-Gauss, les $h^{(i)}$ sont des réponses impulsionnelles de longueur finie $\ell + 1$, et telles que $h_0^{(i)} = 1$, $*$ dénote le produit de convolution, et w est un bruit additif gaussien, indépendant des $x^{(i)}$.

Ce modèle est apte à représenter, par exemple, des signaux biomédicaux de type électromyographiques [1]. Ces signaux sont constitués d'une superposition d'ondelettes élémentaires, les potentiels d'unité motrice, se propageant le long des muscles. La forme de ces ondelettes correspond aux réponses impulsionnelles $h^{(i)}$, et les paramètres de la loi de Bernoulli-Gauss renseignent sur la probabilité d'apparition de telles ondelettes. Ces divers paramètres peuvent être utilisés pour caractériser et diagnostiquer certaines pathologies neuromusculaires.

L'objectif est donc l'estimation des réponses impulsionnelles et des paramètres des lois de probabilité (identification), ainsi que la reconstruction des séquences $x^{(i)}$ (séparation et déconvolution), à partir d'un enregistrement de longueur finie $y_{1:N} = (y_n)_{1\leq n\leq N}$.

Un problème analogue a été traité par Sénécal et Amblard, dans le cas d'un mélange instantané et déterminé de sources discrètes [2]. Dans le contexte convolutif, Cheng *et al.* propose une méthode de déconvolution de signaux sismiques, dans un cas mono-source [3]. Dans les deux cas, une approche de simulation bayesienne par MCMC est proposée [4]. Nous proposons dans cet article d'étendre les résultats de Cheng au contexte multi-source mono-capteur.

2 Approche bayesienne

Soit θ l'ensemble des paramètres à estimer, et soit $\rho : \theta \mapsto \rho(\theta)$ la densité de θ conditionnellement à l'observation $y_{1:N}$. L'estimateur de l'espérance *a posteriori* s'écrit :

$$\hat{\theta} = \mathcal{E}\{\theta|y_{1:N}\} \quad (1)$$

$$= \int \theta\, \rho(\theta)\, d\theta \quad (2)$$

où $\mathcal{E}\{\}$ désigne l'espérance mathématique. Nous approchons cet estimateur par Monte-Carlo, en tirant T valeurs de θ suivant la loi de densité ρ.

$$\hat{\theta} \approx \frac{1}{T}\sum_{t=1}^{T} \theta^{(t)} \quad (3)$$

Pour effectuer cette opération, nous utilisons un algorithme de Gibbs, qui consiste, lorsque le vecteur paramètre θ est scindé sous la forme $(\theta_1, \ldots, \theta_d)$, à tirer récursivement chaque sous-paramètre θ_k conditionnellement à $(\theta_\ell)_{\ell\neq k}$. En pratique, il s'agit d'exprimer préalablement la densité ρ, puis d'isoler pour chaque k les termes dépendant de θ_k, les autres paramètres étant fixés. Par la règle de Bayes, on obtient ρ par la formule :

$$\rho(\theta) = p(\theta|y_{1:N}) \propto p(y_{1:N}|\theta)\, p(\theta) \quad (4)$$

Il faut donc exprimer la densité de la mesure conditionnellement aux paramètres, ainsi que les *a priori* sur ces paramètres.

3 Formalisation

Posons $g(.,V)$ la densité de la loi de Gauss centrée de variance-covariance V. La densité des sources s'écrit alors, pour tout n et pour tout i, en notant $\lambda_1^{(i)}$ le coefficient de mélange pour la i^{e} source :

$$p(x_n^{(i)}|\lambda_1^{(i)}\mu^{(i)}\nu^{(i)}) = \sum_{k=0}^{1} \lambda_k^{(i)}\, g(x_n^{(i)} - \mu_k^{(i)}, \nu_k^{(i)}) \qquad (5)$$

avec :

$$\lambda_0^{(i)} = 1 - \lambda_1^{(i)} \qquad \text{et} \qquad \mu_0^{(i)} = \sqrt{\nu_0^{(i)}} = 0 \qquad (6)$$

Et la densité du bruit additif s'écrit, si r est sa variance :

$$p(w_n|r) = g(w_n, r) \qquad (7)$$

Pour mener à bien les calculs, il est préférable de procéder à une augmentation de variable, en marquant les $(x_n^{(i)})$ par la composante $k_n^{(i)} \in \{0,1\}$ suivant laquelle $x_n^{(i)}$ a été tirée. La loi des variables marquées s'écrit donc :

$$p(x_n^{(i)}k_n^{(i)}|\mu_1^{(i)}\nu_1^{(i)}\lambda_1^{(i)}) = \lambda_{k_n^{(i)}}^{(i)}\, g(x_n^{(i)} - \mu_{k_n^{(i)}}^{(i)}, \nu_{k_n^{(i)}}^{(i)}) \qquad (8)$$

Le vecteur des paramètres à estimer vaut donc :

$$\theta = ((x_n^{(1:s)})_n, (k_n^{(1:s)})_n, h^{(1:s)}, r, \mu_1^{(1:s)}, \nu_1^{(1:s)}, \lambda_1^{(1:s)}) \qquad (9)$$

avec n variant de $1-\ell$ à N, pour éviter toute troncature, comme le montre plus bas la formule 12.

La loi suivant laquelle il faudra tirer les paramètres s'écrit, en l'absence de tout *a priori* sur $\mu_1^{(1:s)}, \nu_1^{(1:s)}, \lambda_1^{(1:s)}$:

$$\rho(\theta) \propto p(y_{1:N}|\theta)\, p(\theta) \qquad (10)$$

$$\propto \quad p(y_{1:N}|(x_n^{(1:s)})_n, (k_n^{(1:s)})_n, h^{(1:s)}, r) \times p((x_n^{(1:s)})_n, (k_n^{(1:s)})_n|\mu_1^{(1:s)}, \nu_1^{(1:s)}, \lambda_1^{(1:s)}) \qquad (11)$$

Posons :

$$\epsilon_n = y_n - \sum_{i=1}^{s}\sum_{d=0}^{\ell} h_d^{(i)} x_{n-d}^{(i)} \qquad (12)$$

Comme fait classiquement en filtrage linéaire, on obtient facilement que le premier terme de l'équation 11 s'écrit :

$$p(y_{1:N}|(x_n^{(1:s)})_n, (k_n^{(1:s)})_n, h^{(1:s)}, r) \qquad (13)$$

$$= \quad \prod_{n=1}^{N} p_{w_n|r}(\epsilon_n) \qquad (14)$$

$$= \quad \prod_{n=1}^{N} g(\epsilon_n, r) \qquad (15)$$

D'autre part, en tirant parti de la mutuelle indépendance des sources, et de leur caractère i.i.d.

$$p((x_n^{(1:s)})_n, (k_n^{(1:s)})_n|\mu_1^{(1:s)}, \nu_1^{(1:s)}, \lambda_1^{(1:s)}) \qquad (16)$$

$$= \quad \prod_{i=1}^{s}\prod_{n=1-\ell}^{N} p(x_n^{(i)}, k_n^{(i)}|\mu_1^{(i)}, \nu_1^{(i)}, \lambda_1^{(i)}) \qquad (17)$$

$$= \quad \lambda_{k_n^{(i)}}^{(i)}\, g(x_n^{(i)} - \mu_{k_n^{(i)}}^{(i)}, \nu_{k_n^{(i)}}^{(i)}) \qquad (18)$$

On obtient finalement :

$$\rho(\theta) \propto \prod_{n=1}^{N} g\,(\epsilon_n, r) \times \prod_{i=1}^{s}\prod_{n=1-\ell}^{N} \lambda_{k_n^{(i)}}^{(i)}\, g(x_n^{(i)} - \mu_{k_n^{(i)}}^{(i)}, \nu_{k_n^{(i)}}^{(i)}) \qquad (19)$$

4 Lois conditionnelles

Le calcul des densités conditionnelles consiste alors à exhiber, dans la formule 19, les termes dépendant de chaque paramètre, les autres étant fixés.

4.1 Paramètres de mélange $\lambda_1^{(i)}$

En isolant les termes de la formule 19 dépendant de $\lambda_1^{(i)}$, on obtient (les produits et sommes sur n portent sur l'intervalle $\{1-\ell, \ldots, N\}$) :

$$\rho(\lambda_1^{(i)}|\text{reste}) \propto \prod_n \lambda_{k_n^{(i)}}^{(i)} \qquad (20)$$

$$\propto \quad (1-\lambda_1^{(i)})^{N+\ell-\sum_n k_n^{(i)}}\, (\lambda_1^{(i)})^{\sum_n k_n^{(i)}} \qquad (21)$$

Il faut donc tirer $\lambda_1^{(i)}$ suivant une loi beta, soit $\mathcal{B}\text{e}\Big(\sum_n k_n^{(i)} + 1, N + \ell - \sum_n k_n^{(i)} + 1\Big)$.

4.2 Moyennes $\mu_1^{(i)}$

En isolant les termes de la formule 19 dépendant de $\mu_1^{(i)}$, on obtient (les produits et sommes sur n portent sur l'intervalle $\{1-\ell, \ldots, N\}$) :

$$\rho(\mu_1^{(i)}|\text{reste}) \propto \prod_{\substack{n \\ k_n^{(i)}=1}}^{N} g(x_n^{(i)} - \mu_1^{(i)}, \nu_1^{(i)}) \qquad (22)$$

$$\propto \quad \exp\left[-\frac{1}{2\nu_1^{(i)}} \sum_n k_n^{(i)}(x_n^{(i)} - \mu_1^{(i)})^2\right] \qquad (23)$$

$$\propto \quad \exp\left[-\frac{\sum_n k_n^{(i)}}{2\nu_1^{(i)}} \left(\mu_1^{(i)} - \frac{\sum_n k_n^{(i)} x_n^{(i)}}{\sum_n k_n^{(i)}}\right)^2\right] \qquad (24)$$

Il faut donc tirer $\mu^{(i)}$ suivant une loi normale, soit $\mathcal{N}\left(\frac{\sum_n k_n^{(i)} x_n^{(i)}}{\sum_n k_n^{(i)}}, \frac{\nu_1^{(i)}}{\sum_n k_n^{(i)}}\right)$.

4.3 Variances $\nu_1^{(i)}$

En isolant les termes de la formule 19 dépendant de $\nu_1^{(i)}$, on obtient (les produits et sommes sur n portent sur l'intervalle $\{1-\ell, \ldots, N\}$) :

$$\rho(\nu_1^{(i)}|\text{reste}) \propto \prod_{\substack{n \\ k_n^{(i)}=1}}^{N} g(x_n^{(i)} - \mu_1^{(i)}, \nu_1^{(i)}) \qquad (25)$$

$$\propto \quad \frac{1}{{\nu_1^{(i)}}^{\frac{1}{2}\sum_n k_n^{(i)}}} \exp\left[-\frac{1}{2\nu_1^{(i)}} \sum_n k_n^{(i)}(x_n^{(i)} - \mu_1^{(i)})^2\right] \qquad (26)$$

Il faut donc tirer $\nu^{(i)}$ suivant une loi gamma inverse, soit $\mathcal{IG}\left(\frac{1}{2}\sum_n k_n^{(i)} - 1, \frac{1}{2}\sum_n k_n^{(i)}(x_n^{(i)} - \mu_1^{(i)})^2\right)$.

4.4 Variance r

En isolant les termes de la formule 19 dépendant de r, on obtient (les produits et sommes sur n portent sur

l'intervalle $\{1,\ldots,N\}$) :

$$\rho(r|\text{reste}) \propto \prod_n g(\epsilon_n, r) \quad (27)$$

$$\propto \frac{1}{r^{\frac{N}{2}}} \exp\left[-\frac{1}{2r}\sum_n \epsilon_n^2\right] \quad (28)$$

Il faut donc tirer r suivant une loi gamma inverse, soit $\mathcal{IG}(\frac{N}{2}-1,\ \frac{1}{2}\sum_n \epsilon_n^2)$.

4.5 Réponses impulsionnelles $h^{(i)}$

Définissons les vecteurs suivants :

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix} \quad x^{(i)} = \begin{bmatrix} x_1^{(i)} \\ \vdots \\ x_N^{(\ell)} \end{bmatrix} \quad h^{(i)} = \begin{bmatrix} h_1^{(i)} \\ \vdots \\ h_\ell^{(i)} \end{bmatrix} \quad (29)$$

ainsi que les matrices Toeplitz circulantes suivantes :

$$\mathcal{X}^{(i)} = \begin{bmatrix} x_0^{(i)} & \ldots & x_{1-\ell}^{(i)} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & x_0^{(i)} \\ x_{N-1}^{(i)} & \ldots & x_{N-\ell}^{(i)} \end{bmatrix} \quad (30)$$

L'équation 12 peut alors s'écrire sous forme vectorielle :

$$\epsilon = y - \sum_j x^{(j)} - \sum_j \mathcal{X}^{(j)} h^{(j)} \quad (31)$$

$$= \underbrace{y - \sum_j x^{(j)} - \sum_{j\neq i} \mathcal{X}^{(j)} h^{(j)}}_{\tilde{\epsilon}^{(i)}} - \mathcal{X}^{(i)} h^{(i)} \quad (32)$$

Le terme $\tilde{\epsilon}^{(i)}$ ne dépend pas de $h^{(i)}$. En isolant, dans la formule 19, les termes dépendant de $h^{(i)}$, on obtient :

$$\rho(h^{(i)}|\text{reste}) \propto \prod_n g(\epsilon_n, r) \quad (33)$$

$$\propto \exp\left[-\frac{1}{2r}\sum_n \epsilon_n^2\right] \quad (34)$$

$$\propto \exp\left[-\frac{1}{2r}\epsilon^{\mathrm{T}}\epsilon\right] \quad (35)$$

$$\propto \exp\left[-\frac{1}{2r}(\tilde{\epsilon}^{(i)} - \mathcal{X}^{(i)}h^{(i)})^{\mathrm{T}}(\tilde{\epsilon}^{(i)} - \mathcal{X}^{(i)}h^{(i)})\right] \quad (36)$$

En faisant apparaître les carrés nécessaires, on obtient que $h^{(i)}$ doit être tiré suivant une loi normale, soit $\mathcal{N}\left(\left(\mathcal{X}^{(i)\mathrm{T}}\mathcal{X}^{(i)}\right)^{-1}\mathcal{X}^{(i)\mathrm{T}}\tilde{\epsilon}^{(i)}, r\left(\mathcal{X}^{(i)\mathrm{T}}\mathcal{X}^{(i)}\right)^{-1}\right)$

4.6 Entrées marquées $(x_n^{(i)}, k_n^{(i)})$

Rappelons que n doit varier ici de $1-\ell$ à N. Définissons, en complément des vecteurs définis en 29, les vecteurs suivants, à $N+\ell$ lignes :

$$\bar{x}^{(i)} = \begin{bmatrix} x_{1-\ell}^{(i)} \\ \vdots \\ x_N^{(i)} \end{bmatrix} \quad (37)$$

ainsi que les matrices Toeplitz circulantes suivantes, à N lignes :

$$\mathcal{H}^{(i)} = \begin{bmatrix} h_p^{(i)} & \ldots & 1 & & 0 \\ & \ddots & & \ddots & \\ 0 & & h_p^{(i)} & \ldots & 1 \end{bmatrix} \quad (38)$$

En outre, nous noterons $\mathcal{H}_n^{(i)}$ la n^{e} colonne de $\mathcal{H}^{(i)}$. L'équation 12 peut alors s'écrire sous la nouvelle forme vectorielle :

$$\epsilon = y - \sum_j \mathcal{H}^{(j)}\bar{x}^{(j)} \quad (39)$$

$$= y - \sum_{j\neq i} \mathcal{H}^{(j)}\bar{x}^{(j)} - \mathcal{H}^{(i)}\bar{x}^{(i)} \quad (40)$$

$$= \underbrace{y - \sum_{j\neq i} \mathcal{H}^{(j)}\bar{x}^{(j)} - \sum_{m\neq n} \mathcal{H}_m^{(i)} x_m^{(i)}}_{\bar{\epsilon}^{(i)}} - \mathcal{H}_n^{(i)} x_n^{(i)} \quad (41)$$

Le terme $\bar{\epsilon}^{(i)}$ ne dépend pas de $x_n^{(i)}$ et $k_n^{(i)}$. En isolant, dans la formule 19, les termes dépendant de $x_n^{(i)}$ et $k_n^{(i)}$, on obtient :

$$\rho(x_n^{(i)}, k_n^{(i)}|\text{reste}) \quad (42)$$

$$\propto \prod_{m=1}^{N} g(\epsilon_m, r)\, g(x_n^{(i)} - \mu_{k_n^{(i)}}^{(i)}, \nu_{k_n^{(i)}}^{(i)})\, \lambda_{k_n^{(i)}}^{(i)} \quad (43)$$

$$\propto g(\epsilon, rI_N)\, g(x_n^{(i)} - \mu_{k_n^{(i)}}^{(i)}, \nu_{k_n^{(i)}}^{(i)})\, \lambda_{k_n^{(i)}}^{(i)} \quad (44)$$

$$\propto g(\bar{\epsilon}^{(i)} - \mathcal{H}_n^{(i)}x_n^{(i)}, rI_N)\, g(x_n^{(i)} - \mu_{k_n^{(i)}}^{(i)}, \nu_{k_n^{(i)}}^{(i)})\, \lambda_{k_n^{(i)}}^{(i)} \quad (45)$$

En faisant apparaître les carrés nécessaires, le premier terme conduit à la densité de la loi de Gauss de moyenne $\bar{\mu}_n^{(i)}$ et de variance $\bar{\nu}_n^{(i)}$ par rapport à $x_n^{(i)}$, définies par :

$$\bar{\mu}_n^{(i)} = \frac{\mathcal{H}_n^{(i)\mathrm{T}}\bar{\epsilon}_n^{(i)}}{\mathcal{H}_n^{(i)\mathrm{T}}\mathcal{H}_n^{(i)}} \qquad \bar{\nu}_n^{(i)} = \frac{r}{\mathcal{H}_n^{(i)\mathrm{T}}\mathcal{H}_n^{(i)}} \quad (46)$$

La densité conditionnelle s'écrit donc, en utilisant les règles de produit de densités de Gauss :

$$\rho(x_n^{(i)}, k_n^{(i)}|\text{reste}) \quad (47)$$

$$\propto g(x_n^{(i)} - \bar{\mu}_n^{(i)}, \nu_n^{(i)})\, g(x_n^{(i)} - \mu_{k_n^{(i)}}^{(i)}, \nu_{k_n^{(i)}}^{(i)})\, \lambda_{k_n^{(i)}}^{(i)} \quad (48)$$

$$\propto g\left(x_n^{(i)} - \frac{\bar{\nu}_n^{(i)}\mu_{k_n^{(i)}}^{(i)} + \nu_{k_n^{(i)}}^{(i)}\bar{\mu}_n^{(i)}}{\nu_{k_n^{(i)}}^{(i)} + \bar{\nu}_n^{(i)}}, \frac{\nu_{k_n^{(i)}}^{(i)}\bar{\nu}_n^{(i)}}{\nu_{k_n^{(i)}}^{(i)} + \bar{\nu}_n^{(i)}}\right) \times \underbrace{g(\mu_{k_n^{(i)}}^{(i)} - \bar{\mu}_n^{(i)}, \nu_{k_n^{(i)}}^{(i)} + \bar{\nu}_n^{(i)})\, \lambda_{k_n^{(i)}}^{(i)}}_{\bar{\lambda}_{k_n^{(i)}}^{(i)}} \quad (49)$$

L'étiquette $k_n^{(i)}$ doit donc prendre la valeur 1 avec la probabilité $\frac{\bar{\lambda}_1^{(i)}}{\bar{\lambda}_0^{(i)}+\bar{\lambda}_1^{(i)}}$, 0 avec la probabilité $\frac{\bar{\lambda}_0^{(i)}}{\bar{\lambda}_0^{(i)}+\bar{\lambda}_1^{(i)}}$. $x_n^{(i)}$ est ensuite tirée suivant une loi normale, soit $\mathcal{N}\left(\frac{\bar{\nu}_n^{(i)}\mu_{k_n^{(i)}}^{(i)} + \nu_{k_n^{(i)}}^{(i)}\bar{\mu}_n^{(i)}}{\nu_{k_n^{(i)}}^{(i)} + \bar{\nu}_n^{(i)}}, \frac{\nu_{k_n^{(i)}}^{(i)}\bar{\nu}_n^{(i)}}{\nu_{k_n^{(i)}}^{(i)} + \bar{\nu}_n^{(i)}}\right)$. On constate immédiatement que dans le cas Bernoulli-gaussien, si $k_n^{(i)}$ prend la valeur 0, $x_n^{(i)}$ est obligatoirement nul.

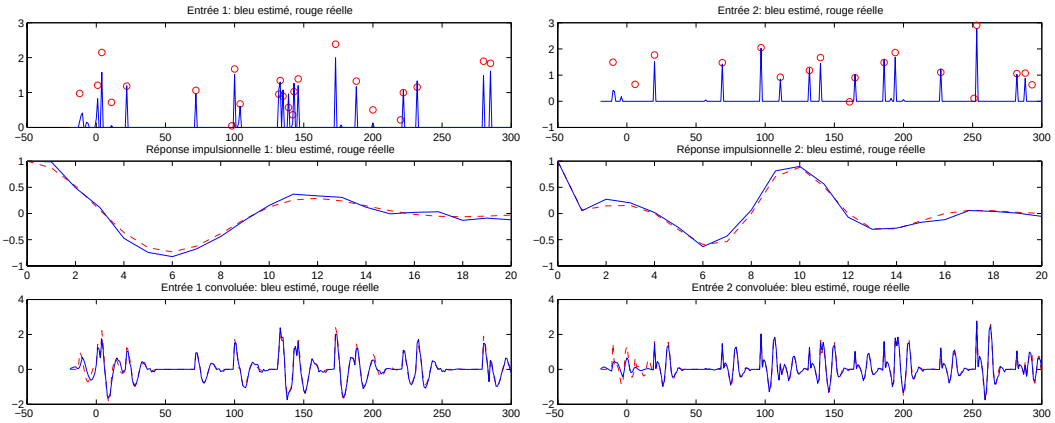


FIG. 1 – Deux entrées de Bernoulli-Gauss, une sortie, bruit 5dB

5 Simulations

Sur la figure 1, sont représentées les résultats dans le cas de 2 sources bernoulli-gaussiennes, dans le cas où les ondelettes $h^{(i)}$ sont de longueur $\ell = 20$. Sont tracées, pour chaque source i, la séquence $x^{(i)}$, l'ondelette $h^{(i)}$, et le filtrage $h^{(i)} * x^{(i)}$, dans leurs versions simulée ($\circ$ ou pointillé) et reconstruite (trait plein).

	Simulé	Estimé
Bruit	$\sqrt{r} = 0.0489$	$\sqrt{r} = 0.0546$
Entrée 1	$\lambda = 0.0750$	$\lambda = 0.0638$
	$\mu_1 = 1.105$	$\mu_1 = 1.178$
	$\nu_1 = 0.346$	$\nu_1 = 0.231$
Entrée 2	$\lambda = 0.0562$	$\lambda = 0.0546$
	$\mu_1 = 1.237$	$\mu_1 = 1.233$
	$\nu_1 = 0.490$	$\nu_1 = 0.637$

6 Conclusion et perspectives

Les résultats en simulation semble démontrer la faisabilité d'une approche de type MCMC pour la séparation de sources bernoulli-gaussiennes filtrées linéairement, pour lesquelles on ne dispose que d'un seul mélange. Les résultats peuvent s'étendre au cas multi-sortie (sous-déterminé ou non) dans le cadre des mélanges convolutifs. L'extension au cas de filtres à réponse impulsionnelle infinie, voire à non minimum de phase (en utilisant la formalisation générale des modèles ARMA [5]), est envisagée. En ce qui concerne les signaux électromyographiques, il est nécessaire de compliquer le modèle, car l'hypothèse de sources indépendantes s'avère abusive ; il est en effet nécessaire de prendre en compte un temps de latence entre les instants de mise à feu de chaque source ; l'implantation implique alors une modification de l'algorithme décrit ci-dessus, qui devra réaliser une mise à jour de type Metropolis-Hastings.

Remerciements

Je tiens à remercier Nicolas BOURDAUD, pour l'aide précieuse qu'il m'a apportée lors de son stage de DEA à l'IRCCyN, en 2004.

Références

[1] C.J. De Luca. Physiology and Mathematics of Myoelectric Signals. *IEEE Transactions on Biomedical Engineering*, 26(6) :313–325, juin 1979.

[2] S. Sénécal et P.O. Amblard. Séparation de sources par méthodes de simulation. *18e colloque GRETSI*, volume 2, pages 319–322, Toulouse, France, septembre 2001.

[3] Q. Cheng, R. Chen, et T.H. Li. Simultaneous wavelet estimation and deconvolution of reflection seismic signals. *IEEE Transactions on geoscience and remote sensing*, 34(2) :377–384, mars 1996.

[4] C.P. Robert. *The bayesian Choice*. Springer, 2001.

[5] O. Shalvi et E. Weinstein. Maximum likelihood and lower bounds in system identification with non-gaussian inputs. *IEEE Trans. on Inform. Theory*, 40(2) :328–339, 1994.

Construction d'estimateurs oracles pour la séparation de sources

Emmanuel VINCENT[1], Rémi GRIBONVAL[2]

[1]Centre for Digital Music, Queen Mary, University of London
Mile End Road, London E1 4NS, United Kingdom

[2]Projet METISS, IRISA-INRIA
Campus de Beaulieu, 35042 Rennes Cedex, France
emmanuel.vincent@elec.qmul.ac.uk, remi.gribonval@irisa.fr

Résumé – La séparation de sources pour des mélanges sous-déterminés et/ou convolutifs est un problème difficile pour lequel de nombreux algorithmes ont été proposés. Afin d'étudier leur performance, nous définissons des estimateurs oracles permettant de calculer la performance maximale théorique de différentes classe d'algorithmes de séparation dans un cadre d'évaluation où les sources de référence sont disponibles. Nous implémentons ces estimateurs pour deux classes (les algorithmes de séparation par filtrage stationnaire et ceux procédant par masquage temps-fréquence) et nous étudions leur performance sur quelques exemples de mélanges audio.

Abstract – Source separation of under-determined and/or convolutive mixtures is a difficult problem that has been addressed by many algorithms. In order to study their performance, we define oracle estimators that compute the maximal theoretical performance achievable for various classes of algorithms in an evaluation framework where the reference sources are available. We implement these estimators for two classes (stationary filtering separation algorithms and time-frequency masking separation algorithms) and we study their performance on a few audio mixture examples.

1 Introduction

La majorité des signaux audio, vidéo et biomédicaux sont des mélanges de plusieurs sources actives simultanément, acquis directement ou par mélange synthétique de signaux sources. En toute généralité, l'opération de mélange est une transformation non linéaire et non stationnaire des signaux sources. Cependant, la plupart des mélanges sont modélisables par un filtrage linéaire stationnaire. Le $i$ème canal du mélange observé s'exprime alors sous la forme $x_i = \sum_j a_{ij} \star s_j$, où (s_j) sont les signaux sources, (a_{ij}) les filtres de mélange et $\star$ représente la convolution. Le mélange est dit instantané lorsque les filtres de mélange sont de simples gains et convolutif dans le cas contraire. Il est aussi dit sur-déterminé lorsque le nombre de canaux observés est supérieur au nombre de sources et sous-déterminé sinon. L'étude des signaux mélangés soulève le problème de la séparation de sources, qui consiste à estimer le signal de chaque source avec la meilleure qualité possible (*i.e.* sans interférences provenant des autres sources ni distortion supplémentaire).

De nombreux algorithmes ont été proposés pour résoudre ce problème. Les mélanges sur-déterminés instantanés ou convolutifs sont généralement séparés par convolution des canaux de mélange par des filtres de démixage stationnaires à réponse impulsionnelle finie [1]. L'Analyse en Composantes Indépendantes (ACI) estime ces filtres de démixage en modélisant les sources par des processus i.i.d. non gaussiens [2, 1] ou des processus gaussiens de variance non stationnaire [3] indépendants. D'autres méthodes utilisent des modèles de sources plus complexes spécifiques à un type de sources donné [4, 5]. La séparation de mélanges sous-déterminés est plus souvent effectuée par masquage binaire temps-fréquence [6] ou par filtrage de Wiener adaptatif [7]. Les masques temps-fréquence sont alors estimés en fonction de la différence d'intensité [6] et de phase [8] entre les canaux du mélange observé en chaque point temps-fréquence, ou bien à l'aide de méthodes plus complexes prenant en compte la structure du spectre de puissance à court terme des sources [9, 7, 10].

La performance de ces diverses méthodes dépend beaucoup du type de mélange rencontré, voire des propriétés des sources et des filtres de mélange dans le mélange particulier étudié. Par exemple la performance de l'ACI sur des mélanges sur-déterminés convolutifs chute généralement lorsque la taille des filtres de mélange dépasse quelques centaines d'échantillons. Trois raisons sont en mesure d'expliquer cette baisse de performance. Premièrement, l'algorithme utilisé pour estimer les filtres de démixage peut échouer à trouver les filtres optimaux en pratique (problèmes de maxima locaux par exemple). Deuxièmement, les filtres de démixage optimaux en terme de performance peuvent ne pas correspondre aux filtres optimaux estimés en fonction du modèle de sources de l'ACI. Troisièmement, les contraintes imposées par le choix de la méthode de séparation (par exemple la relativement faible taille des filtres de séparation) peuvent limiter dans l'absolu la performance optimale possible. Pour améliorer les algorithmes d'ACI existants, il est indispensable de comprendre laquelle de ces raisons est prépondérante. En effet, un tel diagnostic permettrait de savoir s'il faut prioritairement construire de meilleurs algorithmes d'optimisation, de meilleurs modèles de sources, ou modifier l'hypothèse de séparation par filtrage stationnaire. Le même type de raisonnement s'applique aux autres méthodes de séparation.

Dans cet article, nous apportons une réponse partielle à cette question en construisant des estimateurs oracles des sources,

c'est-à-dire des estimateurs idéaux au sein d'une classe d'estimateurs possibles. Ces estimateurs ne peuvent par nature être utilisés que dans un contexte d'évaluation où les sources de référence sont disponibles. L'étude de la performance des estimateurs oracles permet de prédire l'adéquation d'une classe d'algorithmes de séparation à un mélange donné, ainsi que de situer les performances d'un algorithme donné par rapport à l'optimum possible dans la classe dont il fait partie. On aboutit ainsi également à une sorte d'indice mesurant la difficulté de séparation d'un mélange.

La structure de l'article est la suivante. Nous définissons les estimateurs oracles dans la partie 2. Puis dans la partie 3 nous proposons des algorithmes pour calculer les oracles de séparation par filtrage stationnaire et par masquage temps-fréquence. Enfin nous étudions leur performance sur des exemples audio dans la partie 4.

2 Définition d'un estimateur oracle

Supposons que nous disposons des canaux d'un mélange observé $(x_i)_{1\leq i\leq I}$ et de la $j$ième source de référence s_j. Les nombreuses méthodes de séparation de sources citées dans la partie 1 se ramènent à deux classes générales : filtrage stationnaire et masquage temps-fréquence. Au sein d'une classe donnée, tous les algorithmes estiment la $j$ième source sous la forme

$$\widehat{s_j} = f_{\theta_j}(x_1, \ldots, x_I), \tag{1}$$

où θ_j représente les paramètres de séparation permettant de calculer la source séparée en fonction des canaux observés, *i.e.* les coefficients des filtres de démixage ou des masques temps-fréquence utilisés. La différence entre les algorithmes tient à la façon dont ces paramètres sont estimés à partir des données observées $x_1, \ldots, x_I$.

Puisque la source de référence est supposée connue, il est possible d'évaluer la performance de séparation en mesurant la qualité de la source estimée à l'aide d'une mesure $q(\widehat{s_j}, s_j)$. L'estimateur oracle de la $j$ième source est alors la fonction qui calcule la meilleure estimation de la source parmi un ensemble de paramètres de séparation possibles Θ :

$$\widetilde{s_j}(\Theta) = \arg\max_{\theta_j\in\Theta} q(f_{\theta_j}(x_1, \ldots, x_I), s_j). \tag{2}$$

L'ensemble Θ fixe des contraintes sur les paramètres de séparation, comme la taille des filtres de démixage ou la largeur des trames utilisées pour le masquage temps-fréquence. L'étude de l'oracle consiste à calculer sa performance

$$\tilde{q}(\Theta) = q(\widetilde{s_j}(\Theta), s_j) = \max_{\theta_j\in\Theta} q(f_{\theta_j}(x_1, \ldots, x_I), s_j) \tag{3}$$

en fonction du mélange observé et des contraintes posées.

Notons que le calcul de l'oracle ne nécessite pas de connaître toutes les données du problème, mais seulement les données à estimer (dans ce cas la $j$ième source de référence mais pas les autres sources ni les filtres de mélange). Il est possible de définir de la même façon des estimateurs oracles pour d'autres quantités, par exemple pour l'image de la $j$ième source sur le $i$ème capteur $a_{ij} \star s_j$.

3 Calcul de l'estimateur

La première question qui se pose en pratique pour calculer un estimateur oracle est de savoir comment mesurer la qualité d'une source estimée. Généralement les sources séparées sont destinées à l'écoute et la quantité $q(\widehat{s_j}, s_j)$ devrait donc être proportionnelle à la distortion auditive perçue entre $\widehat{s_j}$ et s_j [11]. Cette quantité reste difficile à calculer actuellement, c'est pourquoi nous utilisons par la suite un simple Rapport Signal-à-Distortion (RSD) quadratique exprimé en déciBels (dB) par

$$q(\widehat{s_j}, s_j) = 10\log_{10}\frac{\|s_j\|^2}{\|\widehat{s_j} - s_j\|^2}, \tag{4}$$

où $\|a\|^2 = \sum_{t=0}^{T-1} a(t)^2$ est l'énergie totale d'un signal a de taille T. L'estimateur oracle de la $j$ième source est alors défini par $\widetilde{s_j}(\Theta) = \arg\min_{\theta_j\in\Theta} \|f_{\theta_j}(x_1, \ldots, x_I) - s_j\|^2$.

3.1 Oracle de filtrage stationnaire

Dans le cas d'une séparation par des filtres de démixage stationnaires (causaux) $(w_{ij})_{1\leq i\leq I}$ de taille L, l'estimée de la $j$ième source vaut $\widehat{s_j}(t) = \sum_{i=1}^{I}\sum_{\tau=0}^{L-1} w_{ij}(\tau)x_i(t-\tau)$. En notant $(x_{i\tau})_{1\leq i\leq I,\, 0\leq\tau\leq L-1}$ les versions retardées des canaux observés définies par $x_{i\tau}(t) = x_i(t-\tau)$ et en représentant le couple d'indices (i,τ) par un seul indice η variant entre 1 et $D = IL$, l'estimée prend aussi la forme

$$\widehat{s_j} = \sum_{\eta=1}^{D} w_{\eta j}\, x_\eta. \tag{5}$$

Le calcul des coefficients $(w_{\eta j})_{1\leq\eta\leq D}$ qui maximisent le RSD est un problème de moindre carrés linéaire dont la solution est donnée classiquement par les coefficients de la projection orthogonale de s_j sur le sous-espace vectoriel engendré par les signaux observés et leurs versions décalées $(x_\eta)_{1\leq\eta\leq D}$. Plus précisément, en notant par $\langle a, b\rangle = \sum_{t=0}^{T-1} a(t)b(t)$ le produit scalaire de deux signaux a et b de taille T, le vecteur des coefficients optimaux $\widetilde{\mathbf{w}_j} = [\widetilde{w_{\eta j}}]_{1\leq\eta\leq D}$ vaut

$$\widetilde{\mathbf{w}_j} = \mathbf{G}^{-1}\mathbf{d}, \tag{6}$$

où $\mathbf{G} = [\langle x_\eta, x_{\eta'}\rangle]_{1\leq\eta\leq D,\, 1\leq\eta'\leq D}$ est la matrice de Gram des canaux retardés et $\mathbf{d} = [\langle s_j, x_\eta\rangle]_{1\leq\eta\leq D}$.

3.2 Oracle de masquage temps-fréquence

L'estimation d'une source (par exemple la $j$ième) par masquage temps-fréquence consiste à représenter l'un des canaux observés (par exemple le $i$ème) dans une base de signaux localisés en temps-fréquence, puis à multiplier chaque coefficient de la représentation par un coefficient de masquage réel compris entre 0 et 1, et enfin à estimer la source désirée par inversion de la représentation au sens des moindres carrés. Dans la suite nous utilisons une base orthonormale de cosinus locaux (MDCT), qui présente l'avantage d'être facilement inversible. En appelant $(\phi_n)_{1\leq n\leq N}$ les éléments de la base et $(w_{nj})_{1\leq n\leq N}$ les coefficients de masquage, l'estimée de la $j$ième source vaut $\widehat{s_j}(t) = \sum_{n=1}^{N} w_{nj}\langle x_i, \phi_n\rangle\phi_n(t)$. Puisque la base est orthonormale, l'erreur quadratique sur la source estimée se décompose sous la forme d'une somme

$$\|\widehat{s_j} - s_j\|^2 = \sum_{n=1}^{N} |w_{nj}\langle x_i, \phi_n\rangle - \langle s_j, \phi_n\rangle|^2. \tag{7}$$

Cette somme est minimale lorsque chacun de ses termes est minimal. En notant $r_{nj} = \langle s_j, \phi_n \rangle / \langle x_i, \phi_n \rangle$, le $n$ième coefficient de masquage optimal vaut par conséquent

$$\widetilde{w_{nj}} = \begin{cases} 0 & \text{si } r_{nj} < 0, \\ r_{nj} & \text{si } 0 \leq r_{nj} \leq 1, \\ 1 & \text{si } r_{nj} > 1. \end{cases} \quad (8)$$

4 Étude de performance

Nous étudions la performance des deux estimateurs oracles développés sur quelques mélanges audio musicaux en fonction des contraintes posées (taille des filtres de démixage ou taille de fenêtre de MDCT). Les mélanges utilisés ont été choisis pour rendre compte de divers niveaux de difficulté liés soit à la nature du mélange (déterminé ou sous-déterminé, instantané ou convolutif à filtres courts ou longs) soit à celle des sources (indépendantes ou synchrones et en harmonie).

4.1 Données expérimentales

Nous considérons deux ensembles de sources échantillonnées à 22050 Hz : d'une part un ensemble de sources indépendantes provenant d'extraits solo de CD différents (un violon, un violoncelle, une clarinette), d'autre part un ensemble de sources synchrones et en harmonie synthétisées à partir des pistes d'un même fichier MIDI [1] (un violon, un alto, un violoncelle). Bien que le son des instruments MIDI soit de pauvre qualité, le deuxième ensemble est beaucoup plus réaliste car les sources suivent les règles de l'harmonie musicale.

Pour chaque ensemble de sources, nous effectuons créons deux séries de mélanges stéréo. La première série est composée de mélanges déterminés de deux sources (violon et violoncelle pour l'ensemble de sources indépendantes, violon et alto pour les sources MIDI synchrones). La seconde série est composée de mélanges sous-déterminés des trois sources de l'ensemble considéré.

Dans chaque série, trois mélanges différents de durée égale à 10 s sont engendrés à partir de chaque ensemble de sources par mélange instantané, convolutif court et convolutif long. Les mélanges convolutifs longs sont effectués par convolution avec des réponses impulsionnelles de salle enregistrées à l'IRCAM (temps de réverbération $T_{60} \simeq 800$ ms, soit environ 18000 échantillons) et les mélanges convolutifs courts par convolution avec ces mêmes réponses tronquées à leurs 512 premiers échantillons.

4.2 Résultats

La plupart des algorithmes de séparation de sources à partir de mélanges convolutifs estiment plutôt la contribution de chaque source sur chaque capteur que les sources elles-mêmes. Afin d'obtenir des bornes de performance pour ces algorithmes, nous utilisons les estimateurs oracles de filtrage stationnaire et de masquage temps-fréquence pour estimer l'image de chaque source sur le premier canal. Pour ce faire, nous appliquons la formule (6) (respectivement (8)) avec $\mathbf{d} = [\langle a_{1j} \star s_j, x_\eta \rangle]_{1 \leq \eta \leq D}$ (resp. $r_{nj} = \langle a_{1j} \star s_j, \phi_n \rangle / \langle x_i, \phi_n \rangle$), de façon à optimiser la mesure de qualité $q(\widehat{s_j}, a_{1j} \star s_j)$.

La performance de l'oracle, c'est-à-dire le RSD calculé sur chaque source estimée puis moyenné entre les sources, est calculée pour différentes tailles L des filtres de démixage (oracle de filtrage stationnaire) et différentes tailles L' du support des éléments de la base MDCT (oracle de masquage temps-fréquence). Les résultats pour les mélanges déterminés sont présentés dans le tableau 1, ceux pour les mélanges sous-déterminés dans le tableau 2.

Sans surprise, nous constatons que la performance de tous les oracles considérés est systématiquement meilleure sur les mélanges déterminés que sur les mélanges sous-déterminés : la différence entre les meilleurs oracles est d'environ 4 dB pour les mélanges convolutifs longs, 15 dB pour les mélanges convolutifs courts, et la performance des meilleurs oracles est parfaite pour les mélanges instantanés déterminés.

De même, les oracles ont toujours plus de difficulté à séparer des sources synchrones et en harmonie que des sources indépendantes, la différence de performance étant de l'ordre de 2 dB pour les ensembles de sources considérés.

En général, les mélanges convolutifs sont plus difficiles à séparer que les mélanges instantanés, et ce d'autant plus que les filtres de mélange sont longs. La différence de performance entre les meilleurs oracles pour les mélanges convolutifs courts et longs est importante (de l'ordre de 20 dB) dans le cas déterminé avec des sources indépendantes, plus faible mais significative (de l'ordre de 10 dB) dans le cas déterminé avec des sources synchrones, et négligeable dans le cas sous-déterminé.

Globalement, les seules conditions dans lesquelles une séparation de qualité semble réalisable avec les classes d'algorithmes considérés correspondent à des mélanges déterminés où les filtres de mélange sont courts (ou bien instantanés). Dans ce cas la séparation par filtrage stationnaire peut fournir de meilleurs résultats que le masquage temps-fréquence. Dans les autres cas, les facteurs de difficulté semblent avoir un effet cumulatif sur la dégradation des performances des meilleurs oracles, et le masquage temps-fréquence idéal est aussi bon ou meilleur que le filtrage stationnaire idéal.

5 Conclusion

Les estimateurs oracles proposés dans cet article fournissent un cadre pour calculer les performances idéales de séparation réalisables en théorie avec différentes classes d'algorithmes de séparation de sources. Ils sont disponibles, accompagnés d'exemples sonores, sous forme d'une boîte à outils MATLAB distribuée sous license GPL à l'adresse `http://www.irisa.fr/metiss/bss_oracle/`

En pratique, les algorithmes ne peuvent approcher ces performances que si les modèles de sources sont pertinents, les estimateurs (InfoMax, Maximum A Posteriori, ...) bien choisis et les techniques numériques d'optimisation (descente de gradient, ...) efficaces. D'après les expériences réalisées, le masquage temps-fréquence est potentiellement plus performant que la séparation par filtrage stationnaire sur des mélanges réalistes qui sont souvent sous-déterminés et convolutifs. Le filtrage stationnaire, qui correspond aux méthodes les plus classiques basées sur l'ACI, n'est potentiellement le meilleur que

[1] Il s'agit du fichier *Classical Music n° 15* de la base de données *RWC* [12].

TAB. 1 – Performance moyenne (RSD en dB) des estimateurs oracles sur des mélanges stéréo déterminés de deux sources.

Type d'oracle	Mélange de sources indépendantes			Mélange de sources synchrones		
	Instantané	Convolutif court	Convolutif long	Instantané	Convolutif court	Convolutif long
Filtrage $L = 128$	$+\infty$	19	10	$+\infty$	15	7
Filtrage $L = 512$	$+\infty$	31	13	$+\infty$	25	10
Filtrage $L = 2048$	$+\infty$	42	15	$+\infty$	28	13
Masquage $L' = 128$	12	12	13	10	10	10
Masquage $L' = 512$	16	15	17	15	14	14
Masquage $L' = 2048$	21	20	20	19	15	16

TAB. 2 – Performance moyenne (RSD en dB) des estimateurs oracles sur des mélanges stéréo sous-déterminés de trois sources.

Type d'oracle	Mélange de sources indépendantes			Mélange de sources synchrones		
	Instantané	Convolutif court	Convolutif long	Instantané	Convolutif court	Convolutif long
Filtrage $L = 128$	12	9	6	9	9	6
Filtrage $L = 512$	15	11	7	11	12	8
Filtrage $L = 2048$	16	13	9	12	13	10
Masquage $L' = 128$	10	9	9	7	8	7
Masquage $L' = 512$	15	12	13	10	10	10
Masquage $L' = 2048$	19	16	17	12	11	12

dans les cas déterminés instantanés ou convolutifs courts.

La performance de séparation par filtrage stationnaire a déjà été étudiée à l'aide d'un estimateur oracle supposant connus les filtres de mélange [13] sous l'hypothèse que les sources sont des bruits blancs indépendants. Cependant l'oracle que nous proposons en supposant connues les sources met en évidence un effet non négligeable de la nature des sources (selon qu'elles sont indépendantes ou synchrones). Une comparaison poussée des deux approches est donc nécessaire, et il faut notamment chercher dans les deux cas à mesurer et prédire la dégradation de performance constatée pour des estimateurs réels proches de l'estimateur oracle. Enfin, nous réfléchissons au calcul effectif des oracles pour d'autres mesures de qualité rendant mieux compte de la qualité perçue des résultats.

Références

[1] S. Amari, S. Douglas, A. Cichocki, and H. Yang, "Novel online algorithms for blind deconvolution using natural gradient approach," in *Proc. SYSID*, 1997, pp. 1057–1062.

[2] J.-F. Cardoso, "Blind source separation : statistical principles," *Proceedings of the IEEE*, vol. 9, no. 10, pp. 2009–2025, 1998.

[3] D.-T. Pham and J.-F. Cardoso, "Blind separation of instantaneous mixtures of non stationary sources," *IEEE Trans. on Signal Processing*, vol. 49, no. 9, pp. 1837–1848, 2001.

[4] N. Murata, S. Ikeda, and A. Ziehe, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[5] M. Reyes-Gomez, B. Raj, and D. Ellis, "Multi-channel source separation by factorial HMMs," in *Proc. ICASSP*, 2003, pp. 664–667.

[6] Ö. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

[7] L. Benaroya, "Séparation de plusieurs sources sonores avec un seul microphone," Ph.D. dissertation, Université Rennes I, 2003.

[8] N. Roman, D. Wang, and G. Brown, "Speech segregation based on sound localization," *Journal of the ASA*, vol. 114, no. 4, pp. 2236–2252, 2003.

[9] S. Roweis, "One microphone source separation," in *Proc. NIPS*, 2000, pp. 793–799.

[10] E. Vincent and X. Rodet, "Underdetermined source separation with structured source priors," in *Proc. ICA*, 2004, pp. 327–332.

[11] E. Vincent, C. Févotte, R. Gribonval, L. Benaroya, X. Rodet, A. Röbel, É. LeCarpentier, and F. Bimbot, "Comment évaluer les algorithmes de séparation de sources audio ?" in *Proc. GRETSI*, 2003, pp. 27–32.

[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database : Popular, classical, and jazz music databases," in *Proc. ISMIR*, 2002, pp. 287–288. [Online]. Available : http ://staff.aist.go.jp/m.goto/RWC-MDB/

[13] R. Balan, J. Rosca, and S. Rickard, "Robustness of parametric source demixing in echoic environments," in *Proc. ICA*, 2001, pp. 144–148.

Index des auteurs

Abadi Mohammed 367-370
Abbadi Sami 1013-1016
Abed-meraim Karim 65-68, 1233-1236
Abeida Habti 69-72
Abergel Alain 133-136
Aboutajeddine Driss 1065-1068
Abrial Pierrick 303-306
Abry Patrice 153-156, 279-282, 323-326, 549-552
Achard Catherine 5-8
Adde Patrick 795-798
Adib Abdellah 1065-1068
Adnet Claude 1181-1184
Aeschlimann Fabien 827-830
Aime Claude 265-268
Ait-el-fquih Boujemaa 213-216
Aknin Patrice 101-104
Alasseur Clémence 209-212
Albouy Benjamin 385-388
Alexandre Rolim Fernandes Carlos 57-60
Alleysson David 347-350
Allier Emmanuel 827-830
Almansa Andrés 513-516
Alvarado Victor 719-722
Amara Amara 9-12
Amat Philippe 1045-1048
Amblard Pierre-olivier 89-92
Amiel Frédéric 9-12
Amoud Hassan 367-370
Andriamisaina Caaliph 823-826
André Thomas 1001-1004
André-obrecht Régine 473-476
Anthoine Sandrine 307-310
Antoni Jérôme 461-464, 1097-1100
Antonini Marc 1001-1004, 1013-1016
Aoudi Walid 363-366
Ardon Roberto 707-710
Argentieri Sylvain 85-88
Armspach Jean-paul 331-334
Assemat Louis 1141-1144
Astruc Jean-pierre 221-224
Ataman Stefan 835-838
Atli Benediktsson Jon 505-508
Attilio Pegoraro Paolo 315-318
Aubert Gilles 1113-1116
Aufort Gabriel 429-432
Augereau Bertrand 1017-1020
Augusto Negri Pablo 679-682
Aupetit-berthelemot Christelle 839-842

Auvray Vincent 157-160
Ayache Alain 521-524
Aïssa-el-bey Abdeldjalil 1233-1236
Babel Marie 1021-1024
Badier Jean-michel 433-436
Baillet Sylvain 189-192
Bailly Gérard 1137-1140, 1221-1224
Bali Nadia 1061-1064
Baras Cléo 1205-1208
Barbaresco Frédéric 921-924, 1181-1184, 1209-1212
Barbier Daniel 253-256
Barbot Jean-pierre 619-622
Bardinet Eric 189-192
Barlaud Michel 1001-1004, 1113-1116
Barnault Loic 631-634
Bartelmaos Steve 65-68
Bas Patrick 1129-1132
Basset Olivier 21-24, 53-56
Batatia Hadj 521-524
Baudais Jean-yves 735-738
Baudoin Geneviève 1025-1028
Bay Stéphanie 619-622
Baylou Pierre 145-148
Beaumesni Brice 489-492
Beauseroy Pierre 1145-1148
Beghdadi Azeddine 29-32
Belaroussi Rachid 671-674, 941-944
Belkacemi Hocine 1177-1180
Bellanger Jean-jacques 425-428
Bellanger Maurice 45-48
Bellard Elisabeth 185-188
Belloir Fabien 397-400
Ben Amar Chokri 1013-1016
Benboudjema Dalila 909-912
Bendjoya Philippe 257-260
Benedetto F. 615-618
Benois Pineau Jenny 97-100
Benslimane Anis 465-468
Bensrhair Abdelaziz 683-686
Berar Maxime 1137-1140
Bernard Olivier 359-362
Bertin Emmanuel 299-302
Bertolino Pascal 339-342, 889-892
Besserer Bernard 1161-1164
Biard Mathieu 249-252
Bichot Elodie 1149-1152
Bimbot Frédéric 1229-1232
Blacher Silvia 413-416
Blanchard Frédéric 17-20
Blasi D. 615-618
Bleuzzen Aurore 249-252

Bloch Isabelle 283-286, 937-940
Bobillet William 1085-1088
Bombardier Vincent 695-698
Bomel Pierre 779-782
Bon Nicolas 1169-1172
Bonacci David 579-582
Bonnefont Michel 1049-1052
Borgnat Pierre 323-326
Boubchir Larbi 233-236
Boucheret Marie-laure 961-964
Bouchoux Sophie 803-806
Boudraa Abdel-ouahab 29-32, 77-80, 559-562
Boujemâa Hatem 747-750
Bounsiar Abdenour 1145-1148
Bourdon Pascal 1017-1020
Bourennane El-bay 803-806
Bourennane Salah 33-36, 73-76, 77-80
Boustany Roger 1097-1100
Bouthemy Patrick 157-160
Boutillon Emmanuel 853-856
Bouveyron Charles 711-714
Bozkurt Baris 481-484
Brie David 1053-1056, 1077-1080, 1081-1084
Brochard Jacques 465-468
Brossier Jean-marc 89-92, 619-622
Brunel Nicolas 921-924
Brusseau Elisabeth 287-290
Buzuloiu Vasile 493-496
Bérar Maxime 161-164
C. M. Mota Joao 81-84
Cachard Christian 417-420
Caillault Emilie 687-690
Cammas Nathalie 997-1000
Canals Raphael 137-140
Cancès Jean-pierre 839-842
Canu Stéphane 763-766, 949-952
Capelle Anne-sophie 703-706
Caplier Alice 481-484
Capman François 1025-1028
Cappé Olivier 201-204
Capricelli Thomas 655-658
Cardoso Jean-françois 303-306
Cariou Claude 1121-1124
Cariou Laurent 743-746
Carminati Lionel 97-100
Caron François 217-220
Carrault Guy 241-244, 377-380
Carteret Cédric 1053-1056
Cartron Mickaël 787-790
Casseau Emmanuel 823-826
Castella Marc 1073-1076

Cavaro-menard Christine 525-528, 533-536, 639-642
Cesar Moura Mota João 57-60
Cexus Jean-christophe 559-562
Chabah Myriam 1169-1172
Chabert Marie 1033-1036
Chabrier Sébastien 917-920
Chainais Pierre 541-544
Chaix De Lavarène Brice 347-350
Chakchouk Moez 371-374
Chanussot Jocelyn 117-120, 505-508
Chapeau-blondeau François 595-598
Chapron Michel 1141-1144
Charbonnier Pierre 1109-1112, 1213-1216
Chareyron Gael 1037-1040
Charpentier Patrick 695-698
Charrier Christophe 913-916
Charron Cyril 351-354
Chassande-mottin Eric 261-264
Chateau Thierry 791-794
Chatellier Christian 1017-1020
Chatellier Laurence 457-460
Chaufray Jean-marie 65-68, 843-848
Chaumont Marc 489-492, 675-678
Chauvet Wilfried 849-852
Chaux Caroline 229-232
Chehdi Kacem 335-338, 659-662, 667-670, 1121-1124
Chenoune Yasmina 389-392
Chenu-tournier Marc 869-872
Chevalier Pascal 1057-1060
Chevreuil Antoine 1057-1060
Chibane Alexandre 401-404
Chihab Najat 221-224
Chonavel Thierry 477-480
Christov Nicolai 1089-1092
Chupin Marie 189-192
Ciblat Philippe 169-172
Ciuciu Philippe 295-298
Cladel Nicolas 1225-1228
Clady Xavier 679-682
Claire Jonchery 1157-1160
Clarysse Patrick 287-290
Claude Isabelle 409-412
Coatrieux Jean-louis 197-200
Coche Emmanuel 529-532
Cointault Frédéric 727-730
Collewet Christophe 1-4
Colot Olivier 703-706
Coltuc Dinu 1037-1040, 1197-1200
Combettes Patrick 655-658
Comon Pierre 1069-1072, 1093-1096
Condat Laurent 237-240

Conseil Simon 33-36
Coquin Didier 493-496
Corre Gwenole 779-782
Cosandier-rimélé Delphine 433-436
Coudray Renan 1161-1164
Coulais Alain 133-136, 311-314
Coulon Pierre-yves 1221-1224
Courboulay Vincent 639-642
Courtellemont Pierre 933-936, 1149-1152
Coussy Philippe 779-782
Couvreur Laurent 481-484
Cros Lionel 575-578
Crussière Matthieu 735-738
Cygan Hervé 101-104
Czyz Jacek 205-208
D'hondt Olivier 1185-1188
D'hooge Jan 359-362
Danes Patrick 85-88
Daoudi Khalid 723-726
Daout Franck 1173-1176
Dare-emzivat Délphine 77-80
Daubechies Ingrid 307-310
Davignon Franck 21-24, 53-56
Davoine Franck 327-330, 1201-1204
Davy Manuel 217-220, 973-976, 985-988
De Almeida L. F. André 81-84
De Craene Mathieu 1029-1032
De Roover Cédric 1105-1108
De Stéfano Pierre 775-778
Declercq David 607-610
Declercq David 631-634
Defontaine Antoine 377-380
Delabrouille Jacques 303-306
Delaigle Jean-françois 1005-1008
Delmas Jean-pierre 1181-1184
Deléchelle Eric 389-392
Dervin Mathieu 961-964
Desbat Laurent 193-196, 401-404
Desbouvries François 213-216
Descombes Xavier 509-512
Desurmont Xavier 1005-1008
Desvignes Michel 161-164, 783-786, 1137-1140
Devaux Marie-françoise 397-400
Devillard François 807-810
Dillenseger Jean-louis 355-358
Discher Antoine 1153-1156
Do Minh-tan 465-468
Dobigeon Nicolas 105-108
Dole Hervé 133-136
Domingues Antonio 469-472
Doncarli Christian 245-248

Doncescu Andrei 393-396
Dousse Olivier 269-272
Doussot Michel 367-370
Driouchi Driss 469-472
Dubois Corentin 985-988
Dubost Stéphanie 457-460
Ducasse Alain 961-964
Duchêne Bernard 635-638
Duchêne Cédric 89-92, 109-112
Duchêne Jacques 181-184, 367-370
Duffner Stefan 1113-1116
Duflos Emmanuel 217-220
Dugué Pierre 381-384
Duhamel Pierre 221-224, 953-956
Dumas Jean-michel 839-842
Dumoulin Jean 659-662
Durand Rémi 113-116
Dutilleux Guillaume 1109-1112
Dutoit Thierry 481-484
Duval Laurent 229-232
Duverdier Alban 849-852
Décréau Pierrette 517-520
Déforges Olivier 1021-1024
Ea Thomas 9-12, 819-822
Echelard Antoine 545-548
Eglin Véronique 691-694
Elhafid Abdelkamel 397-400
Elkefi Akran 1013-1016
Ellouze Noureddine 663-666
Emile Bruno 699-702
Ernst Harald 209-212
Eudes Jacques 631-634
Facciolo Gabriele 513-516
Fadaili Elmostafa 981-984
Fadili Jalal 233-236
Faisan Sylvain 331-334
Falcou Joel 791-794
Farah Georges 181-184
Fattaccioli Dominique 587-590
Faucon Gérard 381-384
Faure Laetitia 393-396
Fauvel Mathieu 505-508
Favier Gérard 57-60, 81-84, 177-180
Fernandez-maloigne Christine 1-4, 703-706
Ferrari André 265-268
Ferro-famil Laurent 501-504, 1185-1188
Fesquet Laurent 827-830
Feuilloy Mathieu 185-188
Fijalkow Inbar 623-626, 835-838, 869-872, 957-960
Fiorio Christophe 1193-1196
Flandrin Patrick 153-156, 755-758, 969-972

Fliess Michel 173-176, 571-574
Foidart Jean-michel 413-416
Foret Guillaume 889-892
Forster Philippe 113-116, 121-124, 141-144, 767-770
Fortrat Jacques-olivier 185-188
Fossorier Marc 607-610
Foster Stuart 363-366
Foulonneau Alban 1213-1216
Fournier-prunaret Danièle 1049-1052
François Serignat Jean 485-488
Françoise Dibos 1157-1160
Friboulet Denis 359-362
Frogerais Paul 425-428
Frélicot Carl 933-936
Féron Olivier 635-638
Fuchs Jean-Jacques 759-762
Gac Nicolas 783-786
Gacon Pierre 1221-1224
Gadat Sébastien 945-948
Ganoun Ali 137-140
Garda Patrick 819-822
Garello René 1169-1172
Garnero Line 189-192
Garot Jérôme 389-392
Garreau Mireille 197-200
Gasser Alain 429-432
Gaussier Philippe 1217-1220
Gazengel Bruno 1041-1044
Gelle Guillaume 599-602
Geller Benoît 619-622
Georges Koepfler 1157-1160
Germain Christian 145-148
Germain Guy 567-570
Germond Cécile 1181-1184
Ghogho Mounir 169-172
Gimenez Gérard 21-24, 53-56
Ginolhac Guillaume 113-116, 767-770, 1173-1176
Giovannelli Jean-françois 311-314, 457-460
Girard Stéphane 711-714
Girault Jean-marc 249-252
Giunta G. 615-618
Gosselin Bernard 529-532, 901-904
Gouiffès Michèle 1-4
Goursaud Claire 839-842
Gouze Annabelle 1105-1108
Goy François 253-256
Graba Tarik 819-822
Graffigne Christine 371-374
Grall-maës Edith 1145-1148
Granado Bertrand 819-822
Grandidier Frédéric 807-810

Grangeat Pierre 193-196, 401-404
Grenier Pierre 553-556
Grenier Thomas 21-24, 53-56
Grenier Yves 1233-1236
Gribonval Rémi 1229-1232, 1237-1240, 1245-1248
Grivel Eric 1085-1088, 1089-1092
Gros Patrick 37-40
Guglielmi Véronique 1049-1052
Guidorzi Roberto 1085-1088
Guigue Vincent 719-722, 763-766, 949-952
Guigues Laurent 885-888
Guillard Gwenaël 291-294
Guillemaud Régis 253-256
Guillemin Fabrice 273-276
Guillemot Christine 37-40, 731-734
Guillemot Ludovic 1189-1192
Guillet Valery 65-68
Guérin Alexandre 381-384
Guérin Dominique 727-730
Hadda Sami 861-864
Hafemeister Laurence 1217-1220
Hamioud Farida 715-718
Hamitouche Chafiaâ 291-294
Hamlaoui Soumya 327-330
Hammal Zakia 481-484
Hamrouni Kamel 663-666
Hamza Radhi 371-374
Harba Rachid 429-432
Hasboun Dominique 189-192
Hassani Moha 177-180
Hazart Aurélien 457-460
Heit Bernard 807-810
Heitz Fabrice 1213-1216
Herbin Michel 17-20
Herbulot Ariane 1113-1116
Hernandez Alfredo 241-244, 377-380, 425-428
Herve Nicolas 811-814
Hewson David 181-184, 367-370
Heyrman Barthélémy 771-774
Histace Aymeric 525-528, 639-642
Ho Ba Tho Marie-christine 409-412
Honeiné Paul 969-972
Houacine Amrane 441-444
Houcke Sébastien 877-880
Huart Jérémy 339-342, 889-892
Huet Sylvain 823-826
Huillery Julien 381-384
Husson Lionel 209-212
Hélard Jean-françois 743-746
Hérault Jeanny 347-350, 1117-1120
Idier Jérôme 295-298, 651-654, 1077-1080, 1081-1084

Ikhlef Aïssa 873-876
Ioana Cornel 575-578
Ionescu Bogdan 493-496
Iordache Razvan 29-32
Istrate Dan 485-488
Ivert Blaise 253-256
J.j. Michel Olivier 257-260
Jabloun Meryem 977-980
Jacquelet Yann 437-440, 893-896
Jaffard Stephane 549-552
Jallon Pierre 1057-1060
Jansen Maarten 225-228
Jarboui Slaheddine 861-864
Jarifi Safaa 897-900
Jaïdane Mériem 371-374
Jbari Atman 1065-1068
Jeanne Marion 603-606
Jehan-besson Stéphanie 1113-1116
Jennane Rachid 429-432
Joannis Claude 715-718
Join Cédric 173-176, 571-574
Jolion Jean-michel 925-928
Joséphine Castaing 965-968
Joumaa Hussein 1201-1204
Journet Nicholas 691-694
Julien-vergonjanne Anne 839-842
Jurie Frédéric 791-794
Jutten Christian 1101-1104
Jégou Hervé 731-734
Kaddous Taha Abdel 1049-1052
Kahlifa Nawres 663-666
Kerouedan Sylvie 795-798
Khalil Hachicha 815-818
Khanfouci Mourad 831-834
Kharbouche Saïd 929-932
Khenchaf Ali 1169-1172
Khoudeir Majdi 465-468
Kibangou Alain 177-180
Kieffer Michel 953-956
Kinie Abel 437-440, 421-424, 893-896
Klaine Luc 335-338, 667-670
Kouamé Denis 249-252
Krstulovic Sacha 1237-1240
Labarre David 1089-1092
Labat Christian 651-654
Labbani-igbida Ouiddad 351-354
Lacaze Bernard 849-852, 1033-1036
Lacoume Jean-louis 449-452, 587-590
Lafarge Florent 509-512
Lagrange Jean-michel 707-710
Laguitton Soizic 355-358

Lambert Patrick 493-496
Lamine Ndiaye Mamadou 893-896
Lamy-bergot Catherine 957-960
Lanchantin Pierre 905-908
Langlais Charlotte 627-630
Lantéri Henri 129-132
Larrieu Nicolas 319-322, 323-326
Larue Anthony 1101-1104
Larzabal Pascal 121-124, 141-144
Lashermes Bruno 549-552
Lathauwer Lieven De 965-968
Lathuilière Alexandra 343-346
Laurent Benhamou Claude 429-432
Laurent Christophe 25-28
Laurent Hélène 699-702, 917-920
Lavallou François 647-650
Le Bec Gaël 375-376
Le Bihan Nicolas 13-16, 61-64
Le Bouquin Jeannès Régine 381-384
Le Breton Hervé 197-200
Le Callet Patrick 611-614
Le Duff Alain 1041-1044
Le Gal Bertrand 823-826
Le Guen Benjamin 997-1000
Le Guennec Daniel 873-876
Le Hégarat-mascle Sylvie 93-96
Le Maire Sophie 413-416
Le Masson Jérôme 627-630
Le Touzé Grégoire 587-590
Lebossé Jérôme 537-540
Lebrun Gilles 913-916
Lecellier François 537-540
Lecomte Christèle 929-932
Leconge Rémy 533-536
Leducq Paul 501-504
Lee Chang-ming 953-956
Lefèvre Sébastien 1125-1128
Lejeune Gérard 583-586
Leray Pierre 799-802
Lesage Sylvain 1237-1240
Lesturgie Marc 113-116
Letellier Laurent 771-774
Levy Vehel Jacques 553-556
Li Jian-jin 541-544
Liebgott Hervé 363-366
Lienard Jean 157-160
Lopera Olga 643-646
Lopez-martinez Carlos 445-448
Loubaton Philippe 1057-1060
Louradour Jérôme 723-726
Louët Yves 865-868

Lucas Marie-françoise 245-248
Lucas Yves 385-388, 469-472
Luthon Franck 489-492
Lutton Evelyne 553-556
Lédée Roger 517-520, 529-532, 533-536
Léonard François 977-980
Lévy-vehel Jacques 545-548
Lézoray Olivier 913-916
Maaoui Choubeila 699-702
Macq Benoît 205-208, 529-532, 643-646, 1029-1032, 1105-1108
Mailhes Corinne 579-582, 583-586
Maillard Mickaël 1217-1220
Maitrot Aude 245-248
Makni Salima 295-298
Mallet Gregory 719-722
Mancas Matei 901-904
Mancas-thillou Céline 901-904
Mancini Stéphane 783-786
Mannoni Valérian 599-602, 603-606
Marc Chassery Jean 1129-1132
Marché Pierre 469-472
Marcos Sylvie 831-834, 1177-1180
Marque Catherine 567-570
Mars Jérôme I. 61-64, 449-452, 497-500, 587-590, 1101-1104
Martial Mari Jean 417-420
Martin Eric 779-782, 823-826
Martin Lionel 33-36
Martin Nadine 583-586, 977-980
Martin Vincent 1033-1036
Marzani Franck 343-346
Mascarilla Laurent 1149-1152
Massiani Arnaud 775-778
Massot Corentin 1117-1120
Mathieu-marni Sandrine 509-512
Maussang Frédéric 117-120
Mayrargue Sylvie 739-742
Mazet Vincent 1077-1080, 1081-1084
Mboup Mamadou 173-176, 571-574
Meessen Jérôme 1005-1008
Menard Daniel 811-814
Menard Michel 639-642
Mesnager Gilles 961-964
Miche Pierre 929-932
Michel Lagrange Jean 647-650
Michel Valérie 367-370
Michelet Franck 145-148
Milgram Maurice 5-8, 671-674, 679-682, 941-944
Milisavljevic Nada 643-646
Miron Sebastian 61-64
Mohammad-djafari Ali 635-638, 1061-1064
Mohammad-khani Gholam-reza 953-956

Moisan Lionel 93-96
Montanvert Annick 237-240
Montois Jean-jacques 421-424, 437-440, 893-896
Moreau Eric 981-984
Moreau Nicolas 1205-1208
Morin Luce 997-1000
Mostafaoui Ghilès 5-8
Moudden Yassir 303-306
Moulin Céline 253-256
Moulines Eric 201-204
Moureaux Jean-marie 1189-1192
Moussaoui Saïd 1053-1056
Mullot Rémy 691-694
Mustapha Mouaddib El 351-354
Ménard Daniel 49-52
Nafkha Amor 853-856
Najim Mohamed 1085-1088, 1089-1092
Narduzzi Claudio 315-318
Ndiaye Mamadou 421-424, 437-440
Needles Andrew 363-366
Neri A. 615-618
Nicolas Barbara 449-452, 587-590
Nicolas Pascal 185-188
Normand Nicolas 611-614
Nouvel Fabienne 775-778
Nuzillard Danielle 397-400
Olivier Amblard Pierre 109-112
Olivier Christian 1017-1020
Ouazine Boulem 441-444
Oudin Marc 1181-1184
Ovarlez Jean-philippe 121-124
Owezarski Philippe 319-322, 323-326
Ozerov Alexey 1229-1232
Padellini Marc 1025-1028
Paindavoine Michel 771-774
Palicot Jacques 799-802, 865-868, 873-876, 881-884
Pardo Alvaro 513-516
Parisot Christophe 1005-1008
Parrein Benoit 611-614
Pascal Frédéric 121-124
Pastor Dominique 125-128, 897-900
Patard Jean-jacques 355-358
Pateux Stéphane 997-1000
Patrick Garda 815-818
Paul Yonnet Jean 375-376
Paulus Caroline 497-500
Payan Yohan 1137-1140
Pellerin Denis 1165-1168
Pesquet Jean-christophe 229-232, 1009-1012, 1073-1076
Pesquet-popescu Beatrice 1009-1012
Petit Eric 389-392

Petrisor Teodora 1009-1012
Petrochilos Nicolas 397-400
Philipp-foliguet Sylvie 885-888
Philippe Delahaye Jean 799-802
Philippe Pierrick 1229-1232
Picinbono Bernard 751-754
Pieczynski Wojciech 909-912, 921-924
Pierpaoli Elena 307-310
Pierre Delmas Jean 69-72
Pinel Pierre 1049-1052
Pinna Andrea 819-822
Pinquier Julien 473-476
Poggi Sylvain 185-188
Poline Jean-baptiste 295-298
Pollet Claire 559-562
Pothin Jean-baptiste 1133-1136
Pottier Eric 445-448, 501-504, 1185-1188
Pouletaut Philippe 409-412
Poulliat Charly 957-960
Prevost Lionel 671-674, 941-944
Pronzato Luc 149-152
Prêteux Françoise 1153-1156
Puech William 1045-1048, 1193-1196
Pyndiah Ramesh 405-408
Quellec Jean-michel 1169-1172
Quinquis André 453-456, 575-578
Radoi Emanuel 453-456
Ragot José 715-718
Rahmouni Alain 389-392
Rajhi Hatem 371-374
Rajih Myriam 1093-1096
Rakotomamonjy Alain 683-686, 719-722, 763-766, 917-920, 949-952
Ramasso Emmanuel 1165-1168
Ramel Jean-yves 691-694
Randrianarisolo Solofohery 389-392
Raoof Kosai 375-376
Rauch Jean-louis 517-520
Regis Sebastien 393-396
Renaudin Marc 827-830
Renaux Alexandre 141-144
Renoult Adrien 869-872
Revenu Marinette 537-540
Revol-muller Chantal 21-24, 53-56
Richard Cédric 41-44, 969-972, 1133-1136
Rilling Gabriel 755-758
Ristic Branko 205-208
Robin Amandine 93-96
Rocher Romuald 49-52
Rodet Thomas 133-136
Rojo Guerra Pablo 257-260
Roland Christian 853-856

Romain Olivier 819-822
Romaniuk Barbara 161-164, 1137-1140
Rombaut Michèle 1165-1168
Rondao Alface Patrice 1029-1032
Ronsin Joseph 1021-1024
Roques Sylvie 563-566
Ros Julien 25-28
Rosec Olivier 477-480, 897-900
Rosenberger Christophe 917-920
Rossant Florence 9-12, 937-940
Rostaing Philippe 857-860
Rota Ludwig 1069-1072
Roue Benoit 1129-1132
Rougon Nicolas 1153-1156
Roumy Aline 623-626
Rousseau David 595-598
Rousseau Lionel 253-256
Roux Christian 291-294
Roux Philippe 449-452
Roux Stephane 549-552
Roux Sébastien 193-196
Roviras Daniel 849-852
Rémy Frédérique 563-566
Saadani Ahmed 843-848
Sahli Kheira 583-586
Said Ghada 287-290
Saidi Zineb 73-76, 77-80, 559-562
Salamatian Kavé 277-278
Saloux Eric 537-540
Salzenstein Fabien 905-908
Saoudi Samir 795-798
Sayadi Bessem 835-838
Scalart Pascal 49-52
Scalise Sandro 209-212
Schang Daniel 185-188
Scheers Bart 643-646
Scherrer Antoine 279-282
Schmid Cordelia 711-714
Schmit Renaud 771-774
Schmitt Emmanuel 695-698
Schmitt Françoise 1173-1176
Sebag Guy 409-412
Sebbe Raphael 529-532
Sedoglavic Alexandre 173-176
Seguier Renaud 1225-1228
Sellami Noura 623-626
Semani Dahbia 933-936
Semira Hichem 1177-1180
Senhadji Lotfi 241-244
Senn Eric 779-782
Sentieys Olivier 49-52, 787-790, 811-814

Serradel Eduard 265-268
Sevestre-ghalila Sylvie 371-374
Siala Mohamed 747-750
Sicot Guillaume 405-408, 877-880
Simand Isabelle 925-928
Simard Patrice 101-104
Simon Antoine 197-200
Simon Carine 739-742
Siohan Pierre 603-606
Sira-ramirez Hebertt 571-574
Souchon Gregory 249-252
Soueres Philippe 85-88
Soulat Damien 429-432
Soule Augustin 277-278
Stark Jean-luc 303-306
Strauss Olivier 165-168
Suard Frédéric 683-686
Suraud Xavier 517-520
Sérot Jocelyn 791-794
Tafinine Farid 441-444
Taft Nina 277-278
Taralova Ina 545-548
Tauber Clovis 521-524
Terrien Jérémy 567-570
Theuillon Gwladys 575-578
Theys Céline 129-132
Thierry Eric 149-152
Thiran Patrick 269-272
Thirion Laetitia 113-116
Thirion-moreau Nadège 981-984
Thoraval Laurent 331-334
Tillier Christophe 1009-1012
Tonnin Francois 37-40
Torres Eslava Manuel 9-12
Totir Felix 453-456
Tourneret Jean-yves 105-108
Toutant Jean-luc 1193-1196
Touzri Makram 795-798
Tranquart François 249-252
Tremeau Alain 1197-1200
Treuillet Sylvie 137-140, 385-388
Trémeau Alain 1-4, 1037-1040
Uhlig Steve 315-318
Unay Devrim 481-484
Unser Michael 557-558
Uribelarea Jean-louis 393-396
Vacher Michel 485-488
Vanheeghe Philippe 217-220
Vannoorenberghe Patrick 929-932
Vanstraceele Christophe 619-622
Vautrot Philippe 17-20

Vazquez Emmanuel 591-594
Verdier François 607-610
Verly Jacques 413-416
Viard-gaudin Christian 687-690
Vieira Michelle 977-980
Villain Nicolas 707-710
Vincent Damien 477-480
Vincent Emmanuel 1245-1248
Vincent Nicole 1125-1128
Vinet Hugues 989-996
Voda Alina 401-404
Voicila Adrian 607-610
Voisin Yvon 343-346
Vozel Benoit 335-338, 667-670, 659-662
Vray Didier 363-366
Vrigneau Baptiste 857-860
Walter Eric 591-594
Wendling Fabrice 241-244, 421-424, 425-428, 433-436, 437-440
Wendt Herwig 153-156
Winzenrieth Renaud 409-412
Wissem El Bahri M. 747-750
Wolfe Patrick 973-976
Wolsztynski Eric 149-152
X. Nguyen Hung 269-272
Yang Victor 363-366
Yao Koffi 857-860
Younes Laurent 945-948
Yves Tourneret Jean 265-268
Zabré Sidkièta 881-884
Zarzoso Vicente 1069-1072
Zayen Bassem 1205-1208
Zergaïnoh Anissa 221-224
Zerubia Josiane 509-512
Zhang Yu 319-322
Zozor Steeve 89-92, 109-112
Le Carpentier Éric 1241-1244