COmpression et REprésentation des Signaux Audiovisuels

Lyon, 26-27 octobre 2010

CORESA 2010 LYON

LIRIS

http://liris.cnrs.fr/coresa2010

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE

CNRS

RhôneAlpes Région

RHÔNE LE DÉPARTEMENT

Cluster ISLE Informatique, Signal, Logiciel Embarqué — Projet LIMA — RhôneAlpes Région

Lyon 1

UNIVERSITÉ LUMIÈRE LYON 2

CENTRALE LYON

CPE LYON

imaginove

SOMMAIRE

Présentations orales

Affiches

Présentations orales

Graphe de proximité invariant en échelle pour la détection rapide d'objets

Jerome Revaud[1] Guillaume Lavoué[1] Ariki Yasuo[2] Atilla Baskurt[1]

[1] Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France
{prenom.nom}@liris.cnrs.fr

[2] CS17, Kobe University, Japon
ariki@kobe-u.ac.jp

Résumé

Une procédure d'appariement de graphes pseudo-hiérarchique dédiée à la reconnaissance d'objets est présentée dans cet article. A partir d'une image modèle, un graphe est construit automatiquement en extrayant des caractéristiques invariantes locales et en les reliant selon une règle dite de proximité. Le graphe résultant présente plusieurs propriétés intéressantes dont l'invariance en échelle, la robustesse à diverses déformations non-rigides et la linéarité du nombre d'arêtes par rapport au nombre de nœuds. Le processus d'appariement est effectué de manière hiérarchique afin d'augmenter la vitesse et les résultats de détection. Par conséquent, même un appariement entre des graphes contenant des milliers de nœuds est très rapide (quelques millisecondes). Des expériences démontrent que la méthode surpasse les détecteurs d'objets spécifiques de l'état de l'art en termes de mesures rappel-précision et de temps de détection.

Mots clefs

Détection d'objets, appariement de graphes, relaxation probabiliste, hiérarchie.

1 Introduction

L'utilisation de points d'intérêt invariants pour la reconnaissance d'objets (e.g. [1]) présente de nombreux avantages : la détection est invariante en translation, en échelle, en rotation et en occultation, cela sans augmentation significative de la complexité grâce à la puissance descriptive élevée des points d'intérêt ; l'apprentissage est inexistant ; ces méthodes sont proches du temps réel ; et enfin elles sont simples à mettre en oeuvre.

Cependant, extraire les points d'intérêt est une chose ; mais détecter l'objet complet en est une autre. On peut grossièrement distinguer deux catégories de méthodes pour ce faire : (1) Les méthodes utilisant une transformation globale, et (2) les méthodes issues de l'appariement de graphe (i.e. utilisant une transformation locale). Jusqu'à présent, les méthodes utilisant une transformation globale ont produit des résultats très convaincants [2, 1, 3, 4]. Idéalement, une transformation projective 3D devrait être systématiquement utilisée, mais le trop grand nombre de paramètres requis conduit souvent à se servir d'une transformation simplifiée (par exemple affine dans [1]) pour approximer la réalité. Un autre problème des transformations globales est leur incapacité à traiter les déformations non rigides, comme par exemple avec un magazine ou un visage. D'un autre côté, l'appariement de graphes apparait comme une solution logique : après avoir extrait quelques points saillants, à la fois l'objet modèle et la scène peuvent être représentés sous forme de graphes. En outre, la comparaison de couples de sommets ou d'arêtes à une échelle locale évite la nécessité d'une transformation globale et donne dans le même temps plus de flexibilité au modèle [5]. En somme, le seul problème avec l'appariement de graphe est qu'il est NP-complet. Malgré tout, les méthodes de relaxation comme [6, 7], historiquement assez vieilles, sont rapides et restent compétitives dans la pratique [8] même si aucune garantie théorique n'assure leur convergence. Puisque nous nous concentrons ici sur une sous-classe de problèmes pour lesquels nous disposons de caractéristiques locales invariantes, la détection peut encore être optimisée grâce à une hiérarchie exploitant les informations complémentaires fournies par les caractéristiques (i.e. leur orientation et leur échelle).

En effet, les hiérarchies se sont révélées être un moyen efficace de réduire la charge de calcul en répartissant les contraintes spatiales sur plusieurs niveaux d'échelle, ce qui améliore en plus la robustesse aux variabilités intra-classe [9]. Christmas et al. [6] ont décrit en 1995 une méthode de relaxation probabiliste qu'ils ont ensuite adapté en une pseudo-hiérarchie dans un article connexe [10]. Elle présente plusieurs avantages : le cadre est minimaliste et simple à utiliser, la méthode est robuste au bruit, et l'algorithme converge rapidement (généralement en moins de 5 itérations). Malheureusement, la hiérarchie mise en oeuvre était simpliste et difficile à généraliser car l'objet recherché ne devait être présent qu'une seule fois dans la scène, le nombre de niveaux utilisés (i.e. 2) était minimal et non-modifiable, et elle nécessitait malgré tout d'utiliser une transformation globale. Même si nous avons utilisé le même cadre théorique, notre méthode est parfaitement adaptée à la détection multi-objets et étend la hiérarchie à un nombre arbitraire de niveaux, sans transformation glo-

bale.
La suite de l'article est organisée comme suit : nous commençons par présenter brièvement la théorie originelle de Christmas et al. [6]. Ensuite, nous introduisons la notion de graphe de proximité dans la section 3. La procédure d'appariement pseudo-hiérarchique est décrite en détail dans la section 4. Enfin, nous démontrons l'efficacité de la méthode dans la section 5 et concluons en section 6.

2 Relaxation Probabiliste

Dans cette section, nous résumons pour le lecteur le cadre probabiliste développé par Christmas et al. dans [6]. Soit deux graphes complets G^m et G^s (respectivement, le graphe modèle le graphe scène), l'objectif de l'appariement est de trouver la meilleure correspondance entre chaque sommet du modèle et de la scène. Dans notre formalisme, $G = (V, E, X)$ où E représente l'ensemble des arêtes, V l'ensemble des sommets et X l'ensemble de leurs mesures unaires associées (dans notre cas, un descripteur SIFT). Le cas de l'isomorphisme de sous-graphes est traité en ajoutant le nœud nul $v_0 \in V^m$ au graphe modèle. En d'autres termes, tous les nœuds étrangers au modèle dans la scène sont tout simplement étiquetés nuls.
Comme dans des travaux similaires, la méthode a besoin de deux types de mesures probabilistes pour estimer la probabilité de correspondances entre les noeuds de la scène et du modèle : (a) la probabilité $p(u_\alpha \leftarrow v_i|x_\alpha)$ d'un appariement noeud-à-noeud en utilisant les attributs unaires uniquement ($u_\alpha \in V^s$, $x_\alpha \in X^s$ et $v_i \in V^m$), et (b) une fonction de compatibilité entre arêtes qui décrit l'affinité entre deux paires locales présumées :

$$p(e_{\alpha\beta}|u_\alpha \leftarrow v_i, u_\beta \leftarrow v_j) \tag{1}$$

avec $e_{\alpha\beta} \in E^s$. Après avoir initialisé les probabilités avec la mesure (a), la relaxation itère jusqu'à convergence selon la règle de mise à jour suivante :

$$p^{(n+1)}(u_\alpha \leftarrow v_i) = \frac{p^{(n)}(u_\alpha \leftarrow v_i)Q^{(n)}(u_\alpha \leftarrow v_i)}{\sum_{v_j \in V^m} p^{(n)}(u_\alpha \leftarrow v_j)Q^{(n)}(u_\alpha \leftarrow v_j)} \tag{2}$$

où

$$Q^{(n)}(u_\alpha \leftarrow v_i) = \prod_{u_\beta \in V^s \setminus u_\alpha} \sum_{v_j \in V^m} p^{(n)}(u_\beta \leftarrow v_j) p(e_{\alpha\beta}|u_\alpha \leftarrow v_i, u_\beta \leftarrow v_j). \tag{3}$$

Pour plus de détails, nous renvoyons le lecteur à l'article original [6].

3 Graphe de proximité

Bien que Christmas et al. [6] aient formulé le problème d'appariement avec des graphes complets (i.e. $\forall i \neq j$, $v_i, v_j \in V \times V \Rightarrow e_{ij} \in E$), cela n'est habituellement pas faisable en terme de complexité. Un point très important pour notre système est donc d'être en mesure d'assouplir les contraintes spatiales entre des éléments éloignés. Curieusement, cela reste compatible avec le mécanisme de relaxation de [6] à condition que nous forcions la fonction de densité à valoir zéro lorsque l'arête n'existe pas :

$$\begin{cases} \forall e_{ij} \notin E^m, & p(e_{\alpha\beta}|u_\alpha \leftarrow v_i, u_\beta \leftarrow v_j) = 0 \\ \forall e_{\alpha\beta} \notin E^s, & p(e_{\alpha\beta}|u_\alpha \leftarrow v_i, u_\beta \leftarrow v_j) = 0 \end{cases} \tag{4}$$

Ainsi, on définit simplement le graphe de proximité comme un graphe dans lequel les caractéristiques lointaines ne sont pas connectées. Formellement, nous limitons l'ensemble des arêtes à :

$$E = \left\{ e_{ij} \,\middle|\, \forall i, j \; \frac{\|\mathbf{p}_i - \mathbf{p}_j\|}{\sqrt{\sigma_i \sigma_j}} < \chi \right\} \tag{5}$$

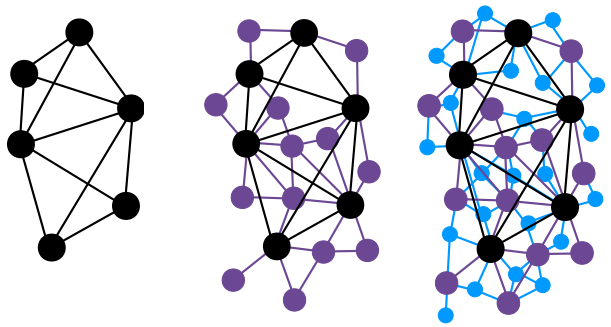
où $\mathbf{p} = (p_x, p_y)$ dénote la position d'un point d'intérêt, σ son échelle et χ est une constante. Cette définition induit plusieurs propriétés intéressantes pour notre application :

- La topologie du graphe est indépendante de l'échelle, c'est à dire que les structures du graphe modèle et du graphe de scène sont invariantes à la taille de l'objet dans l'image.
- Chaque arête du graphe représente une connexion stable. En effet, du point de vue d'un point d'intérêt, le bruit sur la position relative des autres points d'intérêt augmente avec leur distance dans l'espace-échelle pyramidale (i.e. les points plus gros paraissent plus proches).
- Le graphe de proximité permet de réduire sensiblement la charge de calcul tout en améliorant dans le même temps les performances de détection (section 5).
- Globalement, le graphe présente une structure hiérarchique centralisée (voir figure 1.(c)). Cela est du au fait que les patches plus grand possèdent plus de connexions.
- Aucune contrainte de planarité n'est imposée. Contrairement à une triangulation de Delaunay classique [8], notre graphe n'est pas affecté par la disparition de noeuds due au bruit.

4 Appariement pseudo-hiérarchique

Globalement, l'appariement de graphe est traité par une approche descendante qui commence par l'échelle la plus grossière et termine avec la plus petite (contrairement aux vraies approches hiérarchiques). Pour chaque niveau d'échelle, la relaxation probabiliste est exécutée afin de déterminer la meilleure correspondance possible entre un sous-ensemble du graphe modèle et un sous-ensemble du graphe de scène. Grâce à cette restriction, notre méthode est très rapide. L'algorithme complet est détaillée dans l'algorithme 1, mais nous détaillons maintenant les différentes étapes.

4.1 Décomposition du graphe



(a) G_1^m (b) G_2^m (c) $G_3^m = G^m$

FIGURE 1 – *Décomposition du graphe modèle (ici, 3 niveaux). Les caractéristiques plus petites sont incorporées au fur et à mesure.*

Tout d'abord, nous décomposons le graphe modèle en un ensemble de sous-graphes $\{G_l^m\}_{l=1}^L$ en se basant sur l'échelle des points d'intérêt. Pour chaque niveau l, seuls les éléments dont l'échelle est supérieure à un seuil s_l sont conservés (pour le noeud nul, $\sigma_0 = \infty$ par convention). Plus précisément, les seuils sont définis de telle sorte que le plus gros soit égal à une fraction $\rho \in [0,1]$ du rayon w_{obj} de l'objet modèle, et le plus petit à l'échelle minimum σ_{min} :

$$s_l = \sigma_{min} \left(\frac{\rho . w_{obj}}{\sigma_{min}} \right)^{\frac{L-l}{L-1}}$$

Par conséquent, $G_l^m = (V^{m,l}, E^{m,l}, X^{m,l})$ avec$V^{m,l} = \{\forall i \, v_i^m | \sigma_i^m > s_l\}$ (et ainsi de suite pour $E^{m,l}$ et $X^{m,l}$). Un exemple d'une telle décomposition est présenté dans la figure 1. Notez que la topologie graphique ne change pas à travers les niveaux, i.e. $\forall l < l', E^{m,l} \subseteq E^{m,l'} \subseteq E^m$.

4.2 Graphe d'association

Comme dans d'autres articles traitant de sujets similaires [11, 12], nous introduisons la notion de graphe d'association pour décrire l'espace discret des hypothèses de correspondance entre les nœuds du modèle et de la scène.
Formellement, le graphe d'association $A = (V^A, E^A, X^A, Y^A)$ représente les hypothèses candidates examinées durant l'appariement ainsi que leurs relations d'influence réciproque. Ici, V^A est l'ensemble des hypothèses, $X^A = \{p^{(n)}\}$ les probabilités correspondantes estimées à l'itération n, E^A l'ensemble des arêtes et Y^A leur poids associé issu de l'équation (1). Une illustration d'un tel graphe est donné dans la figure 2.(a). Dans la suite de cet article, nous désignons par une hypothèse $h_{i\alpha} \in V^A$ un couple entre un nœud du modèle et un noeud de la scène $h_{i\alpha} = (v_i, u_\alpha)$ et une hypothèse nulle par $h_{0\alpha} = (v_0, u_\alpha)$.
Avant d'expliquer comment construire V^A et E^A à partir du modèle et du graphe scène, nous allons maintenant décrire un ensemble d'opérations communes à tous les niveaux hiérarchiques, exécutées sur le graphe d'association, avant, pendant et après le processus de relaxation :

Élagage dynamique du graphe. Pour augmenter encore les performances, le graphe d'association est élagué à chaque itération de relaxation en éjectant les hypothèses pour lesquelles le noeud scène associé correspond au nœud nul avec une certaine confiance (généralement, plus de 99,9%).

Extraction des détections. Enfin, après l'achèvement du processus de relaxation, le graphe d'association est traité pour en extraire les détections. Premièrement, nous appliquons la règle du MAP pour chaque nœud de la scène, i.e. nous éliminons toute hypothèse non-maximale en terme de probabilité a posteriori. De plus, chaque hypothèse nulle est également supprimé. Il reste un ensemble de composantes connexes $\{C_k = \{h_{i\alpha}\}\}$, chacune d'elles représentant une détection unique dans l'image scène. Notez que le sous-graphe modèle $C_k^m = \{v_i\}$ et le sous-graphe scène $C_k^s = \{u_\alpha\}$ dérivés de C_k sont également connexes dans leur graphe respectif étant donné la construction du graphe d'association (éq. (4), voir section suivante).

4.3 Initialisation de l'appariement

Le sous-graphe grossier G_1^m est utilisé pour l'appariement initial. Comme ce graphe contient un petit nombre de caractéristiques, le calcul est presque instantané. Nous détaillons ici les opérations nécessaires :

Génération des hypothèses. Avant le processus de relaxation, les attributs unaire des nœuds sont utilisés pour fixer les probabilités de départ :

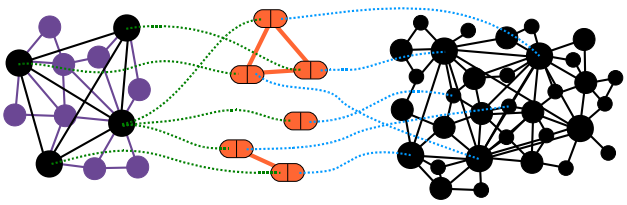
$$\begin{aligned} p^{(0)}(u_\alpha \leftarrow v_i) &= p(u_\alpha \leftarrow v_i | \mathbf{x}_\alpha) \\ &= \frac{p(\mathbf{x}_\alpha | u_\alpha \leftarrow v_i) p(u_\alpha \leftarrow v_i)}{\sum_{v_j \in V^m} p(\mathbf{x}_\alpha | u_\alpha \leftarrow v_j) p(u_\alpha \leftarrow v_j)} \end{aligned}$$

avec $p(u_\alpha \leftarrow v_i) = cste$ puisqu'on ne peut pas l'estimer, et :

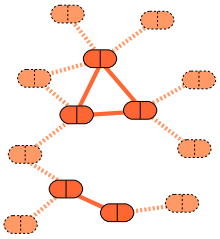
$$p(\mathbf{x}_\alpha | u_\alpha \leftarrow v_i) = \begin{cases} \phi_i(\mathbf{x}_\alpha) & \text{si } \phi_i(\mathbf{x}_\alpha) > \varepsilon_1, \\ 0 & \text{sinon.} \end{cases} \quad (6)$$

Dans le cas où $p^{(0)}(u_\alpha \leftarrow v_i)$ est nulle, alors l'hypothèse n'est pas considérée. Nous avons supposé que le bruit de mesure sur les descripteurs SIFT suit une distribution gaussienne, c'est-à-dire $\phi_i(\mathbf{x}_\alpha) = \mathcal{N}(x_\alpha; x_i, \Sigma)$ avec une variance uniforme. En outre, si v_i est le noeud nul, alors on impose $p(\mathbf{x}_\alpha | u_\alpha \leftarrow v_0) = \eta_1$ (voir section 5.1 pour savoir comment régler ε_1 et η_1).

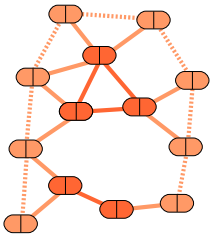
Génération des arêtes. En regardant l'éq. (2), on voit que deux hypothèses ne doivent être connectées que si leur compatibilité d'arêtes n'est pas nulle. Puisque nous avons déjà forcé la compatibilité à être nulle pour chaque paire d'hypothèses dont les noeuds correspondants ne sont pas



(a) Exemple d'un appariement initial (grossier)



(b) Ajouter des nouvelles hypothèses



(c) et les connecter.

FIGURE 2 – *(a) Illustration du graphe d'association (noeuds oranges) entre le graphe modèle (à gauche) et le graphe scène (à droite). (b), (c) : Algorithme de mise à jour (voir le texte).*

liés dans le graphe modèle ou dans le graphe de scène, par définition de (4), il suffit de simplement itérer sur chaque arête du modèle e_{ij} et chaque arête de la scène $e_{\alpha\beta}$, en reliant à chaque fois les hypothèses $h_{i\alpha}$ et $h_{j\beta}$ (à noter que le nœud nul est connecté à tous les autres noeuds dans le graphe modèle, y compris lui-même), afin d'initialiser complètement E^A.

En pratique, la compatibilité d'arêtes $p(e_{\alpha\beta}|u_\alpha \leftarrow v_i, u_\beta \leftarrow v_j) = y_{i\alpha,j\beta} \in Y^A$ est estimée en extrayant 4 invariants de e_{ij} et $e_{\alpha\beta}$:

– La longueur de l'arête $e^{(1)}_{\alpha\beta} = \|\mathbf{p}_\alpha - \mathbf{p}_\beta\| / (\sigma_\alpha + \sigma_\beta)$,

– l'angle de l'arête $e^{(2)}_{\alpha\beta} = \theta_{\alpha\beta} - \theta_\alpha$,

– la différence d'échelle $e^{(3)}_{\alpha\beta} = |\sigma_\alpha - \sigma_\beta| / max(\sigma_\alpha, \sigma_\beta)$ et

– La différence d'angle $e^{(4)}_{\alpha\beta} = \theta_\alpha - \theta_\beta$

où θ désigne l'orientation d'un point d'intérêt ou d'une arête. Nous avons supposé quatre distributions gaussiennes indépendantes par rapport au descripteur d'arête $\left\{e^{(n)}_{ij}\right\}_{n=1}^{4}$ pour calculer la compatibilité finale. De même qu'avant, si le résultat est inférieur à un seuil constant ε_2, l'arête est ignorée, et lorsque l'arête modèle contient le noeud nul, le résultat vaut η_2 (voir la section 5.1).

4.4 Mise à jour du graphe d'association

Après la première relaxation en utilisant G^m_1, on obtient un ensemble de composantes connexes, chacune correspondant à une détection localisée dans l'image scène. La plupart de ces composantes ne contiennent qu'une seule paire, i.e. un descripteur de la scène était semblable à un descripteur du modèle, mais aucun autre paire en accord n'a été trouvée dans le voisinage. Nous estimons que ces détections sont insuffisantes et les éliminons.

Puis, le reste de l'algorithme de mise à jour consiste à raffiner itérativement le modèle (à savoir ajouter les caractéristiques plus petites du modèle) en élargissant les composantes connexes dans le graphe de scène (à savoir essayer d'ajouter les voisins). L'étape d'expansion est elle-même divisée en deux étapes : d'abord, ajouter de nouvelles hypothèses impliquant les voisins de noeuds détectées (fig. 2.(b)) et, ensuite, pour relier les nouvelle hypothèses entre elles (fig. 2.(c)). La procédure complète est résumée dans l'algorithme 1.

Algorithm 1 Algorithme complet de la procédure d'appariement pseudo-hiérarchique.

Initialisation (niveau $l = 1$) :

1. Pour chaque $v_i \in V^m_1$ et pour chaque $u_\alpha \in V^s$:
   Essayer de générer une hypothèse $h_{i\alpha}$ (section 4.3).
2. Pour chaque $e_{ij} \in E^m_1$ et Pour chaque $e_{\alpha\beta} \in E^s$:
   Si $h_{i\alpha} \in V^A$ et $h_{j\beta} \in V^A$: essayer de générer une arête entre elles (section 4.3).

Mise à jour : Pour chaque $l \in [2..L]$:

1. Répéter R fois (nombre d'itération de relaxation) :
   - Exécuter une itération de relaxation (éq. (2)).
   - Élaguer le graphe d'association (section 4.2).
2. Appliquer le MAP et extraire l'ensemble de composantes connexes $\{C_k\}_{k=1}^{C}$ (section 4.2).
3. Si $l = L$: sortir et retourner l'ensemble des $\{C_k\}$.
4. Créer une liste vide T.
5. Pour chaque composante connexe C_k, $k \in C$ (section 4.4) :
   Calculer l'ensemble des noeuds voisins dans la scène $N^s_k = \{u_\beta \in V^s_l | u_\alpha \in C^s_k, u_\beta \notin C^s_k, e_{\alpha\beta} \in E^s\}$.
   Pour chaque $u_\beta \in N^s_k$ et chaque $v_j \in V^m_l$:
   - Essayer de générer une nouvelle hypothèse $h_{j\beta}$.
   - Si succès : connecter $h_{j\beta}$ avec C_k et ajouter $h_{j\beta}$ à T.
6. Pour chaque hypothèse $h_{j\beta} \in T$ (section 4.4) :
   Pour chaque v_k voisin de v_j et chaque u_γ voisin de u_β :
   Si $h_{k\gamma} \in T$: ajouter une arête entre $h_{j\beta}$ et $h_{k\gamma}$

5 Expérimentations

5.1 Apprentissage des paramètres

Paramètres indépendants Le cadre probabiliste développé par Christmas et al. [6] ne nécessite pas d'hyperparamètres (contrairement à RANSAC, par exemple). Toutefois, nous avons à apprendre à la place les constantes ε_1, ε_2, η_1 et η_2 au cours d'une phase de pseudo-apprentissage indépendante du modèle.

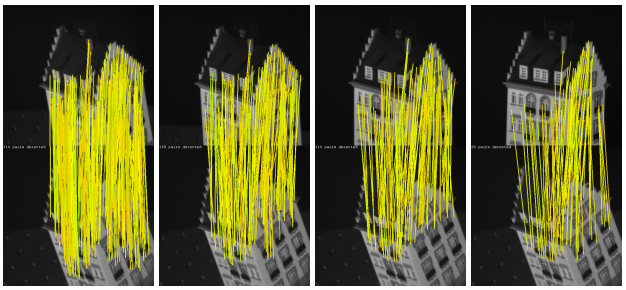
Concrètement, nous avons calibré le seuil ε_1 (éq. (6)) de manière à éliminer 99% des hypothèses candidates. C'est plutôt généreux, puisque cela équivaut virtuellement à utiliser un dictionnaire visuel de seulement $1/1\% = 100$ mots. Pour cela, nous avons extrait un grand nombre de descripteurs SIFT dans des images naturelles et avons effectué des comparaisons aléatoires. Ensuite, η_1 a été fixé à l'espérance

de la formule (6) lorsque deux descripteurs aléatoires sont utilisés, car cela correspond à une comparaison entre un descripteur connu et un inconnu (le nœud nul).
Pour fixer la valeur de η_2, nous avons supposé une répartition uniforme sur les intervalles des quatre invariants (section 4.3), respectivement 2, 2π, 1 et 2π, de sorte que $\eta_2 = 1/(8\pi^2)$. Nous avons alors fixé arbitrairement le seuil ε_2 à $\eta_2/10$.
Enfin, le nombre d'itérations de relaxation R a été fixé à 2 sans observer de perte notable de performances, preuve que le processus de relaxation converge très rapidement.

Paramètres dépendants du modèle Le paramètre χ contrôle le compromis entre un graphe de proximité densément connecté et une rapidité de détection élevée. En conséquence, nous fixons ce paramètre à sa valeur minimale à condition que les caractéristiques du modèle sont suffisamment connectées (i.e. $|E^m| \,/\, |V^m| \approx 8$). Dans la plupart des cas, une valeur de $\chi = 1$ produit de bons résultats lorsque σ correspond au rayon d'un patch SIFT. L'influence de ρ et L est étudiée dans les expériences suivantes.

5.2 Robustesse aux déformations 3D

La robustesse au changement de point de vue 3D est présentée à travers la base CMU-hotel [13]. Nous avons comparé des paires d'images séparées par un nombre d'images allant de $\Delta = 20$ à $\Delta = 80$ en utilisant les points d'intérêt SIFT. Les résultats de la figure 3 montrent que la méthode proposée réussit à reconnaître les points présents sur les différentes façades malgré un changement de point de vue important.



$\Delta = 20$	$\Delta = 40$	$\Delta = 60$	$\Delta = 80$
314 paires	158 paires	116 paires	69 paires

FIGURE 3 – *Résultats d'appariements entres des paires d'images séparées par Δ frames de la base CMU-hotel [8] (des points SIFT sont utilisés au lieu de balises posées manuellement). L'approche proposée reste robuste à un changement de point de vue 3D important.*

5.3 Comparaison avec des méthodes existantes

Étant donné que notre méthode est à cheval sur deux domaines (à savoir, l'appariement des graphes et la détection d'objets), il est difficile de se comparer avec des méthodes d'appariement de graphes existantes. En effet, notre algorithme nécessite l'existence, pour chaque nœud d'une échelle et d'une orientation - en plus de leur position dans l'espace et de leur descripteur. En outre, nos graphes modèle et scène doivent avoir une structure spécifique (i.e. un graphe de proximité). Malheureusement, ces conditions ne sont pas remplies dans la plupart des base de test, comme par exemple la base CMU-hotel [13, 8] (30 points d'intérêt manuellement définis, échelle non disponible). Au lieu de cela, nous avons comparé contre certaines méthodes de détection d'objets plus traditionnelles de l'état de l'art :
– un RANSAC basique [2] (avec une homographie)
– Locally Optimized RANSAC (LO-RANSAC) [14, 4] (similitude 2D suivie d'une homographie)
– la méthode de Lowe [1] (vote/Hough suivi d'une transformation affine)

Base d'évaluation. Nous avons filmé deux courtes vidéos avec une caméra Sony Handycam (720x480 px). Comme les vidéos ont été prises dans des conditions réalistes pour un robot d'intérieur, les vidéos contiennent naturellement une variété de bruits divers, dont du flou de bougé, de l'entrelacement vidéo, un éclairage criard. Les vidéos ont été échantillonnées pour obtenir un ensemble de 400 images (1160 noeuds par graphe scène en moyenne). Deux objets ont été utilisés pour évaluer notre méthode (fig. 5), chacun d'eux apparaissant environ 200 fois dans l'ensemble de données de test. Une image en gros plan de chaque objet a été utilisée pour construire le modèle (respectivement 225 et 1093 sommets dans les graphes modèles) et pour initialiser les autres méthodes.

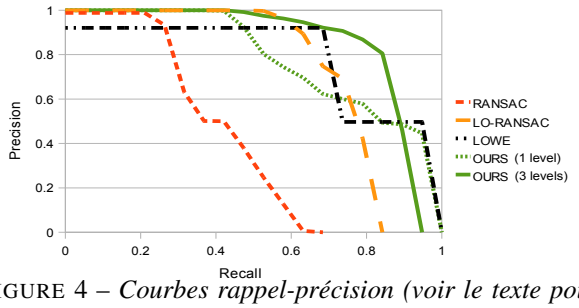


FIGURE 4 – *Courbes rappel-précision (voir le texte pour plus de détails).*

Résultats expérimentaux. Les résultats sont indiqués dans la fig. 4 en termes de courbes rappel-précision. Précision et rappel sont définis comme N_c/N_d et N_c/N_g, respectivement, avec N_c le nombre de détections correctes, N_d le nombre total de détections et N_g le nombre de boîtes de la vérité terrain (le plus haut est la courbe, meilleur est le résultat). Nous avons utilisé un seuil sur la cardinalité des composantes connexes (i.e. nombre de paires) pour générer les courbes avec notre méthode. Quelques exemples de détection sont présentés dans la fig. 5.
Globalement, la méthode proposée surpasse les autres. Nous expliquons ce fait principalement par notre procédure hiérarchique et par la distance utilisées entre les points d'intérêt. En effet, l'utilisation d'une hiérarchie (courbe "OURS 3 levels" dans la fig. 4) augmente notoirement les performances par rapport à la même méthode sans hié-

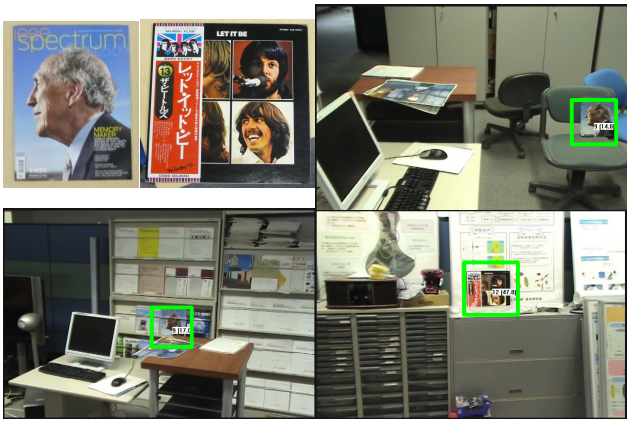


FIGURE 5 – *Objets modèles (en haut à gauche) et exemple de détections (seuil fixé à 95% de précision).*

rarchie (courbe "OURS 1 level"). Par ailleurs, une distance absolue entre keypoints est plus robuste au bruit, bien qu'elle génère plus d'hypothèses de paires.

Influence des paramètres L et ρ. Nous avons tour à tour fait varier ρ et L, chaque fois en fixant l'autre paramètre à sa valeur optimale. Fait intéressant, les performances de détection maximale sont atteintes pour des valeurs intermédiaires de ρ et L, à savoir entre 0.2 et 0.3 pour ρ et entre 3 et 6 pour L. Pour des valeurs élevées de ρ, il ne reste plus assez de caractéristiques dans G_1^m et la détection devient logiquement impossible. Le nombre de niveaux n'a pas une grande importance tant que $L \geq 3$, fixer L à 3 semble donc être le meilleur compromis car le temps de détection augmente linéairement avec L.

Temps d'exécution. Nous avons mesuré les temps moyens de traitement pour détecter les deux objets du modèle (deux graphes modèle de 225 et 1093 noeuds contre un graphe scène de 1160 nœuds en moyenne) avec différents niveaux d'optimisation :
– avec des graphes complets [1] comme dans [6] : $\approx 10^5$ s,
– avec des graphes de proximité ($L = 1$) : 2,58 s,
– avec des graphes de proximité et une pseudo-hiérarchie ($\rho = 0.2$, $L = 3$) : 0,027 s.
Comme on peut le remarquer, il y a une différence de 5 ordres de grandeur entre la première et la deuxième option, et encore un écart de 2 ordres de grandeur entre la deuxième et la troisième option. Au final, la vitesse de détection de l'article original [6] a été améliorée d'un facteur 10^6. En outre, notre méthode semble très compétitive par rapport aux détecteurs de l'état-de-l'art qui affiche chacun des temps de détection moyens de 100 ms environs.

6 Conclusion

Nous avons montré qu'une relaxation pseudo-hiérarchique peut être efficace en termes de temps de calcul et de performances de détection. Elle surpasse plusieurs méthodes de l'état de l'art en termes de courbes rappel-précision, et le temps de détection a été réduit de plusieurs ordres de grandeurs par rapport à l'approche originale grâce au graphe de proximité et à une procédure d'appariement multi-niveau novatrice. Nous pensons que ces résultats sont très encourageants et nous essayerons d'améliorer cet aspect ainsi que d'étendre notre méthode à la reconnaissance de classes d'objets dans des travaux futurs.

1. Ce résultat a été extrapolé à partir du nombre de noeuds et d'arêtes dans le graphe. Pour référence, un appariement entre deux graphes complets de 163 et 120 sommets prend environ 13 s.

Références

[1] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2) :91–110, 2004.

[2] Martin A. Fischler et Robert C. Bolles. Random sample consensus : a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6) :381–395, 1981.

[3] Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, et Jean Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *International Journal of Computer Vision*, 66(3) :231–259, 2006.

[4] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, et Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. Dans *Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[5] V. Ferrari, T. Tuytelaars, et L.J. Van Gool. Simultaneous object recognition and segmentation by image exploration. Dans *ECCV*, 2004.

[6] William J. Christmas, Josef Kittler, et Maria Petrou. Structural matching in computer vision using probabilistic relaxation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17 :749–764, 1995.

[7] S. Gold et A. Rangarajan. A graduated assignment algorithm for graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18 :377 – 388, 1996.

[8] Tiberio S. Caetano, Julian J. McAuley, Li Cheng, Quoc V. Le, et Alex J. Smola. Learning graph matching. Dans *International Conference on Computer Vision*, 2007.

[9] Boris Epshtein et Shimon Ullman. Feature hierarchies for object classification. Dans *International Conference on Computer Vision*, pages 220–227, 2005.

[10] W. J. Christmas, J. Kittler, et M. Petrou. Matching of road segments using probabilistic relaxation : Reducing the computational requirements. Dans *Sensing, Imaging and Vision for control and guidance of aerospace vehicles, volume SPIE 2220*, pages 169–179, 1994.

[11] Sergey Melnik, Hector Garcia-Molina, et Erhard Rahm. Similarity flooding : A versatile graph matching algorithm and its application to schema matching. Dans *ICDE*, 2002.

[12] Lorenzo Torresani, Vladimir Kolmogorov, et Carsten Rother. Feature correspondence via graph matching : Models and global optimization. Dans *European Conference on Computer Vision*, pages 596–609, 2008.

[13] CMU 'hotel' dataset : http ://vasc.ri.cmu.edu/idb/html/motion/hotel/index.

[14] Ondřej Chum, Jií Matas, et Josef Kittler. Locally optimized ransac. *Pattern Recognition*, pages 236–243, 2003.

Décomposition des manuscrits anciens en graphèmes et construction des codes book basée sur la coloration de graphe

H. Daher[1] D. Gaceb[1] V. Eglin[1] S. Bres[1] N. Vincent[2]

[1] LIRIS (Laboratoire d'Informatique en Image et Systèmes d'information)
UMR 5205 CNRS - INSA de Lyon – 69621 Villeurbanne Cedex

[2] CRIP5 (Centre de Recherche en Informatique Paris 5)
Université René Descartes - Systèmes Intelligents de Perception - 75270 Paris Cedex

{hani.daher,djamel.gaceb1,veronique.eglin,stephane.bres}@insa-lyon.fr

Nicole.vincent@math-info.univ-paris5.fr

Résumé

Nous présentons dans cet article une nouvelle méthode d'analyse et de décomposition de l'écriture manuscrite en graphèmes pour la construction d'un code book. Les diverses techniques développées sont inspirées de méthodes empruntées à l'imagerie au sens large et de modèles mathématiques issus de la coloration de graphes. Nos approches apportent d'une part une caractérisation fine et rapide basée sur le suivi de mouvement de la plume (courbure, épaisseur, direction, etc.) et d'autre part une méthodologie d'analyse très performante pour la catégorisation des formes de base: les graphèmes. Les outils que nous avons produits permettent aux paléographes d'étudier rapidement et avec plus de précision un grand volume de manuscrits et d'extraire un nombre important de caractères spécifiques d'un individu, ou d'une époque.

Mots clefs

Clustering, Coloration de graphe, Moments de Zernike, Segmentation des manuscrits anciens.

1 Introduction

Cet article s'inscrit dans le cadre du projet ANR GRAPHEM. Il représente une contribution méthodologique applicable à l'analyse automatique des écritures manuscrites anciennes pour assister les experts en paléographie dans le délicat travail d'étude et de déchiffrage de ces écritures. Nous nous sommes intéressés en particulier aux anciens manuscrits latins du Moyen Âge qui précède la période de la Renaissance, avant l'émergence de l'imprimerie. Sur ce type de manuscrits, on se trouve confronté à plusieurs contraintes à cause du vieillissement des supports et des encres : imprégnation irrégulière des encres, plissement, déchirement, cassures et autres dégradations du papier. De plus, les règles d'exécution des écritures en paléographie sont très strictes : certaines lettres et combinaison de lettres ne peuvent être produites que selon une unique dynamique d'exécution. Il est donc nécessaire de tenir compte dans notre étude de toutes ces contraintes et ces particularités d'exécution d'écritures. Notre objectif est :

- de produire une décomposition de l'écriture en graphèmes cohérents, en évitant notamment de produire des graphèmes qui correspondraient à certains gestes de rebroussement (retour en arrière du mouvement de la plume)
- de produire une classification robuste et modulable de l'ensemble de graphèmes obtenus sous la forme d'un code book (table de similarités de graphèmes).

Nous montrerons comment ces éléments peuvent servir à la reconnaissance des styles d'écriture des manuscrits. Nous présenterons, dans les sections suivantes, les insuffisances des méthodes conventionnelles ainsi que le principe de nos approches originales conçues pour être plus adaptées aux conditions citées ci-dessus. Nos approches se basent sur :

- une décomposition en graphèmes qui est basée sur la détection de l'axe médian directement réalisée sur l'image en niveau de gris.
- une classification de graphèmes par coloration de graphe.

2 Décomposition des manuscrits en graphèmes

Cette décomposition se réalise en deux étapes :

- Suivi de tracé et détection de l'axe médian
- Découpage du tracé en graphèmes

2.1 Suivi du tracé et détection de l'axe médian

Il existe actuellement une grande variété de méthodes de suivi et de détection de l'axe médian (ou squelette) de tracé. Dans la littérature, ces méthodes peuvent être

regroupées selon quatre catégories basées sur (pour plus de détails voir [1]) :

- ***L'amincissement morphologique :*** l'amincissement consiste à retirer au fur et à mesure les points du contour de la forme, tout en préservant ses caractéristiques topologiques. Ces méthodes nécessitant une étape préalable de binarisation des images en niveaux de gris. Celle-ci conduit à une grande perte d'informations lorsque les documents sont anciens (de mauvaise qualité) [2,3] et donne des traits binaires dégradés : caractères cassés, fusionnés ou biaisés (trous, bruit). Ces dégradations déforment souvent le squelette des traits et provoquent des erreurs significatives dans l'appariement de traits [4]. Ces limites ont donné naissance à d'autres méthodes de squelettisation qui s'appliquent directement sur les images en niveaux de gris comme la méthode basée sur les champs de potentiels 2D [5]. Ce type d'approche nécessite de couteuses opérations de lissage, mais il est plus robuste aux dégradations des images. Comme exemples, nous pouvons citer :

- ***La transformée de distance :*** la carte de distances d'un objet consiste à associer à chacun de ses points sa distance au point de contour le plus proche. Les maxima locaux de la carte de distance correspondent exactement aux points du squelette de l'objet. Plusieurs distances ont été utilisées dans ce cadre (Euclidienne [6], Chamfer [7], etc.) appliquées le plus souvent sur des images binaires mais aussi sur des images en niveaux de gris.

- ***Des heuristiques :*** ces méthodes s'appliquent directement sur les images en niveaux de gris en utilisant des heuristiques pour régler un grand nombre de paramètres qui gèrent la détection de l'axe médian. Elles ont été développées à l'origine pour extraire le squelette des traits sur des empreintes digitales et leurs résultats sont nettement plus robustes et efficaces sur des images dégradées que ceux de deux familles de méthodes précédentes [3,8].

- ***La détection des contours :*** Ces méthodes utilisent les contours pour naviguer le long du trait et détecter l'axe médian par corrélation entre une ligne et ses deux bords. Dans ce cadre, une approche itérative intéressante est proposée dans [9] pour détecter l'axe médian dans des images de neurones. D'autres méthodes, fondées sur le même principe, sont utilisées dans le suivi des routes sur des images satellitaires [10,11]. Ce type de méthodes ne peut pas être facilement appliqué sur des manuscrits dégradés où les contours de traits sont souvent déformés et discontinus et le suivi de contours risque de se perdre dans des petites chaines parasites.

2.2 Notre méthode de détection de l'axe médian

Notre approche s'applique directement sur les images en niveaux de gris de manuscrits anciens. Elle s'adapte mieux aux dégradations des manuscrits comme la discontinuité des traits tout en préservant et reconstituant, dans la mesure de possible, la forme initiale des traits. Grâce à la grande diversité technique utilisée combinant les 3 premières catégories citées ci-dessus, notre méthode offre un suivi et une détection de l'axe médian plus précise. La combinaison des méthodes se déroule de la façon suivante :

- la méthode de Frangi [14] (utilisée à l'origine pour détecter des vaisseaux sanguins de propriétés comparables à celles des traits noirs de l'écriture) est appliquée directement sur l'image initiale pour mettre en évidence le tracé par rapport au fond. Le résultat subit un lissage gaussien pour enlever les discontinuités, les trous, les bruits ou les déformations qui peuvent être présents sur le tracé.
- la transformé de distance de Chamfer est appliquée sur la carte de Frangi pour déterminer d'une manière automatique le rayon (distance par rapport au point contour le plus proche) de chaque point du tracé. Ceci nous libère de tous les inconvénients liés à l'utilisation d'un rayon avec une taille fixe dans la méthode de Xu et offre à notre méthode une meilleure adaptation au changement d'épaisseur de traits.
- le suivi de l'axe médian repose, d'une part, sur une squelitisation par diffusion [21] de la carte de Frangi et d'autre par sur le principe de la méthode de Xu adapté à notre application. Nous avons amélioré la méthode de Xu, en utilisant à chaque point la combinaison de deux directions complémentaire. a) la direction géométrique pour assurer une certaine robustesse aux situations indésirables de bifurcations et de croisement. b) la direction dynamique basée sur l'intensité lumineuse pour garantir un suivi robuste au changement soudain d'orientation ou d'épaisseur le long des traits. Cette dernière utilise une fenêtre de taille dynamique pour chercher le point suivant qui va appartenir à l'axe médian. La taille de la fenêtre varie automatiquement selon que l'on se situe dans le cas d'une bifurcation, d'un croisement ou d'une ligne droite.

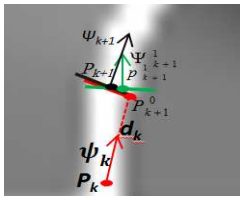


Figure 1 - *Suivi du tracé et extraction de l'axe médian*

Les étapes de notre méthode de suivi et de détection de l'axe médian sont résumées dans l'algorithme suivant :

a) ***Initialisation***

- Détecter les points de départ la carte issue de la diffusion (un point est considéré comme point de départ si son intensité lumineuse est maximale par rapport à ses voisins),
- Extraire le rayon de chaque point de départ,
- Commencer à partir de premier point de départ rencontré,
- Déterminer le point suivant P^0_{k+1} en utilisant d_k « Look ahead Distance »,
- Calculer sa direction Ψ^0_{k+1} (direction de gradient de diffusion),

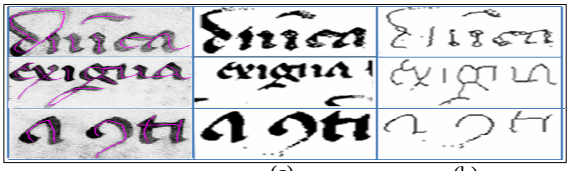
b) Déterminer le point suivant et ajuster sa position

- Tracer perpendiculairement à la direction Ψ^0_{k+1} un profil de la densité g_{k+1}du point p^0_{k+1} .
- Mise à jour de la direction dans Ψ^1_{k+1} afin de déterminer le point suivant p^1_{k+1} qui aura la valeur du rayon calculée avant,
- Procéder de la même façon par calcul du profil de densité pour avoir le point P_{k+1}, point de convergence au centre du trait. Calculer sa direction Ψ_{k+1},
- Marquer ce point comme un point visité, de cette façon il ne sera pas visité une autre fois par notre suivi de tracé.

C) ***Critères d'arrêt***

- Dans le cas où on rencontre un point de bifurcation et que ce point a déjà été visité, on arrête le suivi du tracé si on arrive à un point marqué comme point de départ, il sera enlevé de la liste des points de départ.
- -Réitération du processus le long du trait jusqu'à ce qu'un critère d'arrêt soit atteint.

La figure suivante montre le résultat de l'extraction de l'axe médian par notre méthode qui est nettement meilleur que celui obtenu par la méthode classique de Zhang. On remarque que notre méthode a bien détecté l'axe médian, même dans les situations où l'encre était dégradée ou très claire.



(a) (c) (b)

Figure 2 - *Extraction de l'axe médian par : (a) notre méthode, (b) la méthode de Zhang. (c) binarisation par la méthode de Sauvola*

2.3 Découpage de tracé en graphèmes

D'un point de vue méthodologique, la segmentation des traits en graphèmes cohérents est réalisée de la façon suivante : entre chaque point de départ et d'arrêt, tous les points impliqués dans la formation d'un trait sont sauvegardés dans une liste avec leurs directions et épaisseurs. Les points d'épaisseur minimale (minimum local) sont ensuite marqués et proposés comme point de découpage, comme cela est effectivement le cas dans la formation d'un trait. Chaque segment du tracé a une couleur différente. On identifie par cette approche les zones de croisements, les points de levers et de posers de plume (voir le zoom de la figure 3). La décomposition présentée dans la figure suivante montre que les lettres sont constituées de fragments adjacents rattachés aux points d'épaisseur minimale supposée correspondre à des points de poser et de lever de plume. Afin de construire avec robustesse un code book pour chaque page manuscrite à partir de ces graphèmes découpés, nous nous sommes intéressés à une approche innovante de regroupement des graphèmes similaires utilisant le concept de coloration de graphes, jamais exploitée dans un tel contexte. Cette méthode n'exige pas la connaissance a priori du nombre de classes, ni un apprentissage préalable et permet d'offrir une bonne homogénéité intra-classes et une bonne disparité inter-classes. Elle s'adapte mieux à la nature des graphèmes réellement présents, à l'inverse des méthodes conventionnelles qui nécessitent l'introduction préalable du nombre de classes et l'étiquetage manuel des graphèmes durant l'apprentissage en introduisant des erreurs à plusieurs niveaux d'intervention de l'utilisateur. On peut citer, à titre d'exemple, la méthode de Zhu [15] et la méthode de Kumar [16] qui sont basées sur les SVM et la méthode de Schomaker [17] basée sur la carte de Kohonen.

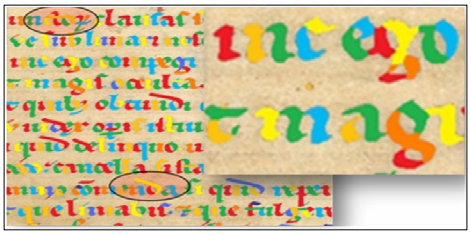


Figure 3 - *Exemple de décomposition des traits en graphèmes par notre méthode.*

3 Principe théorique de la coloration de graphes

La coloration de graphe constitue une branche très importante de la théorie de graphes. Ses applications sont nombreuses dans différents domaines scientifiques (optimisation des réseaux de transports ou de communication, des formules chimiques, …). Les définitions de la coloration sont simples et de véritables problèmes de recherche peuvent être posés sous une forme bien structurée dont la formulation peut recouvrir de grandes difficultés pratiques. Ce modèle a été introduit la première fois dans le domaine de documents par Gaceb et Eglin [18] qui l'ont adapté à toutes les étapes d'analyse des documents (de l'extraction de la structure physique et la localisation à la reconnaissance) pour consolider la coopération et assouplir les échanges d'information entre les différents modules. Grâce à sa simplicité et son potentiel en matière de classification, nous avons pu imaginer une méthode originale de construction de code book représentatif de la distribution des graphèmes de l'écriture et de leurs fréquences d'apparition.

3.1 Types de coloration existantes

Il existe ainsi plusieurs types de colorations de graphe : la coloration de sommets à laquelle nous nous sommes intéressés, la coloration d'arêtes, la coloration par liste, etc. [18,20]. Une coloration de graphe *G(V,E)* est une fonction qui affecte une couleur à chaque sommet, et qui est telle que deux sommets reliés par une arête (adjacents ou voisins) n'ont pas la même couleur (contrainte de propreté). Les couleurs (ou entiers) attribuées aux sommets du graphe servent uniquement à regrouper les sommets en classes.

3.2 Modalisation du problème de classification de graphèmes en termes de coloration

Le regroupement d'un ensemble $X=\{x_1,..,\ x_n\}$ de n graphèmes en plusieurs groupes homogènes se base sur le principe que chaque groupe doit réunir le plus de graphèmes similaires. Les regroupements portent sur un critère de similarité S. Ce critère spécifie que certaines paires de graphèmes $\{x_i,\ x_j\}$ ne peuvent être fusionnées au sein d'un même groupe. Pour résoudre ce problème de partitionnement (ou de classification), on peut partir du point de vue inverse et formuler la question suivante, à savoir : quel est le plus petit nombre de groupes homogènes que l'on peut former en respectant la contrainte S. L'intérêt de formuler le problème de cette manière, est qu'il devient alors possible de le modéliser en termes de coloration de graphe. Le positionnement du problème est alors le suivant : nous représentons chaque graphème xi par un sommet $v_i \in V$ d'un graphe simple G et nous ajoutons une arête $E(v_i,v_j)$ entre chaque paire de graphèmes dissemblables (qui ne respectent pas la contrainte S).

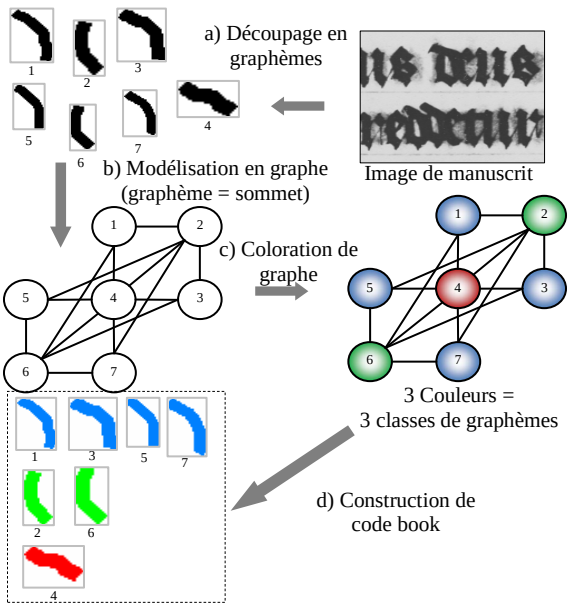


Figure 4 - *Etapes de construction de code book par coloration de graphe.*

La coloration des sommets du graphe $G(V,E)$ consiste alors à affecter à tous ses sommets une couleur de telle sorte que deux sommets adjacents (dissemblables) ne puissent pas porter la même couleur. Ces couleurs vont correspondre aux différents groupes homogènes qui constituent les différentes classes de graphèmes. Dans ce problème de regroupement, la question de la détermination du plus petit nombre de groupes homogènes, revient à rechercher le plus petit k pour lequel le graphe G correspondant admet une k-coloration : c'est donc précisément le nombre chromatique $\chi(G)$ du graphe G qu'il faut déterminer. En plus, cette modélisation présente l'avantage de gérer facilement plusieurs sortes d'ambiguïtés inhérentes à la forme de graphèmes par rapport aux mécanismes de regroupements classiques.

4 Construction du codes book

A partir de l'ensemble des graphèmes extraits dans l'étape précédente, nous procédons à la construction du code book de l'écriture (appelé aussi table de similarités).

4.1 Extraction des caractéristiques

Au départ, on produit une description vectorielle de chacun des graphèmes, qui sera ensuite traitée pour définir les critères de similarité nécessaire à la classification. Nous avons utilisé selon le besoin deux types de descripteurs : un descripteur de 15 caractéristiques topologiques et un descripteur de 25 moments de Zernike.

4.1.1 Caractéristiques topologiques

Dix caractéristiques de forme ont été extraites à partir des graphèmes binaires. La longueur et la largeur du graphème sont utilisées pour différencier les styles d'écriture ainsi que la plume utilisée. L'orientation permet de connaître l'inclinaison du graphème et de différencier les différents mouvements d'exécution d'écriture. L'excentricité permet de savoir la forme globale du graphème. L'épaisseur permet connaitre l'épaisseur de la plume et le style d'écriture. Par exemple dans les écritures gothiques l'épaisseur de la plume est grande ce qui n'est pas le cas des écritures modernes utilisant des stylos fins. Les trois dernières caractéristiques sont reliées à la surface qu'un graphème occupe dans un manuscrit par rapport aux autres.
Cinq caractéristiques de courbure (directions de la plus grande et de la plus petite courbure, les courbures Gaussienne et moyenne et le Laplacien de la courbure) sont extraites à partir des graphèmes en niveau de gris et sont calculées à partir de la matrice Hessienne [8]. Elles représentent la forme des courbes de graphèmes et reflète les propriétés structurelles telles que la convexité et concavité. Leur rôle est important, car elles sont des indicateurs essentiels permettant de différencier le style et l'époque de l'écriture.

4.1.2 Moments de Zernike

Les 25 moments de Zernike utilisés pour décrire les graphèmes sont classés parmi les moments orthogonaux (géométrique, de Legendre, etc.) car ils possèdent la propriété d'invariance à la rotation.

Le moment de Zernike d'ordre n avec la répétition m (n - $|m|$ est paire et $|m|<n$) est définie par :

$$A_{nm} = \frac{n+1}{\pi} \sum_x \sum_y f(x,y) \cdot V^*_{nm}(\rho,\theta)$$

$V_{nm}(\rho,\theta)$ est un ensemble de polynômes complexes dans l'espace à deux dimensions qui forme un ensemble orthogonal sur l'intérieur du cercle unité $(x^2+y^2=1)$, avec :

$$V_{nm}(\rho,\theta) = R_{nm}\, e^{im\theta}$$

Où ρ est la longueur du vecteur d'origine au point de coordonnées (x, y). θ est l'angle entre le vecteur ρ et l'axe des abscisses. $R_{n,m} = R_{n,-m}$ est un polynôme radial défini comme suit :

$$R_{nm}(\rho) = \sum_{S=0}^{(n-|m|)/2} (-1)^S \frac{(n-S).!}{S!\left(\frac{n+|m|}{2}-S\right)!\left(\frac{n-|m|}{2}-S\right)!}\rho^{n-2S}$$

Pour calculer les moments de Zernike, le centre de chaque graphème binaire est pris comme origine du repère et les coordonnées des pixels de l'image sont transformées de manière à être dans le domaine du cercle unité. Comme nous l'avons signalé précédemment les moments de Zernike sont invariants seulement à la Rotation. Pour les rendre invariants au changement d'échelle [19], on doit normaliser l'image binaire du graphème par le moment du premier ordre m_{00} défini comme étant la surface du graphème.

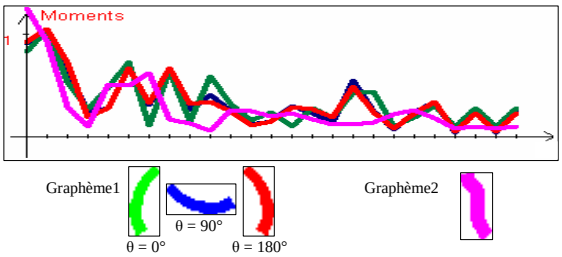


Figure 5 - *Invariance à la rotation des 25 moments de Zernike (n=8).*

4.2 Mesure de similarité

La dissimilarité entre v_i et v_j est donnée par la distance généralisée de Minkowski d'ordre α ($\alpha = 2$: distance euclidienne). $Ds = \left(\sum_{k=1}^{Nc} g_k(v_i^k, v_j^k)^\alpha\right)^{\frac{1}{\alpha}}$

Nc est la longueur des vecteurs des caractéristiques. g_k est la fonction de dissemblance qui compare les caractéristiques deux à deux.

4.3 Construction du graphe

La construction d'un graphe G à colorer à partir d'un ensemble $X=\{x_1,\ldots,x_n\}$ de n graphèmes (où chaque sommet v_i correspond au vecteur descripteur de graphème x_i) est principalement basée sur le calcul de la matrice de distances MD_s . Cette matrice traduit les dissimilarités $Ds(x_i, x_j)$ existantes entre les paires de graphèmes (x_i, x_j) données par la relation suivante : $MDs[v_i,v_j]=Ds(x_i, x_j)$ avec $i\in[1,n]$ et $j\in[1,n] \mid (i\neq j)$. Une fois v calculée, nous associons à X un graphe seuil supérieur $G_{\geq S} = (V=X, E_{\geq S})$ en utilisant la relation suivante :

$$E_{\geq S}[v_i, v_j] = \begin{cases} 1 & \text{si } Ds(x_i,x_j) = Ds(v_i,v_j) \geq S \\ 0 & \text{sinon} \end{cases}$$

Pour ne pas confondre le terme adjacence (ou voisinage) avec le terme similarité, il faut noter que deux sommets sont adjacents s'ils ont une dissimilarité supérieur au seuil S. Le seuil S est également nommé seuil d'adjacence. Ce seuil peut être ajusté manuellement à l'aide des paléographes ou automatiquement en maximisant la qualité de classification ψ donné par [18]:

$$S^{Optimal} = \arg\max(\psi(S_i))\}$$

4.4 Classification des graphèmes

Une fois le graphe G construit à partir de l'ensemble des graphèmes, on applique l'algorithme de coloration de Gaceb et Eglin [18]. Les différentes couleurs résultantes représentent les classes de graphèmes.

5 Résultats et application

Nous avons testé notre méthode de décomposition en graphèmes sur 12 pages de textes manuscrites du Moyen-âge de différents types.

Figure 6 - *Extraits des 12 manuscrits.*

Un ensemble des 4863 graphèmes répartis de la façon suivante a été extrait des douze pages manuscrites. : *{p1=343, p2=583, p3=643, p4=248, p5=398, p6=528, p7=564, p8=316, p9=499, p10=193, p11=269, 12=279}.* Cette décomposition a été soumise à la validation des experts paléographes et a obtenu leur approbation. La figure suivante montre un extrait de code book construit à partir des graphèmes de la page 12 par coloration de graphe (section 4). Il est important de noter ici que les descripteurs topologiques ont permis de mettre en évidence la dynamique dans la formation du tracé : les graphèmes qui représentent ainsi les mêmes particularités de mouvement (arrondis droits ou gauches) de la plume

sont regroupés dans une même classe. Une vérification complète produite par les experts paléographes permet de l'attester.

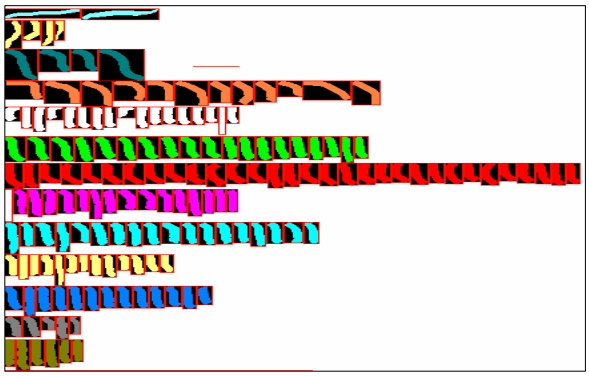


Figure 7 - *Extrait de code book de page p12.*

Pour reconnaitre le style des manuscrits : chacune des 12 pages manuscrites p_i est représentée par son code book cb(p_i) de k_i classes de graphèmes (k_i nombre chromatique de la coloration de graphe) avec $cb(p_i) = \{c_1^i, ..., c_k^i\}$. Chaque classe c_t contient n_t graphèmes avec $c_t = \{x_1^t, ..., x_{nt}^t\}$. On peut donc estimer la similarité entre chaque paire de pages (p_i, p_j) par la distance *dp* suivante :

$$dp\left[cb(p_i), cb(p_j)\right] = \sup\left[dc(c_n, c_m)\right] \mid n = 1...k_i \text{ et } m = 1...k_j$$

Avec *dc*, la distance entre deux classes de graphèmes donnée par : $dc(c_n, c_m) = \min\{Ds(x_i \in c_n, x_j \in c_m)\}$

Les distances *dp* entres les pages 12 sont illustrées dans le tableau suivant :

	1	2	3	4	5	6	7	8	9	10	11	12
1	0											
2	0.7145	0										
3	0.391	0.459	0									
4	0.3063	0.679	0.3464	0								
5	0.1559	0.596	0.3353	0.2483	0							
6	0.5464	0.391	0.7943	0.5372	0.4591	0						
7	0.127	0.422	0.3173	0.3419	0.1755	0.3867	0					
8	0.4871	0.448	0.8065	0.5029	0.5351	0.5487	0.2362	0				
9	0.7056	0.231	0.8526	0.7685	0.6276	0.4637	0.4141	0.7529	0			
10	0.8062	0.206	0.728	0.6466	0.6261	0.4637	0.1435	0.7122	0.153	0		
11	0.3437	0.597	0.2586	0.1511	0.2446	0.1863	0.208	0.3852	0.555	0.5772	0	
12	0.3413	0.603	0.3659	0.1331	0.2381	0.3868	0.1741	0.3883	0.589	0.6152	0.2295	0

Tableau 1 - *Distances entres les codes book de pages.*

On peut estimer à partir cette table que le style de la page 7 et le plus proche de celui de la page 1, que celui de la page 2 est plus proche de celui de la page 10, et ainsi de suite. Compte tenu de l'évidence visuelle facile à obtenir en observant les images de la base, cette distance est un indicateur efficace de similarités entre écriture. Ces premiers tests sont actuellement complétés par des tests en vraie grandeur sur la base paléographiques de 10000 images au sein de laquelle des différences plus difficiles à relever devront être relevées.

6 Conclusion

Nous avons présenté dans cet article une nouvelle méthode de découpage de l'écriture manuscrite en graphèmes et une construction d'un code book robuste bien adaptés aux exigences de notre domaine et basée sur la coloration de graphes. En perspective de ce travail, nous envisageons de reconnaitre automatiquement les styles et l'époque des manuscrits de la base complète paléographique. Ce travail sera intégré prochainement dans un prototype expérimental à l'usage des experts paléographes du projet GRAPHEM.

Références

[1] H. Daher et al. A New approach for centerline extraction in handwritten strokes…, International Workshop on Document Analysis Systems, Boston, June, 2010.

[2] D. Lee, S.W. Lee. A new methodology for gray-scale character segmentation and recognition, ICDAR, vol. 1, pp.524, 1995.

[3] Maio, D. Maltoni, Direct Gray-Scale Minutiae Detection in Fingerprints, IEEE Transactions on PAMI, vol. 19, n°. 1, pp. 27-40, 1997.

[4] Suh et al. Stroke extraction from gray-scale character image, Progress in Handwriting Recognition 593-598, 1997.

[5] Grigorishin et al. Skeletonisation: An Electrostatic Field Based Approach, Pattern Analysis & Applications 98, v.1, pp. 163-177.

[6] P.E. Danielsson. Euclidean Distance Mapping, Computer Graphics and Image Processing, vol. 14, pp. 227-248, 1980.

[7] A. Rosenfeld and J.L. Pfalz. Distance Functions on Digital Pictures, Pattern Recognition, vol. 1, pp. 33-61, 1968.

[8] Yaxuan Qi. Fingerprint Ridge Line Reconstruction. Intelligent Information Processing, pp. 211-220, 2004.

[9] Y. Zhang et al. A novel tracing algorithm for high throughout imaging Screening of neuron-based assays, J Neurosci Methods 160, pp. 149–162, 2007.

[10] D.Poz et al. Automated extraction of road network from medium and high-resolution images, Pattern Recognition and Image Analysis, vol. 16, n°. 2, pp 239-248, 2006.

[11] R. Peteri et al. Detection and extraction of road networks from high resolution satellite images, International Conference on Image Processing, vol.1, pp I-301-4, 2003.

[12] Y. Xu et al. An improved algorithm for vessel centerline tracking in coronary angiograms, Computer Methods and Programs in Biomedicine, Vol. 88, n° 2, Pages 131-143.

[13] T.Y. ZHANG ET C.Y. SUEN: A fast parallel algorithm for thinning digital patterns. Communications of the ACM, 27(3):236–240, mars 1984.

[14] A.F. Frangi et al. Multiscale Vessel Enhancement Filtering, MICCAI '98, pp. 130-137, 1998.

[15] G. Zhu et al. Language Identification for Handwritten Document Images Using A Shape Codebook.

[16] J. Kumar et al. Handwritten Arabic Text Zone Detection using A Shape Codebook. ICPR, 2010.

[17] L. Schomaker et al. using codebooks of fragmented connected-component contours in forensic and historic writer identification. Pattern Recognition Letters 28(6), pp 719-727, 2007.

[18] D. Gaceb et V. Eglin : Improvement of postal mail sorting system. IJDAR, 11(2):67-80, 2008.

[19] M.R.Teague, Image analysis via the general theory of moments, J.opt.soc.Am, vol.70, n°8, pp 920-930, 1980.

[20] V. PASCHOS, book, Optimisation combinatoire5: problèmes paradigmatiques et nouvelles problématiques, Lavoisier, France, pp. 270, 2007.

[21] Z. Yu and C.L. Bajaj, A Segmentation-Free Approach for Skeletonization of Gray-Scale Images via Anisotropic Vector Diffusion, CVPR'04, vol.1, pp. 415-420, June 2004.

Tatouage et Compression Conjoint dans JPEG2000 avec un Algorithme de Quantification Codée par Treillis (TCQ)

D. Goudia[1,2] M. Chaumont[1] W. Puech[1] N. Hadj Said[2]

[1] LIRMM, UMR CNRS 5506, Université Montpellier II

161 rue Ada, 34392, MONTPELLIER cedex 5, FRANCE

{dalila.goudia, marc.chaumont, william.puech}@lirmm.fr

[2] SIMPA (Laboratoire Signal Image Parole)

USTO, Université des Sciences et de la Technologie d'Oran
BP 115, El M'aouael, ORAN , ALGERIE

{min_hadj}@yahoo.fr

Résumé

Dans cet article, nous présentons une méthode conjointe de quantification et de tatouage basée sur la Quantification Codée par Treillis (TCQ). Cette technique a été intégrée au niveau du codeur JPEG2000. Plus précisément, nous cherchons à effectuer conjointement compression et tatouage à l'aide de la TCQ en réalisant simultanément la quantification et l'insertion de la marque durant la compression. Le processus d'extraction de la marque peut être mis en œuvre à la fois durant la décompression ou après celle-ci. Les résultats obtenus ont montré que ce schéma conjoint résiste à une compression JPEG2000 avec variation du taux de compression sans dégradation de la qualité de l'image décompressée.

Mots clefs

Compression d'images, tatouage numérique, JPEG2000, TCQ.

1 Introduction

Le tatouage d'image consiste à insérer (généralement sous forme invisible) une information dans une image puis à tenter de la récupérer après que l'image ait éventuellement subi des manipulations de nature variée. Il existe de nombreuses méthodes d'insertion dans la littérature. Les méthodes de tatouage informées [1, 2, 3] reposent sur l'utilisation de dictionnaires structurés et considèrent le tatouage comme un problème de codage canal. Durant l'insertion, deux phases successives sont exécutées : le *codage informé* et *l'insertion informée*. Le codage informé consiste à construire un dictionnaire de mots de codes et ensuite à trouver le mot de code $\mathbf{c}^*$ le plus proche du signal hôte $\mathbf{x}$. Ce mot de code représente le message $\mathbf{m}$ à insérer. L'insertion informée va déplacer le signal hôte $\mathbf{x}$ vers le mot de code $\mathbf{c}^*$. Le vecteur déplacement est appelé *signal de tatouage*. Cette étape consiste à modifier $\mathbf{x}$ pour le déplacer dans la région de détection introduisant le moins de distorsion. Les méthodes de tatouage par quantification sont une implantation pratique du principe de tatouage informé, et plus précisément du schéma de Costa [4]. En 1999, Chen et Wornell [1] ont proposé sous le nom de QIM (Quantization Index Modulation), l'implantation la plus courante du schéma de Costa. Eggers *et al.* [2] ont présenté sous le nom de Schéma de Costa Scalaire SCS (Scalar Costa Scheme) une implantation sous-optimale explicite du schéma de Costa similaire à celle de la DC-QIM [1].

La compression représente, non seulement un passage obligé du stockage ou du transfert d'images, mais aussi l'une des attaques les plus destructrices vis-à-vis du tatouage. La compression jointe au tatouage a suscité un intérêt récent. Associer tatouage et compression permet d'améliorer la capacité d'insertion du tatouage et d'assurer une meilleure détection tout en maintenant une bonne qualité d'image. En effet, cette approche offre de nombreux avantages qui sont très recherchés dans certaines applications telles que le contrôle d'intégrité, l'authentification ou la production de documents enrichis. La norme de compression JPEG2000 [5] a été développée par le groupe JPEG (Joint Photographic Experts Group). Ce standard offre de nombreuses fonctionnalités et se caractérise par une grande diversité des options d'encodage menant à de bons compromis compression/qualité. Le marquage d'images dans le domaine JPEG2000 a déjà fait l'objet de plusieurs travaux [6, 7, 8].

L'objectif de cet article est de concevoir un nouveau schéma permettant de combiner la compression JPEG2000 et une méthode de tatouage fondé sur la quantification. Ce schéma conjoint repose sur un module hybride de quan-

tification codée par treillis (TCQ) [9] capable de quantifier et d'insérer la marque en une seule fois. Ce module est intégré dans la chaine de codage de JPEG2000 à la place de l'étape de quantification scalaire uniforme avec zone morte. L'algorithme utilise la version TCQ telle que définie dans la partie 2 de la norme [10]. Nous pouvons extraire la marque de deux manières : durant ou bien après le processus de décompression. L'approche conjointe nécessite de prendre en compte le taux de compression obtenu après codage entropique : l'insertion de la marque ne doit pas aboutir à une dégradation sensible des performances du compresseur (taux de compression et qualité de l'image reconstruite). Il est nécessaire de maintenir un compromis débit-distorsion optimal. Le troisième critère à considérer concerne la robustesse du tatouage. La marque doit pouvoir survivre à certaines attaques, et ce, de manière à pouvoir assurer une extraction correcte de celle-ci à partir de l'image compressée/tatouée. Notons qu'actuellement, il n'existe pas de véritables schémas conjoints dans le domaine JPEG2000. Les travaux proposés dans ce domaine se contentent d'intégrer une étape supplémentaire d'insertion/extraction de la marque dans le schéma de fonctionnement de JPEG2000. Notre système permet à la fois de quantifier et de marquer les coefficients d'ondelettes en utilisant un seul et même composant.

Ce papier est organisé de la façon suivante : dans la section 2, nous passons en revue les techniques de tatouage basées-quantification combinées à JPEG2000. La section 3 présente la quantification TCQ dans JPEG2000 ainsi que les techniques de tatouage basées-quantification TCQ. Les détails du schéma conjoint sont donnés dans la section 4. La section 5 est consacrée à la discussion des premiers résultats obtenus. Enfin, la dernière section conclut le présent article.

2 Le tatouage dans JPEG2000

JPSEC (Secure JPEG2000) [11] qui représente la partie 8 du standard, propose des solutions permettant à des applications de générer, décoder et échanger des bit-stream JPEG2000 sécurisés. les mécanismes de sécurisation sont intégrés dans le bit-stream sous la forme de métadonnées. Ces informations peuvent donc être perdues lors d'opérations de conversion de format. Par contre, lorsque la marque est insérée directement durant le processus de codage, celle-ci devient partie intégrante de l'image. Un certain nombre de travaux consacrés à l'aspect compression/tatouage conjoint avec JPEG2000 dans le cadre d'une application d'authentification d'images ou de protection de copie ont été proposés. Seul un petit nombre de ces travaux utilisent une technique de tatouage fondé sur la quantification. Meerwald *et al.* [6] ont proposé une méthode de tatouage aveugle de type QIM intégrée dans le schéma de fonctionnement de JPEG2000. Schlauweg *et al.* [7] ont présenté un schéma de tatouage semi-fragile sécurisé basé sur JPEG2000 utilisant la DM-QIM (Dither Modulation QIM) et des outils cryptographiques tels que les fonctions de hachage et les méthodes de cryptographie. Makhloufi *et al.* [8] ont développé un algorithme de tatouage aveugle basé sur la QIM qui, contrairement aux travaux précédents, effectue l'insertion de la marque avant l'étape de quantification. Le codeur utilisé est un codeur compatible avec la partie 2 de la norme et les auteurs utilisent un décalage non-linéaire dans leur méthode de tatouage pour réduire la distorsion causé par l'insertion de la marque. Aucune de ces techniques ne considèrent la quantification scalaire du schéma de codage JPEG2000 comme étant un bruit connu à l'insertion. Depuis la redécouverte des travaux de Costa [4], nous savons que si nous prenons en compte le bruit connu, nous augmentons la capacité du canal. Par conséquent, nous proposons un système combiné compression/tatouage informé dans JPEG2000 prenant en compte le bruit de quantification.

3 La Quantification Codée par Treillis (TCQ)

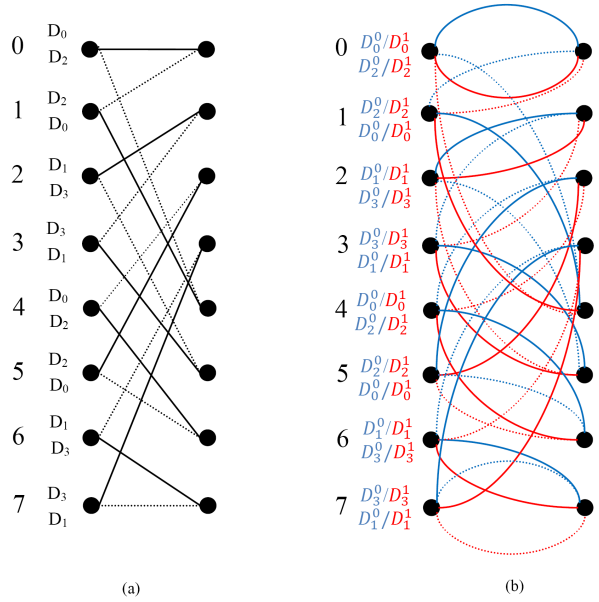


Figure 1 – *La structure du trellis utilisé : a) dans JPEG2000 b) dans notre schema de compression/tatouage conjoint.*

La Quantification Codée par Treillis (TCQ) est une technique de quantification rapide proposée par Marcelin et Fisher [9]. La TCQ est basée sur l'idée de partitionnement d'ensembles proposée par Ungerboeck [12] pour combiner la modulation et le codage de canal. Elle consiste à partitionner un dictionnaire de quantification initial en sous-dictionnaires complémentaires associés aux transitions entre les états d'un code convolutif.

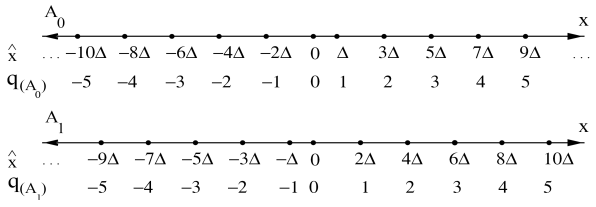


Figure 2 – *Les quantificateurs d'union de la TCQ dans JPEG2000.*

3.1 La TCQ dans JPEG2000

Plusieurs techniques de quantification sont proposées dans la partie 2 du standard. Parmi celles-ci, nous pouvons citer la TCQ [5, 10]. Cette technique permet d'atteindre de meilleures performances en termes de débit-distorsion par rapport à la quantification scalaire, et ce, à différents taux de compression. La complexité du codeur s'en trouve cependant augmentée. Le code convolutif est représenté par un treillis. Le treillis est constitué d'un ensemble d'états reliés entre eux par des arcs ou branches. Il s'agit d'une variante du diagramme de transition avec prise en compte du temps. Chaque branche dans le treillis représente une transition entre deux états à un instant donné. A partir d'un état initial, le chemin dans le treillis peut être spécifié par une séquence binaire puisqu'il n'y a que deux transitions possibles d'un état à un autre. Concernant la variante de la TCQ incluse dans la norme JPEG2000, le dictionnaire de quantification du quantificateur scalaire uniforme de pas de quantification Δ est partitionné en 4 sous-dictionnaires notés D_0, D_1, D_2 et D_3. Chaque branche du treillis est étiquetée par un sous-dictionnaire D_i. La Figure 1.a décrit la structure d'un treillis à 8 états ainsi que les sous-dictionnaires associés aux branches de ce treillis. L'union des quantificateurs associés à chaque état est appelé quantificateur d'union. Les quantificateurs d'union utilisés dans le treillis de la figure 1.a sont $A_0 = D_0 \cup D_2, A_1 = D_1 \cup D_3$. Les dictionnaires de ces quantificateurs sont illustrés au niveau de la Figure 2.
Chaque état dans le treillis est associé à l'un des deux quantificateurs d'union. Afin de quantifier la séquence source **x**, l'algorithme de Viterbi [13] est utilisé pour trouver le chemin optimal à travers le treillis. Le chemin optimal indique la séquence de bits de transition pour laquelle la distorsion totale est minimale, c'est-à-dire que l'on recherche le vecteur quantifié $\hat{\mathbf{x}}$ le plus proche du vecteur source **x** au sens de l'erreur quadratique. En plus de la séquence de bits indiquant le chemin optimal, l'algorithme de Viterbi produit également une séquence d'indices de quantification TCQ, nécessaire pour indiquer les niveaux choisis à l'intérieur des sous-dictionnaires du chemin optimal. A la réception, le décodeur reconstruit la source quantifiée comme suit : la séquence de bits indiquant le chemin optimal à travers le treillis ainsi que la séquence d'indices TCQ sont utilisées comme entrée du codeur convolutif. A chaque transition i, le bit de chemin permet de retrouver l'état suivant et donc le sous dictionnaire de quantification utilisé. L'indice TCQ permet de reconstruire l'échantillon source qui sera présenté en sortie du codeur.

3.2 La TCQ en tatouage

Les techniques de tatouage basé sur la TCQ sont peu nombreuses. Elles reposent toutes sur le principe suivant : durant l'encodage, on quantifie la séquence source x en forçant les transitions du treillis afin qu'elles correspondent au message à encoder. Le rendement de la TCQ est de 1/1 (un bit inséré pour un échantillon source). Cette approche est appelée sélection de chemin TCQ (TCQ-PS : TCQ path selection). Braci *et al.* [14] se sont focalisés sur l'aspect sécurité des schémas de tatouage informé basé sur la QIM et ont proposé une version sécurisée de la TCQ-PS. Ouled Zaid *et al.* [15] ont développé un algorithme de tatouage basé sur la Turbo TCQ dans le domaine ondelettes. Aucune de ces techniques de tatouage n'a été combinée à un codeur JPEG2000. Le principe sur lequel repose notre approche est le suivant : quantifier et marquer les coefficients d'ondelettes en même temps grâce à l'utilisation d'un module hybride de quantification TCQ. Notre technique de tatouage possède des similitudes avec celle des codes à papier sales ou DPTC [3]. Les deux méthodes reposent sur l'utilisation d'un treillis modifié associé à un dictionnaire. Cependant, nous utilisons un dictionnaire de quantification alors que Miller *et al.* [3] utilise un dictionnaire de codage structuré. De plus, l'insertion de la marque se fait différemment. Notre schéma conjoint intègre un système de tatouage informé basé sur une quantification alors que celui de Miller *et al.* est un algorithme d'insertion itératif permettant de construire pas à pas le tatouage jusqu'à atteindre les conditions de transmission robuste.

4 Le schéma conjoint de compression et de tatouage dans JPEG2000

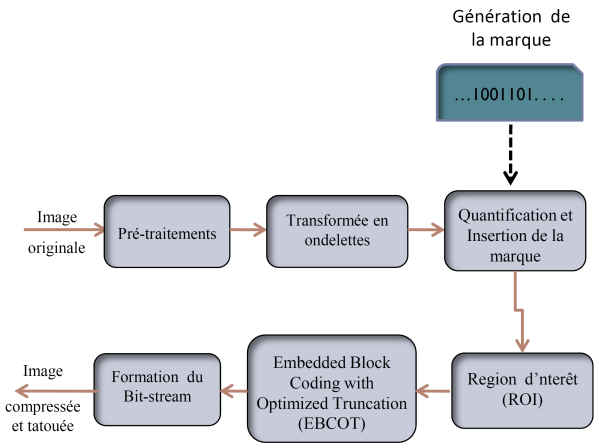


Figure 3 – *Schéma de fonctionnement du système conjoint codage JPEG2000/tatouage.*

La contribution majeure apportée par notre travail concerne l'utilisation d'un module hybride de quantification permettant de quantifier et de tatouer simultanément les coefficients d'ondelettes. Le schéma de fonctionnement de notre système de compression/tatouage conjoint est illustré sur la Figure 3. La TCQ-PS [14] ne peut pas être adaptée à notre approche conjointe car le chemin dans le treillis correspond au message inséré. La sécurité du message n'est donc pas assurée puisque le chemin fait partie intégrante du bit-stream JPEG2000. Nous proposons une nouvelle technique de tatouage basée sur la TCQ qui est indépendante de la sélection du chemin. L'insertion de la marque se fait au niveau des sous-bandes de détails LH, HL et HH issues de la décomposition en ondelettes. Le nombre de sous-bandes qui seront quantifiées/tatouées dépend du nombre de niveaux de décomposition inclus dans le processus de tatouage. La capacité est calculée à partir du nombre de sous-bandes candidates au tatouage. Les coefficients d'ondelettes appartenant aux autres sous-bandes sont quantifiés avec le quantificateur scalaire uniforme classique. Les quantificateurs tels que définis dans la norme JPEG2000 partie 2 sont remplacés par des quantificateurs décalés. Ces quantificateurs décalés possèdent le même pas de quantification Δ que les originaux mais ils diffèrent de ces derniers par l'utilisation d'un décalage $\mathbf{d} \in [-\Delta/2, \Delta/2]$ [1]. Ce décalage est obtenu aléatoirement à partir d'une clé secrète. Le principe de tatouage est le suivant : on associe à chaque bit du message à insérer un décalage **d**. Si le bit à insérer est égal à 0 alors on utilise le quantificateur $D_j^0, j = 0, 1, 2, 3$ avec le décalage $d[0]$. Si le bit est égal à 1 alors cela induit l'utilisation du quantificateur D_j^1 avec le décalage $d[1]$ remplissant la condition : $|d[1] - d[0]| = \Delta/2$. Nous disposerons de deux groupes de quantificateurs d'union dans notre treillis : le groupe 0, $A_0^0 = D_0^0 \cup D_2^0, A_1^0 = D_1^0 \cup D_3^0$ qui représente l'insertion du bit 0 et le groupe 1, $A_0^1 = D_0^1 \cup D_2^1, A_1^1 = D_1^1 \cup D_3^1$ correspondant à l'insertion d'un bit 1. Soit le message binaire **m** à insérer et le signal hôte **x**. L'insertion de la marque guide le processus de quantification c'est-à-dire qu'a chaque transition *i* dans le treillis, la valeur du bit **m**[i] va déterminer quel sera le quantificateur d'union à utiliser pour quantifier le coefficient d'ondelette **x**[i]. Les sous-dictionnaires $D_j^{m[i]}, j = 0, 1, 2, 3$ vont étiqueter les branches du treillis. Le treillis classique tel qu'utilisé dans JPEG2000 comporte deux arcs ou branches par état. A chaque branche est attribuée un bit appelé bit de chemin (0 ou 1). Dans notre treillis, le nombre d'arcs est multiplié par 2. Au niveau de la transition *i*, pour un état e et un bit de chemin donné (0 ou 1), il y a 2 arcs possibles vers un état de la transition *i*+1 comme l'illustre la Figure 1.b. A l'un des arcs est associé un sous-dictionnaire du groupe 0 (par exemple le sous-dictionnaire D_0^0 du groupe A_0^0) et à l'autre arc le sous-dictionnaire appartenant au groupe 1 (le sous-dictionnaire D_0^1 du groupe A_0^1). Avant de procéder à la quantification, le treillis est élagué de façon à supprimer toutes les branches ne correspondant pas aux sous-dictionnaires encodant le message. On se retrouvera avec un treillis à deux branches par état mais celles-ci sont associées aux quantificateurs encodant le message **m**. L'étape de quantification produit la séquence de bit de chemin **p** et la séquence d'indice de quantification TCQ notée **q** donnée par :

$$\begin{aligned} q[i] &= Q_{D_j^{m[i]}}(x[i]) \\ &= sign(x[i] - d_i[m[i]]) \lfloor \frac{|x[i] - d_i[m[i]]|}{\Delta_j} \rfloor \quad (1) \end{aligned}$$

où d_i représente le décalage du quantificateur décalé $D_j^{m[i]}$ à la transition i et Q est la fonction de quantification. En plus des séquences **p** et **q**, nous avons besoin d'une information supplémentaire qui sera codée et stockée au niveau de la séquence **l**. Celle ci va nous permettre de retrouver, durant l'étape de quantification inverse, la structure du treillis modifié utilisé lors de la quantification. Connaissant **p**, **q**, **l** et **d**, les valeurs de reconstruction $\hat{\mathbf{x}}$ sont obtenues de la manière suivante :

$$\begin{aligned} \hat{x}[i] &= \bar{Q}^{-1}_{D_j^{m[i]}}(q[i]) \\ &= sign(q[i])(|q[i]| + \delta)\Delta_j + d_i[m[i]] \quad (2) \end{aligned}$$

où $\bar{Q}^{-1}$ est la fonction de quantification inverse et δ le paramètre de reconstruction avec $0 < \delta < 1$. Le treillis utilisé est le même que celui employé lors de la quantification.

4.1 Quantification et insertion de la marque

L'insertion de la marque est effectuée durant la quantification de manière indépendante au niveau de chaque code-block [5]. Dans un premier temps, le message à insérer est généré puis codé à l'aide d'un code correcteur d'erreur afin d'augmenter la robustesse du message transmis. Pour chaque code-block, les étapes suivantes sont exécutées :

- Tirage de nombres aléatoires à l'aide d'un générateur pseudo-aléatoire initialisé par une clef secrète : ces nombres seront utilisés comme décalage **d** dans les quantificateurs décalés.
- Génération des deux groupes de quantificateurs d'union (le groupe 0 et le groupe 1) pour chaque transition *i*.
- Elagage du treillis : l'état initial est mis à 0. Nous parcourons le treillis et pour chaque transition nous supprimons les branches dont les sous-dictionnaires associés ne correspondent pas au message à encoder (figure 4).
- Quantification et tatouage : exécution de l'algorithme de Viterbi afin de trouver le chemin optimal. La figure 4 illustre un exemple d'insertion du message m = {1, 0, 1}. Le chemin en gras représente le chemin à distorsion minimal déterminé par l'algorithme de Viterbi.

4.2 Extraction de la marque à partir de l'image décompressée

L'image décompressée/marquée est transformée à l'aide d'une transformée en ondelettes discrète. Les coefficients d'ondelettes appartenant aux sous-bandes incluses dans le processus de tatouage sont sélectionnés et placés dans le

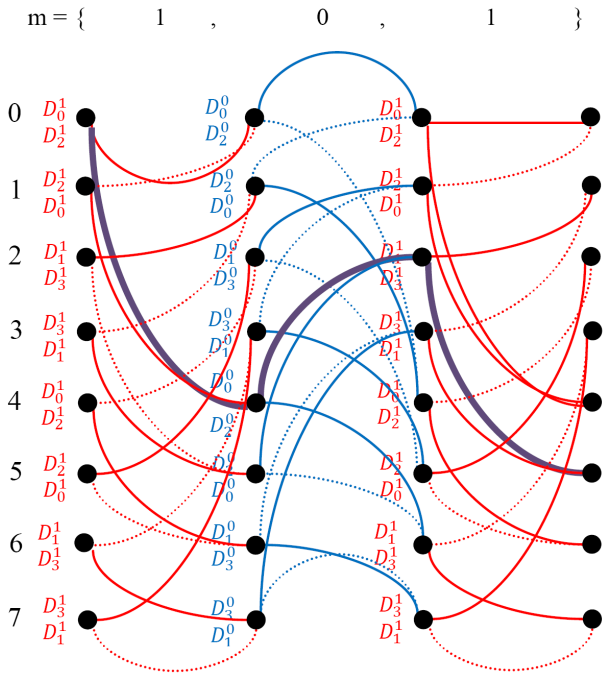


Figure 4 – *Insertion du message m = 1, 0, 1.*

vecteur **y**. L'extraction du message est réalisée en décodant **y** par un algorithme de Viterbi appliqué au treillis complet, c'est à dire sans élagage.

4.3 Extraction de la marque pendant la décompression

On peut également extraire le message inséré durant la décompression JPEG2000 lors de l'étape de quantification inverse. Pour chaque code-block, les étapes sont les suivantes :

– Récupération des décalages **d** à l'aide de la clé secrète et génération des groupes de quantificateurs d'union 0 et 1.
– Extraction de la marque et quantification inverse : La séquence **l** est utilisée pour reconstruire la structure du treillis employée lors de la quantification. Le message inséré peut alors être extrait en examinant le type de quantificateurs qui étiquettent les branches du treillis. Si le groupe 0 est utilisé au niveau de la transition i, cela signifie que le bit inséré $\hat{m}[i]$ est égal à 0. Sinon, c'est le groupe 1 qui est utilisé et $\hat{m}[i] = 1$. Le processus de quantification inverse est ensuite éxecuté afin de reconstruire les coefficients d'ondelettes.

Remarquons qu'il est également possible de décompresser le bit-stream JPEG2000 marqué à l'aide d'un décodeur JPEG2000 classique. Cependant, l'image obtenue sera différente en terme de qualité avec celle obtenue à l'aide du schèma conjoint. De plus, l'information de tatouage sera perdue car le décodeur classique n'utilise pas les quantificateurs décalés. Il sera donc impossible de procéder à une extraction de la marque pendant ou après la décompression.

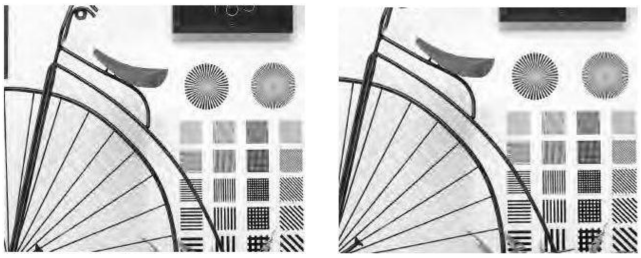
5 Résultats expérimentaux

Afin de procéder à l'implémentation de notre schéma conjoint compression/tatouage, nous avons choisi de travailler avec la librairie OpenJPEG [16]. Cette librairie est un code source libre écrit en langage C implémentant la norme JPEG2000 partie 1. Les paramètres de compression et de tatouage qui ont été utilisés dans nos tests sont les suivants : une transformée en ondelettes 9/7 sur 5 niveaux de décomposition, un découpage de l'image en une seule tuile, cette dernière étant découpée en un certain nombre de code-blocks. La taille des code-blocks est égale à 64 x 64 pour les premier et second niveaux de résolution et 16 x 16 pour les niveaux restants. Un message binaire de longueur égale à 4080 bits est inséré au niveau des sous-bandes de détail de tous les niveaux de résolution sauf le premier niveau. Cela signifie que l'on a un bit inséré pour 64 pixels. Le message est codé à l'aide d'un code convolutif de rendement 1/16. Les vecteurs de décalage sont gardés secret afin de protéger la marque. Plusieurs tests ont étés effectués sur un certain nombre d'images en niveaux de gris. Nous présentons ici les résultats de nos tests sur les images bike et peppers de taille 512 x 512. Nous avons observé les effets induits par une variation du taux de compression comprise entre 2.5 et 0.2 bpp sur la qualité de l'image décompressée/tatouée comme le montre le Tableau 1. Au-delà de 0.2 bpp, il a été constaté que l'image décompressée est trop dégradée pour que la persistance du tatouage soit nécessaire. La qualité de l'image reconstruite est évaluée à l'aide du PSNR. La marque est correctement détectée durant l'étape de décompression, et ce, pour tous les débits binaire testés. La comparaison des mesures de PSNR entre les images compressées et les images compréssées/tatouées montrent que l'insertion de la marque entraîne une légère baisse du PSNR. De même, nous avons remarqué que la qualité visuelle de l'image compressée/tatouée est proche

Image test	Débit (bpp)	PSNR (dB) avec JP2k	PSNR (dB) schéma conjoint
Bike	2.5	43.46	43.09
	2	41.47	41.16
	1.6	39.68	39.56
	1	38.07	38.19
	0.5	36.72	36.57
	0.2	33.31	32.92
Peppers	2.5	44.41	44.23
	2	43.25	42.24
	1.6	40.17	40.10
	1	39.05	38.96
	0.5	36.41	36.41
	0.2	29.04	29.42

Tableau 1 – *PSNR des images compressées en utilisant JPEG2000 avec et sans tatouage*

de celle obtenue avec une compression JPEG2000 classique comme l'illustre la figure 5. L'ajout de la marque ne dégrade pas de manière significative le taux de compression. Les résultats obtenus montrent que le tatouage résiste à la phase de contrôle de taux de JPEG2000 et que l'insertion de la marque conduit à une dégradation minimale des performances du codeur.



(a) (b)

Figure 5 – *Image compressée Bike avec et sans tatouage : a) image compressée à 0.5 bpp , b) image compressée et tatouée à 0.5 bpp.*

6 Conclusion

Nous avons proposé un algorithme de compression/tatouage conjoint avec JPEG2000 basé sur la quantification TCQ. Les résultats expérimentaux ont prouvé que ce système conjoint fournit de bonnes performances de compression. Il permet d'obtenir une bonne qualité d'image en termes de PSNR. De plus, l'insertion de la marque n'a qu'un faible impact sur le taux de compression. Un autre avantage est la faible complexité de l'ensemble du système du fait d'opérations communes au tatouage et à la compression (transformée, quantification). La capacité atteinte est élevée et permet à l'utilisateur d'insérer suffisamment d'informations sur l'image et sur son auteur. Par conséquent, le schéma proposé constitue une bonne solution pour des applications d'enrichissement de contenu ou d'auto-indexation. Les travaux futurs porteront sur l'étude de la robustesse de ce système conjoint face à certaines attaques d'effacement ainsi qu'a la recherche du meilleur compromis possible entre la robustesse du tatouage et la qualité visuelle de l'image obtenue.

Références

[1] B. Chen et G. Wornell. Quantization index modulation : A class of provably good methods for digital water-marking and information embedding. *IEEE Transaction on Information Theory*, 47 :1423–1443, Janvier 2001.

[2] J. J. Eggers, R. Bäuml, R. Tzschoppe, et B. Girod. Scalar costa scheme for information embedding. *IEEE Transaction on Signal Processing*, 51(4) :1003–1019, Janvier 2003.

[3] M. L. Miller, G. J. Doerr, et I. J. Cox. Applying informed coding and informed embedding to design a robust, high capacity watermark. Dans *IEEE Transactions on Image Processing*, 2004.

[4] M. Costa. Writing on dirty paper. *IEEE Transaction on Information Theory*, 29(3) :439–441, Janvier 1983.

[5] D.S. Taubman et M.W. Marcellin. *JPEG2000 image compression fundamentals, standards and practice*. Kluwer Academic Publishers, Boston, 2002.

[6] P. Meerwald. Quantization watermarking in the jpeg2000 coding pipeline. Dans *Communications and Multimedia Security*, pages 69–79, Darmstadt, Germany, Mai 2001.

[7] M. Schlauweg, D. Pröfrock, et E. Müller. Jpeg2000-based secure image authentication. Dans *workshop on Multimedia and security*, pages 62–67, Geneva, Switzerland, Septembre 2006.

[8] A. Makhloufi, A. Ouled Zaid, R. Boualleg, et A. Boualleg. Improved quantization index modulation based watermarking integrated to jpeg2000 coding scheme. Dans *SPIE*, pages 17–21, San Diego, USA, Janvier 2010.

[9] M.W. Marcellin et T.R. Fischer. Trellis coded quantization of memoryless and gauss-markov sources. *IEEE Transaction on communication*, 38 :82–93, Janvier 1990.

[10] JPEG2000 Part II Final Committee Draft Version 1.0. *ISO/IEC JTCI/SC29 WG1*. 2000.

[11] F. Dufaux, S. Wee, J. Apostolopoulos, et T. Ebrahimi. Jpsec for secure imaging in jpeg2000. Dans *SPIE, Application of Digital Image Processing*, pages 319–330, Denver, Août 2004.

[12] G. Ungerboeck. Channel coding with multilevel/phase signals. *IEEE Transaction on Information Theory*, 28(1) :55–67, Janvier 1982.

[13] G. D. Forney Jr. The viterbi algorithm. *IEEE Transaction on Information Theory*, 61 :268–278, Janvier 1973.

[14] S. Braci, R. Boyer, et C. Delpha. Security evaluation of informed watermarking schemes. Dans *ICIP*, Cairo, Egypt, Novembre 2009.

[15] A. Ouled Zaid, A. Makhloufi, et A. Bouallegue. Wavelet domain watermark embedding strategy using ttcq quantization. *IJCSNS*, 7(6) :268–278, Janvier 2007.

[16] The OpenJPEG library is available for download at. *http ://www.openjpeg.org*.

COMPRESSION VIDEO AVEC MARQUAGE D'INDEX DE MODE INTRA DANS LA CHROMA

J.-M. Thiesse[1] J. Jung[1] M. Antonini[2]

[1] Orange Labs
38 rue du G. Leclerc, 92794 Issy les Moulineaux, France
{jeanmarc.thiesse, joelb.jung}@orange-ftgroup.com

[2] I3S Lab. Université de Nice-Sophia Antipolis/CNRS
2000 route des Lucioles, 06903 Sophia Antipolis, France
am@i3s.unice.fr

Résumé

Des nouvelles activités ont été récemment lancées dans le but de succéder au standard de compression vidéo H.264/AVC. Plusieurs améliorations de ce standard ont déjà été proposées, mais l'objectif de 50% de gains sur le débit pour une qualité équivalente n'est pas encore atteint. Dans ce contexte, un précédent travail propose d'utiliser des techniques de marquage afin de réduire le coût de codage d'informations de signalisation résultantes d'une amélioration du codage Inter. L'idée principale est de cacher des index de signalisation de codage dans les coefficients transformés et quantifiés de chrominance et de luminance judicieusement sélectionnés. Afin de minimiser l'erreur de prédiction, la modification est réalisée à l'aide d'une optimisation débit-distorsion.
Dans cet article, la méthode est étendue au codage Intra et au format de couleur 4 :4 :4 dans le but d'explorer les limites mises en évidence par l'étude précédente et de répondre aux questions subsistantes. Une nouvelle méthode d'optimisation construite à partir de la théorie de Pareto est aussi proposée. Les améliorations résultantes (1.5% de gain moyen sur le débit) sont rapportées et analysées objectivement et subjectivement pour plusieurs séquences.

Mots clefs

Compression vidéo, H.264/AVC, marquage, indice de compétition, chroma.

1 Introduction

Des gains significatifs par rapports aux précédents standards de codage vidéo ont été obtenus par H.264/AVC [1] issu des travaux conjoints du Video Coding Experts Group (VCEG/ITU-T SG16-Q6) et du Moving Pictures Experts Group (MPEG / ISO/IEC). Ces gains proviennent de l'amélioration des outils existants et de l'inclusion de nouveaux outils. Ces amélioration concernent notamment l'estimation de mouvement, le codage des informations avec le Context Adaptive Binary Arithmetic Coding (CABAC) et l'ajout de nombreux modes de codage Intra et Inter mis en compétition, qui nécessitent la transmission d'indices pour leur signalisation.
Aujourd'hui, de nouvelles activités de standardisation en codage vidéo sont lancées. VCEG et MPEG travaillent au sein du regroupement JCT-VC (Joint Collaborative Team on Video Coding) et viennent de clore un appel à soumission dont les premiers résultats ont été délivrés en avril 2010. L'objectif est de proposer un standard qui atteigne 50% de réduction du débit pour une même qualité subjective avec une augmentation de la complexité par un facteur 2 ou 3. Plusieurs améliorations sont déjà connues et rassemblées au sein du codeur JM KTA [2] (Key Technical Area). Certaines de ces améliorations et d'autres contributions reconnues ajoutent de nouveaux indices de signalisation de compétition qui sont de plus en plus coûteux. Nous avons proposé dans un précédent travail [3] de réduire le coût de signalisation de l'un de ces outils [4] à l'aide d'une méthode permettant de cacher les indices dans les résiduels de pixels de chroma et de luma. A présent, nous proposons d'explorer les limites mises en évidence par la précédente étude qui a été réalisée sur le codage Inter. Le schéma est tout d'abord étendu pour le codage Intra combiné à une sélection optimale de l'index à cacher et une nouvelle optimisation débit-distortion bâtit sur la théorie de Pareto. Le schéma est aussi testé pour des séquences au format de couleur 4 :4 :4 qui, comme les blocs Intra, permet de cacher davantage d'indices dans la composante de chroma.
Dans la suite de cet article nous présentons dans la partie 2 un état de l'art sur H.264/AVC et le marquage en rappelant les principales conclusions de la précédente étude. Nous détaillons ensuite le schéma proposé dans la partie 3 après une étude préliminaire. Les résultats obtenus sont finalement présentés et analysés de façon objective dans la partie 4 et subjective dans la partie 5.

2 Etat de l'art

2.1 H.264/AVC

H.264/AVC est un codeur vidéo hybride : les prédictions Intra et Inter sont utilisées conjointement afin d'exploiter

les redondances spatiales et temporelles. Pour chaque macrobloc, de nombreux modes sont mis en compétition et nécessitent d'être signalés au décodeur comme d'autres informations de résiduels ou de codage (sous-mode, taille de la DCT, ...).
Toutes ces informations sont séquentiellement rangées dans le bitstream, indépendamment les unes des autres. Dans le but de réduire le coût de codage un Most Probable Mode (MPM) est utilisé : le mode qui est le plus probable est codé avec un seul bit. Bien qu'un codage adaptatif suivant le contexte soit réalisé, le coût de codage demeure élevé. A moyen débit, les informations de compétition du codeur de référence H.264/AVC JM représentent 20% du débit total et atteignent 40% pour les plus bas débits.
Dans cet article, nous proposons d'utiliser le marquage afin de diminuer le coût des indices de compétition, en l'appliquant au cas particulier du MPM Intra.

2.2 Marquage de vidéo

Le marquage et la compression vidéo ont traditionnellement deux buts contradictoires. Le premier ajoute des informations non perceptibles afin de cacher des données tandis que la seconde supprime des informations redondantes afin de réduire la taille de la vidéo. Cette partie introduit le marquage de vidéo.

Théorie. Le marquage ou tatouage traite de la capacité à masquer des données dans un média avec un minimum de dégradation visible. Il y a de nombreuses applications du marquage : signature de vidéos, authentification basée sur le contenu, empreinte digitale ou correction d'erreurs. Pour chaque application, un complexe compromis entre trois paramètres est nécessaire : quantité de données à cacher (aussi appelé le message), fidélité (la distorsion induite par la marque) et robustesse (la résistance aux attaques). Des critères de complexité, de sécurité et de réversibilité sont aussi pris en compte selon le type d'application.
Trois classes d'approches du marquage de vidéos sont répertoriées : la première étend le marquage d'image fixe tandis que la seconde exploite la dimension temporelle de la vidéo dans le but d'augmenter la fidélité et la robustesse. Ces contributions [5] utilisent principalement la théorie de l'étalement de spectre : le message est réparti parmi un large domaine de fréquences du signal hôte.
La dernière classe est basée sur les caractéristiques du standard de compression vidéo. Ainsi, le sujet le plus étudié concerne le masquage d'informations dans les coefficients transformés non-nuls du signal vidéo compressé. Parmi ces approches, le Force Even Watermarking (FEW) a été proposé dans [6] : les informations sont cachées dans la parité des coefficients AC, ceci entraîne une augmentation du débit. Après une étude poussée du fonctionnement de CABAC, les auteurs de [7] ont proposé de cacher le bit de la marque dans le signe des Trailing Ones. Cela n'implique pas de modification du débit mais il y a une dégradation de la qualité visuelle accentuée par la propagation de l'erreur engendrée par la marque. Une solution est proposée dans [8] en cachant chaque bit dans la parité de la somme des coefficients transformés. Cette approche permet de choisir la modification qui induit la plus faible dégradation du signal.

Compression via marquage. Seulement quelques méthodes proposent d'utiliser le marquage dans un but de compression vidéo comme dans cet article. La notion de marquage pour la compression a été formalisée dans [9]. Les auteurs y proposent d'exploiter le marquage pour améliorer l'efficacité de codage en masquant les informations de chrominance dans le domaine transformé en ondelettes de la composante de luminance avant compression de l'image. Un schéma proche est proposé dans [10] en utilisant la DCT à la place de la transformée en ondelette dans le but de marquer les informations de couleurs dans une image JPEG en niveau de gris. Dans [11], le signal à compresser est séparé entre une partie d'image hôte et une partie résiduel. Seulement la portion d'image hôte nécessite d'être compresser après une étape de masquage de la partie résiduelle.
Ces approches confirment que l'utilisation de méthodes de marquage dans un but de compression peut être une piste de recherche prometteuse.

2.3 Marquage pour la compression en codage Inter

Dans notre précédent travail [3], nous avons proposé de cacher des informations de compétition sur le mouvement générées par l'outil Motion Vector Competition (MVComp) [4] dans les coefficients transformés et quantifiés des résiduels de luma et de chroma. Plus précisément, un marquage dans la parité de la somme des coefficients transformés est utilisé car il permet de facilement équilibrer entre la modification du débit et de la distorsion. Bien que l'article résolvait les difficultés initiales dû à cette approche originale de compression vidéo, les gains moyens sur le débit demeurent faibles. Différentes limites de la méthode ont été rapportées et nécessitent d'être approfondies, notamment :

1. Le nombre de coefficients disponible pour le marquage.
2. La sélection de l'information à cacher.
3. La sélection des coefficients à modifier.

Dans cette nouvelle étude, nous explorons ces limites dans le but d'étendre le premier travail et de proposer des gains plus élevés.

3 Approche proposée et extensions

3.1 Etude préliminaire

Comme expliqué dans la partie précédente, nous avons favorisé un schéma qui place la marque directement dans le signal vidéo compressé, plus particulièrement dans les résiduels de pixels. Dans le but de résoudre les problèmes mis en évidence dans [3], nous proposons dans une première

extension de considérer le codage Intra dans la mesure où les blocs Intra contiennent davantage de coefficients transformés non-nuls, particulièrement dans la composante de chroma. Parmi tous les indices de compétition Intra, l'index Most Probable Mode (MPM) a été sélectionné car c'est le plus coûteux (6% du débit Intra total à moyen débit). Il s'agit d'un index binaire qui indique si le prédicteur Intra à coder est égal ou non au mode le plus probable déterminé à partir des blocs voisins.
La seconde extension proposée est une méthode de choix de la modification basée sur la théorie de Pareto afin d'avoir une alternative à la détermination de multiplicateurs de Lagrange adaptés nécessaires pour l'optimisation RD utilisée dans la précédente etude.
Le schéma est enfin appliqué au format de couleur 4 :4 :4 qui est un format important pour l'amélioration future de la qualité des services vidéos. Il permet de masquer plus d'indices dans la chroma. Chaque index étant caché dans plus de coefficients, il y a aussi un impact réduit sur les pixels reconstruits car la marque est mieux répartie.

3.2 Description du schéma

Nous proposons d'utiliser les coefficients transformés et quantifiés de chroma afin de masquer l'index MPM noté i du $k^{ème}$ bloc. Ce masquage est complexe du fait des spécificités du codeur entropique CABAC (utilisation de contextes) et des dégradations possible de la prédiction qui peuvent résulter d'un mauvais choix de marquage. Afin de résoudre ces problèmes, le schéma proposé doit satisfaire les conditions suivantes :

1. Contrôle de la modification du débit : elle doit être inférieure au coût de l'index, $R_k^w - R_k < \varsigma(i)$, avec R_k et R_k^w les débits original et modifié du $k^{ème}$ bloc, et $i \in \{0,1\}$ l'index MPM avec $\varsigma(i)$ son coût.
2. Minimisation de la dégradation de la prédiction : la modification des coefficients transformés doit être invisible et l'impact sur la prédiction des blocs suivants et des images suivantes doit être minimal.

Dans le but de respecter ces conditions, nous utilisons la parité de la somme des coefficients pour marquer l'index MPM. La première condition est respectée en rejetant toutes les modifications qui induisent une augmentation du débit original supèrieur au coût de l'index. Pour la seconde condition, une optimisation débit-distorsion est appliquée.
Notons a_n, $n \in \{1, ..., N\}$ les coefficients transformés et quantifiés avant transmission. La somme des coefficients pour le $k^{ème}$ bloc est notée S_k :

$$S_k = \sum_{n=1}^{N} a_n. \tag{1}$$

La somme après application de la marque, noté S_k^w est obtenue ainsi :

$$S_k^w = \begin{cases} S_k & ; |S_k| \, mod \, 2 = i \\ S_k + m_k & ; |S_k| \, mod \, 2 \neq i, \end{cases} \tag{2}$$

où $i \in \{0,1\}$ est l'index MPM et m_k la modification des coefficients. Les blocs qui ne contiennent que des coefficients DC ne sont pas utilisés afin de limiter la dégradation.

Nous proposons d'utiliser une optimisation de Pareto pour sélectionner la meilleure modification lorsque les coefficients nécessitent d'être modifiés, c'est à dire lorsque :

$$i \neq |S_k| \, mod \, 2. \tag{3}$$

Nous considérons N coefficients transformés a_n, $a_n \neq 0$. Pour chaque coefficient, six couples débit distorsion $(R_n^{w_j}, D_n^{w_j})$ sont calculés après addition d'une valeur impaire m_j :

$$a_n^{w_j} = a_n + m_j. \tag{4}$$

Nous avons choisis de limiter les valeurs de m_j à $\{-5, -3, -1, 1, 3, 5\}$ car des valeurs plus élevées induisent une modification trop importante du signal. $R_n^{w_j}$ est le débit généré par le codage des coefficients modifiés et $D_n^{w_j}$ est la distorsion associée. Les coefficients transformés et quantifiés égaux à zéro ne sont pas considérés afin de ne pas casser les séries de zéro dont le codage est optimisé. Chaque index peut être caché dans l'une des composante U ou V de la chroma.
En alternative à l'optimisation débit-distorsion classique utilisée dans l'étude précédente [3], nous appliquons dans cet article une méthode basée sur la théorie de l'optimisation multicritère afin de choisir la meilleure modification. Cette méthode est construite sur la théorie de Pareto et elle est divisée en trois étapes :

1. Calcul des couples débit-distorsion $(\overline{R}_n^{w_j}, \overline{D}_n^{w_j})$ correspondant à un couple Pareto optimal $(\overline{a}_n, \overline{m}_j)$. Un couple donné de variables, $(\overline{a}_n, \overline{m}_j)$, est optimal au sens de Pareto si et seulement si il n'est pas dominé par un autre couple (a_n, m_j), où (a_n, m_j) domine $(\overline{a}_n, \overline{m}_j)$ signifie :

$$soit \left\{ (R_n^{w_j} \leq \overline{R}_n^{w_j}) \; et \; (D_n^{w_j} < \overline{D}_n^{w_j}) \right\},$$

$$soit \left\{ (R_n^{w_j} < \overline{R}_n^{w_j}) \; et \; (D_n^{w_j} \leq \overline{D}_n^{w_j}) \right\}.$$

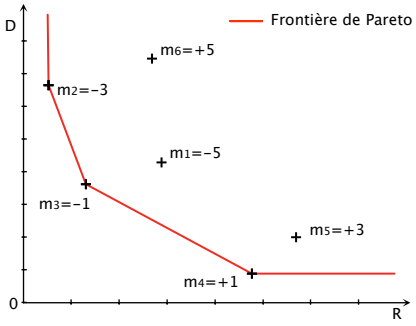


Figure 1 – *Frontière de Pareto obtenue pour un ensemble donné de couples RD.*

La figure 1 représente un exemple de frontière de Pareto. Les points représentent les six modifications possibles pour un coefficient a_n donné, la courbe rouge est la frontière de Pareto sur laquelle la modification va être sélectionnée.

2. Normalisation de ces couples :

$$\hat{R}_n^{w_j} = \frac{(\overline{R}_n^{w_j} - \overline{R}_{min})}{(\overline{R}_{max} - \overline{R}_{min})},\ \hat{D}_n^{w_j} = \frac{(\overline{D}_n^{w_j} - \overline{D}_{min})}{(\overline{D}_{max} - \overline{D}_{min})}$$

où $(\hat{R}_n^{w_j}, \hat{D}_n^{w_j})$ est le couple normalisé, $(\overline{R}_{min}, \overline{R}_{max})$ et $(\overline{D}_{min}, \overline{D}_{max})$ sont les valeurs minimales et maximales du débit et de la distorsion parmi les couples Pareto optimaux.

3. Sélection du meilleur couple parmi les couples $(\hat{R}_n^{w_j}, \hat{D}_n^{w_j})$ par minimisation de la norme euclidienne :

$$(R_{opt}, D_{opt}) = min\left\{\left\|(\hat{R}_n^{w_j}, \hat{D}_n^{w_j})\right\|_2\right\}_{n\in[1,N],\ j\in[1,6]}$$

où la norme euclidienne du couple (R, D) est définie par :

$$\|(R, D)\|_2 = \sqrt{R^2 + D^2}$$

Finalement, la modification sélectionnée (a_n, m_j) est celle associée au couple optimal (R_{opt}, D_{opt}). Notons que les couples ayant une distorsion trop élevée sont rejetés par la méthode afin de ne pas choisir une modification donnant un bloc reconstruit trop différent du bloc original, cela permet de préserver l'efficacité de la prédiction dans la mesure ou cette modification a un impact sur la prédiction des blocs suivants. Le principal avantage de ce schéma comparé à l'optimisation RD classique est qu'il est indépendant du calcul de multiplicateurs de Lagrange et ainsi adaptatif pour chaque bloc. Cette approche permet de comparer tout les meilleurs couples avec un ordre de grandeur correct lors de l'étape 3. Cependant, ce schéma ne tient pas compte des paramètres de quantification (QP), contrairement aux multiplicateurs de Lagrange qui sont fonctions du QP.

Enfin, l'extraction de l'index au niveau du décodeur est exprimée par $i = |S_k|\ mod\ 2$, avec i qui représente l'index extrait du $k^{ème}$ bloc et S_k la somme des coefficients a_n décodés. L'index MPM est fixé à i seulement si il y a au moins un coefficient AC non nul. Autrement, l'index est décodé de façon classique.

4 Résultats expérimentaux

4.1 Protocole de test

Le schéma proposé a été implémenté dans le codeur JM KTA [2] version 2.1. Les résultats de référence ont été générés sur un profil Intra avec les modes Intra 16×16 et Intra 8×8 activés. La DCT 8×8 est activée et le codage entropique CABAC est sélectionné. La méthode est implémentée pour la transmission des index MPM de l'Intra 8×8.
Le premier test a été réalisé sur huit séquences YUV 4 :2 :0 ayant des résolutions de QWVGA à 1080p. Deux ensembles de QPs ont été sélectionnés pour étudier les résultats à haut et moyen débit : QP set 1 : 12-17-22-27 et QP set 2 : 22-27-32-37. Les résultats représentent le gain sur le débit calculé avec la métrique de Bjontegaard [12] comme cela est recommandé par VCEG. Un test sur plusieurs séquences au format de couleur 4 :4 :4 est aussi présenté dans une seconde partie.

4.2 Etude objective

Ensembles de QP	Set 1	Set 2
Mobisode2 qwvga	2.3	1.3
RaceHorses qwvga	0.9	1.1
RaceHorses wvga	1.0	1.0
DucksTakeOff 720p	1.4	1.7
BlueSky 1080p	1.3	2.1
DucksTakeOff 1080p	1.6	1.8
SunFlower 1080p	2.8	1.1
Tractor 1080p	1.9	1.5
Moyenne	**1.6**	**1.5**

Tableau 1 – Pourcentage de gains sur le débit pour les deux ensembles de QPs.

Résultats en YUV 4 :2 :0. Le tableau 1 donne le pourcentage de gains sur le débit pour chaque séquence et pour les deux ensembles de QPs. La méthode proposée donne systématiquement un gain sur le débit pour toutes les séquences étudiées et atteint 2.8% et 2.1% pour les séquences *SunFlower 1080p* et *BlueSky 1080p*. Contrairement au précédent schéma proposé en codage Inter, l'efficacité de la méthode est préservé à moyen débit où le nombre de coefficients transformés est pourtant moins élevé qu'à haut débit avec respectivement 1.5% et 1.6% de gains sur le débit en moyenne. Ce gain est significatif compte tenu du fait que la méthode traite seulement un index de signalisation dont la proportion dans le bitstream reste limitée bien qu'étant l'index le plus coûteux. En considérant de plus que la méthode demeure bornée en Intra par le nombre de blocs ayant des coefficients non nuls, ces gains confirment l'intérêt d'une telle approche pour réduire le coût de transmission d'indices de signalisation.

QP	12	22	32
Chroma DC	19	27	31
Chroma AC	38	33	29
Non-modifié	43	40	40

Tableau 2 – Position de la marque (en pourcentage) pour trois QPs.

Lorsqu'il y a des coefficients non-nuls disponibles, c'est à dire lorsque l'index MPM n'est pas transmis, le tableau 2 donne la position de la marque. Les coefficients sont modifiés dans 60% des cas et les indices sont majoritairement masqués dans la composante AC à haut débit tandis que la composante DC joue un rôle plus important à moyen débit.

Résultats en YUV 4 :4 :4. Nous avons testé la méthode sur un ensemble de douze séquences disponibles à la fois au format de couleur YUV 4 :2 :0 et YUV 4 :4 :4. La figure 2 donne le pourcentage moyen d'indices caché pour le schéma appliqué pour les deux formats pour des QPs de 12 à 37. Le marquage en YUV 4 :4 :4 permet systématiquement de réduire le nombre d'index transmis comparé au marquage pour les séquences en YUV 4 :2 :0. Cette réduction varie entre 10% et 20% et est plus élevée pour les QPs 17 et 22 où le nombre de coefficients transformés commence à décroître pour les séquences YUV 4 :2 :0. Pour la méthode étendue au format de couleur YUV 4 :4 :4, le pourcentage d'index masqué atteint 93% à haut débit. Comme attendu, moins d'indices sont transmis pour ce format. Cela est confirmé par une augmentation du gain sur le débit qui atteint 2.60% pour la séquence *DucksTakeOff 720p* pour le second ensemble de QPs en YUV 4 :4 :4 (contre 1.70% en YUV 4 :2 :0). En moyenne, le gain sur le débit est augmenté de 36% et 47% en YUV 4 :4 :4 comparé au format YUV 4 :2 :0, pour le premier et le second ensemble de QP respectivement.

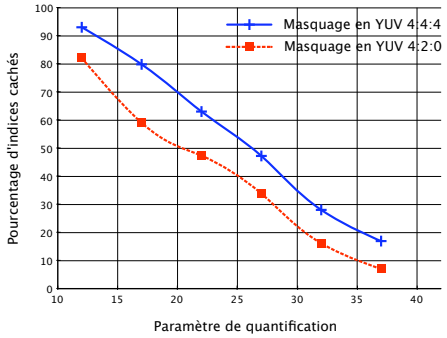


Figure 2 – *Pourcentage d'indices cachés par le schéma proposé en YUV 4 :2 :0 et YUV 4 :4 :4 pour chaque QP.*

5 Etude subjective

5.1 Observation

	Chroma U		Chroma V	
Ensembles de QP	Set 1	Set 2	Set 1	Set 2
Mobisode2 qwvga	-0.49	-0.25	-0.26	-0.11
RaceHorses qwvga	-0.51	-0.58	-0.45	-0.52
RaceHorses wvga	-0.52	-0.39	-0.40	-0.45
DucksTakeOff 720p	-0.65	-0.76	-0.20	-0.17
BlueSky 1080p	-0.54	-0.60	-0.35	-0.29
DucksTakeOff 1080p	-0.64	-0.60	-0.24	-0.14
SunFlower 1080p	-0.82	-0.43	-0.58	-0.33
Tractor 1080p	-0.68	-0.45	-0.54	-0.43
Moyenne	**-0.61**	**-0.51**	**-0.38**	**-0.31**

Tableau 3 – Différences de PSNR pour les composantes U et V pour la méthode proposée par rapport à la référence.

Le tableau 3 donne la différence moyenne de PSNR pour les composantes de chroma U et V pour la méthode proposée par rapport à la référence pour deux ensembles de QPs. Comme attendu, le schéma proposé induit une perte objective systématique pour toutes les séquences. On observe une différence de 0.2 entre les pertes des composantes U et V, indiquant que les coefficients de Chroma U ont été plus déteriorés par la marque. Les pertes augmentent de 0.1 en moyenne à haut débit (set 1) comparé au moyen débit (set 2). Ceci s'explique par le nombre plus élevé de coefficients modifiés à haut débit. Finalement, on peut noter que la séquence ayant la perte la plus élevée (*SunFlower 1080p*) correspond à la séquence ayant le meilleur gain dans le tableau 1.
Cette dégradation objective de la chroma est une conséquence naturelle de la méthode proposée. La principale question est de savoir si elle est perceptible par l'oeil humain. Dans ce but, une évaluation subjective a été réalisée.

5.2 Protocole expérimental

Une évaluation subjective est une méthode expérimentale visant à déterminer des modifications dans la qualité perçue. Chaque test consiste en un couple de stimulis, comprenant la référence et la séquence encodée avec la méthode proposée, affiché dans un ordre aléatoire. Cinq des séquences avec la plus grande perte de chroma calculée objectivement (tableau 3) ont été affichées deux fois : le premier test est pour une évaluation globale et le second pour une évaluation spécifique de la qualité de la couleur. Dans ce second test, on demande explicitement aux évaluateurs de se concentrer sur la distorsion de la chroma. La figure 3 détaille l'échelle utilisée pour l'évaluation de la qualité.

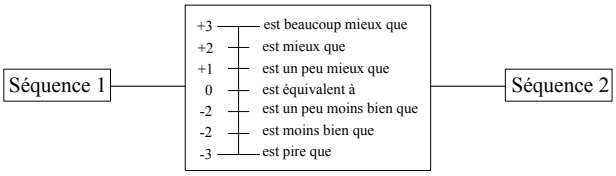


Figure 3 – *Echelle d'évaluation de la qualité.*

5.3 Interprétation

La figure 4 donne les résultats obtenus pour chaque séquences pour les deux phases de tests, les résultats sont donnés à haut (a) et moyen (b) débit. La conclusion majeure est que la qualité des séquence (globale et couleur) n'est pas visuellement affectée par la méthode proposée. En effet, en dehors de la séquence *RaceHorses* dont les couleurs sont très vives, aucun score supérieur à un ("un peu meilleur") n'a été utilisé par les évaluateurs. La dégradation visuelle peut ainsi être considérée comme invisible. On peut cependant noter que l'impact de la méthode est davantage perçu à haut débit qu'à moyen débit, ce qui confirme les résultats observés dans la partie 5.1. Enfin, la modification de la qualité est davantage perçue dans le test axé sur la couleur mais reste faible.
La figure 5 illustre cette analyse avec la séquence *RaceHorses WVGA* qui correspond à la perte la plus élevée sub-

jectivement. La figure présente une capture de la séquence compressée avec le codeur de référence (a), avec la méthode proposée (b) et la différence des deux séquences (c).

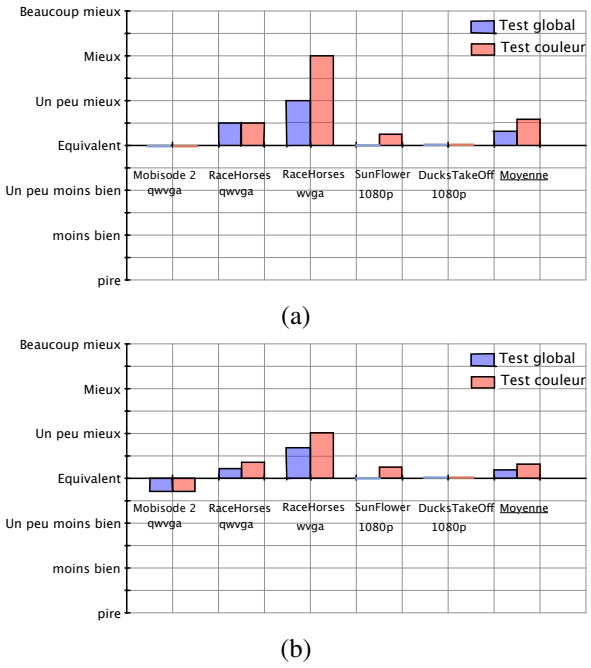


Figure 4 – *Evaluation de séquences testées à haut (a) et moyen (b) débit. Comparaison de la référence avec la méthode proposée.*

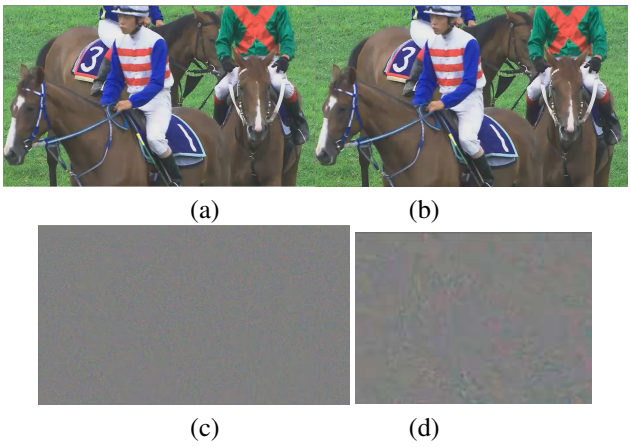


Figure 5 – *Séquence RaceHorces WVGA : (a) Compressée avec le codeur de référence, (b) Avec le masquage dans la chroma, (c) Différence entre (a) et (b), (d) Zoom de (c).*

6 Conclusion

Une application non conventionnelle du marquage a été proposée dans un précédent article dans le but de faire de la compression vidéo. Les indices MVComp étaient cachés dans les composantes de chroma et de luma.
Cet article prolonge ce sujet et résout les principales questions précédemment soulevées. Le schéma est ainsi appliqué à l'index Intra Most Probable Mode. L'application de la méthode au codage Intra permet d'avoir un nombre plus élevé de coefficients transformés non-nuls disponibles pour le marquage. Une seconde extension est proposée avec une optimisation basée sur la théorie de Pareto qui permet de sélectionner la position de la marque sans nécessité de fixer des multiplicateurs de Lagrange. Un gain moyen sur le débit de 1.5% (jusqu'à 2.8%) est rapporté, ce qui est un gain significatif en considérant que seulement un index de signalisation a été considéré. La méthode a aussi été testée pour des séquences au format de couleur 4 :4 :4. Comme attendu, une augmentation de 47% des gains sur le débit précédents est observée.
L'application de techniques de marquage dans un but de compression vidéo soulève plusieurs difficultés techniques. Cet article en résout la plupart, et montre qu'une telle technique est intéressante pour améliorer le standard actuel. Notamment, cette méthode est particulièrement adaptée pour réduire le coût d'indices de signalisation qui tendent à prendre de plus en plus de place dans les codeurs vidéos actuels basés compétition.

Références

[1] "Advanced Video Coding for Generic Audiovisual Services, ITU-T Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4 AVC)," Standard Version 7 : Apr. 2007, ITU-T and ISO/IEC JTC 1.

[2] K. Suehring, "H.264/AVC and KTA software coordination." Available : http ://iphome.hhi.de/suehring/.

[3] J.-M. Thiesse, J. Jung and M. Antonini, "Data hiding of Motion Information in Chroma and Luma Samples for Video Compression," MMSP'10, France, Oct. 2010.

[4] G. Laroche, J. Jung, and B. Pesquet-Popescu, "RD optimized coding for motion vector predictor selection," IEEE Trans. on CSVT, 18(9) :1247-1257, Sep. 2008.

[5] F. Hartung and B. Girod, "Watermarking of Uncompressed and Compressed Video", in Sig. Proc., 66(3) :283-301,1998.

[6] O. Nemethova, G. Calvar Forte, and M.Rupp, "Robust Error Detection for H.264/AVC Using Relation Based Fragile Watermarking," IWSSIP, Budapest, Hungary, Sep. 2006.

[7] S. M. Kim, S. B. Kim, Y. Hong, and C. S. Won, "Data Hiding on H.264/AVC Compressed Video," ICIAR, Berlin, 2007.

[8] L. M. Marvel, G. W. Hartwig, C. Boncelet, "Compression Compatible Fragile and Semi-Fragile Tamper Detection," SPIE Int. Conf. on Security and Watermarking of Multimedia Contents II, vol. 3971, USA, 2000.

[9] P. Campisi, D. Kundur, D. Hatzinakos, and A. Neri, "Compressive Data Hiding : An Unconventional Approach for Improved Color Image Coding," EU. JASP, 152-163, 2002.

[10] M. Chaumont, and W. Puech, "A DCT-Based Data-Hiding Method to Embed the Color Information in a JPEG Grey Level Image," EUSIPCO, Italy, Sep. 2006.

[11] B. Zhu, and A. H. Tewfik, "Media Compression via Data Hiding," 31^{st} Asilomar Conf. on SSC, pp. 647-650, 1997.

[12] J. Jung and S. Pateux, "An Excel add-in for computing Bjontegaard metric and its evolution," ITU-T VCEG contribution VCEG-AE07, Marrakech, Jan. 2007.

Une approche neuronale pour la classification d'actions de sport par la prise en compte du contenu visuel et du mouvement dominant

Moez Baccouche[1,2] Franck Mamalet[1] Christian Wolf[2] Christophe Garcia[1] Atilla Baskurt[2]

[1]Orange Labs - France Télécom R&D
35510 Cesson-Sévigné, France.
{prénom.nom}@orange-ftgroup.com

[2]Laboratoire d'InfoRmatique en Image et Systèmes d'information
Université de Lyon, CNRS, INSA-Lyon, F-69621, France.
{prénom.nom}@insa-lyon.fr

Résumé

Dans cet article, nous proposons une approche de classification automatique de séquences vidéo d'actions de sport. Pour cela, nous extrayons de chaque action des caractéristiques du contenu visuel, en utilisant deux approches, l'une par sac de mots, et l'autre par le mouvement dominant de la scène à chaque instant. La classification de l'évolution temporelle de ces caractéristiques extraites est gérée dynamiquement par un modèle neuronal, basé sur les réseaux de neurones récurrents à large « mémoire court-terme » (LSTM). Les expérimentations faites sur la base « MICC-Soccer-Actions-4 » montrent que l'approche neuronale de classification permet d'obtenir des résultats supérieurs à l'état de l'art ($76\ \%$ *de bonne classification), et que la combinaison des caractéristiques (information visuelle et mouvement dominant) permet un taux de bonne classification de* $92\ \%$.

Mots clefs

Classification d'actions de sport, réseaux de neurones récurrents, *Long Short-Term Memory*, sacs de mots visuels, mouvement dominant.

1 Introduction

Les volumes des contenus audio-visuels mis à disposition ne cessent de croître. Naviguer simplement et rechercher précisément ces contenus au sein de grandes collections devient un problème de première importance. L'un des enjeux majeurs des systèmes d'information s'impose donc comme étant l'indexation et la recherche des vidéos par analyse automatique de leur contenu.

Dans ce contexte, il est important de pouvoir extraire de manière automatique des informations de haut niveau pouvant décrire le contenu sémantique d'une vidéo. Ainsi, de plus en plus de travaux introduisent la notion d'« évènement » ou d'« action » dans des applications diverses (vidéo-surveillance, structuration des contenus télévisuels...). En particulier, les vidéos de sport représentent un type de contenu spécialement intéressant à traiter de part les enjeux commerciaux qui y sont liés.

Plusieurs travaux se sont ainsi intéressés à la classification automatique de séquences vidéo de sport, avec pour objectif la reconnaissance d'évènements de niveau sémantique plus ou moins élevé. On peut en effet distinguer deux catégories de travaux. La première s'intéresse à des évènements de niveau sémantique assez faible, loin de la notion d'action. On peut par exemple citer les travaux de Ekin et al. [1] dans lesquels deux types d'évènements (« phase de jeu » et « pause ») sont identifiés à partir de leurs durées et de leurs angles de prise de vue. Assfalg et al. [2] se basent sur l'extraction de primitives bas-niveau pour classer les plans de dix sports différents en trois catégories. Bien que les résultats relatifs à ces approches soient assez satisfaisants, les évènements identifiés restent néanmoins sémantiquement faibles. Une deuxième catégorie de travaux [3, 4, 5] s'intéresse à des actions complexes mais restent spécifiques vu qu'elles font intervenir des informations relatives au sport étudié (modèle du terrain, règles du jeu...).

Récemment, Ballan et al. [6] ont proposé une approche qui permet de classer des actions de haut niveau sémantique sans faire intervenir d'informations a priori. Le principe est de se baser uniquement sur des primitives décrivant l'aspect visuel de la séquence pour modéliser les actions et d'entrainer un classifieur à reconnaitre ces primitives. La méthode a été testée sur la base « MICC-Soccer-Actions-4 » [6] contenant quatre classes d'actions de football différentes : *Shot-on-goal*, *placed-kick*, *throw-in* et *goal-kick* (cf. figure 4). Les taux de classification obtenus sont de $52,75\ \%$ avec un classifieur K-NN et de $73,25\ \%$ avec un classifieur SVM en se basant sur des primitives visuelles. Néanmoins, cette méthode se base sur une représentation simpliste des actions vu qu'elle ne fait intervenir aucune notion de mouvement.

Dans cet article, nous proposons une nouvelle méthode de classification d'actions de sport. Dans la section suivante,

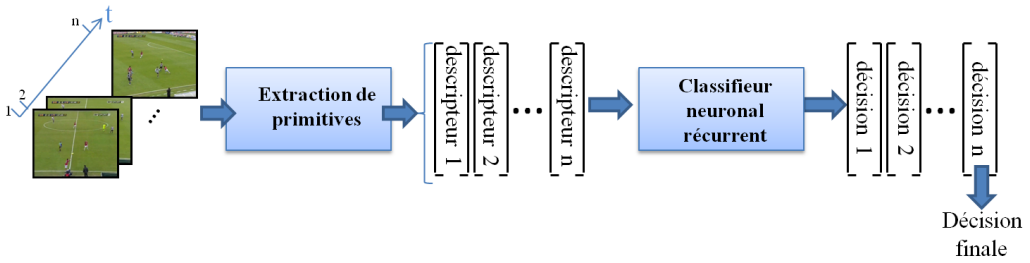


Figure 1 – *Schéma synoptique de la méthode proposée.*

nous présentons l'idée générale de la méthode proposée. Nous décrivons ensuite dans la section 3 les primitives visuelles introduites par Ballan et al. [6], et proposons un ensemble de caractéristiques décrivant le mouvement global à chaque instant de la vidéo, ainsi que la fusion des deux types de caractéristiques. Dans la section 4, nous présentons une méthode de classification, basée sur les réseaux de neurones récurrents. Enfin, nous présentons les résultats expérimentaux sur la base « MICC-Soccer-Actions-4 » dans la section 5.

2 Principe général de la méthode proposée

Le schéma général de la méthode proposée est présenté dans la figure 1. L'idée est de classer des séquences d'actions de sport issues d'une segmentation en plans. Nous calculons, pour chaque image de chaque séquence, un descripteur relatif à une primitive particulière (cf. section 3). Chaque séquence est alors représentée par une succession de descripteurs (un vecteur par image). Le nombre de ces descripteurs peut-être variable selon la longueur de la séquence. La classification est réalisée ensuite en considérant de manière dynamique l'évolution temporelle des primitives extraites. Concrètement, le classifieur devra considérer un à un les descripteurs et prendre une décision finale calculée à partir de plusieurs décisions individuelles accumulées tout au long de la séquence.

3 Extraction de primitives pour la représentation des actions de sport

Afin de décrire le contenu des séquences vidéos, en prenant en compte aussi bien leur aspect visuel que le mouvement, nous avons opté pour une approche par sacs de mots, puis nous avons introduit la notion de mouvement dominant.

3.1 Représentation du contenu visuel : une approche par sacs de mots

Les sacs de mots sont des modèles largement utilisés en traitement d'images en général, et en classification d'objets en particulier. Plusieurs travaux ont essayé d'étendre ces modèles au cas de la vidéo. Par exemple dans [7], les points d'intérêts 2D sont remplacés par des points d'intérêts spatio-temporels, et un modèle par sac de mots spatio-temporels est utilisé pour la classification d'actions humaines. Même si ce type d'approches donne de bons résultats pour les actions simples (action d'une seule personne), la généralisation aux cas des actions de sport, où les mouvements locaux sont en général trop complexes et non représentatifs de la scène, ne donne pas de résultats. Dans cet article, nous avons repris le modèle de Ballan et al. [6] dans lequel une vidéo est représentée par une séquence de sacs de mots visuels (un histogramme de mots SIFT par image). Le dictionnaire de mots est généré en appliquant une classification par k-moyennes sur un large nombre d'images extraites de la base. Le descripteur pour chaque image a donc la taille du dictionnaire et chaque valeur représente la fréquence d'occurrence du mot du dictionnaire dans l'image. Cette représentation permet à la fois de prendre en compte le contenu visuel de la vidéo, mais aussi de modéliser les transitions entre les images à travers l'apparition ou la disparition des mots.
Cette approche va aussi par la suite nous servir de base pour évaluer, sur la base *MICC-Soccer-Actions-4* et dans les mêmes conditions que [6], les performances de la classification neuronale (cf. sous-section 5.2).

3.2 Estimation du mouvement dominant par appariement de points SIFT

Nous proposons d'introduire également un autre type de primitive décrivant le mouvement dominant de la scène. Celui-ci est défini comme étant le mouvement représenté par le plus grand nombre d'éléments de cette scène. Typiquement, pour une action de sport avec une vue globale du terrain (ce qui est le cas pour les actions de la base *MICC-Soccer-Actions-4*), le mouvement dominant se confond avec celui de la caméra, et celui-ci est très caractéristique du type d'actions. L'idée est donc d'estimer ce mouvement puis de l'exploiter pour la classification. Nous avons fait l'hypothèse d'un mouvement affine de la caméra, ce qui est généralement vérifié. Le principe est donc d'estimer la transformation affine T qui permet de passer d'une image I_t d'une vidéo à l'image I_{t+1}.
Pour ce faire, nous effectuons un appariement des points SIFT entre deux images successives de la vidéo. Nous avons recours à l'algorithme *kd-tree* pour une recherche rapide du voisin le plus proche de chaque point. L'algorithme

RANSAC [8] est ensuite appliqué pour séparer le mouvement dominant (celui de la caméra) des mouvements locaux (ceux des joueurs par exemple). Enfin les N paires de points qui ont été considérées comme *inliers* (points conformes) sont utilisées pour estimer les paramètres de la transformation T.
Si l'on pose : $T = [a_1\ a_2\ a_3\ a_4\ t_1\ t_2]^T$ où les a_i sont les coefficients relatifs à la rotation et au facteur d'échelle et les t_i ceux relatifs à la translation, et si l'on note par $\left(x_i^{(t)}, y_i^{(t)}\right)$ pour $i \in \{1, \ldots, N\}$ les N *inliers* relatifs à l'image I_t, la relation entre $\left(x_i^{(t)}, y_i^{(t)}\right)$ et $\left(x_i^{(t+1)}, y_i^{(t+1)}\right)$ sera de la forme :

$$\begin{bmatrix} x_i^{(t+1)} \\ y_i^{(t+1)} \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} x_i^{(t)} \\ y_i^{(t)} \end{bmatrix} + \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}$$

En ré-écrivant l'équation précédente pour les N *inliers*, on peut se ramener à un système linéaire d'inconnue T et de la forme :

$$A\,T = B$$

avec :

$$A = \begin{bmatrix} & & \cdots & & & \\ x_i^{(t)} & y_i^{(t)} & 0 & 0 & 1 & 0 \\ 0 & 0 & x_i^{(t)} & y_i^{(t)} & 0 & 1 \\ & & \cdots & & & \end{bmatrix}$$

et :

$$B = \begin{bmatrix} \cdots \\ x_i^{(t+1)} \\ y_i^{(t+1)} \\ \cdots \end{bmatrix}$$

La résolution de ce système se fait par moindres carrés en décomposant la matrice A en valeurs singulières, puis en calculant sa matrice pseudo-inverse. La figure 2 montre un exemple d'appariement de points entre deux images issues de la même vidéo, ainsi que la compensation du mouvement affine estimé.
Pour éviter que les logos soient pris en considération dans l'estimation, nous effectuons un pré-traitement sur toutes les vidéos pour les détecter et les flouter. Pour ce faire, nous avons recours à une combinaison de deux approches. La première permet de détecter, à partir de l'analyse des statistiques de plusieurs images sélectionnées aléatoirement, les pixels immobiles. La deuxième se base sur l'approche introduite dans [9] et qui permet de détecter les textes horizontaux.
Une fois les transformations estimées, les six coefficients sont normalisés, colonne par colonne, entre -1 et 1. Pour cela, pour chaque colonne, nous calculons la moyenne m et l'écart type σ sur toute la base et nous normalisons et tronquons les valeurs en fixant les extremums à $m \pm 2\sigma$ afin de prendre en compte 98 % de la masse, en faisant l'hypothèse d'une distribution Gaussienne. Les descripteurs pour

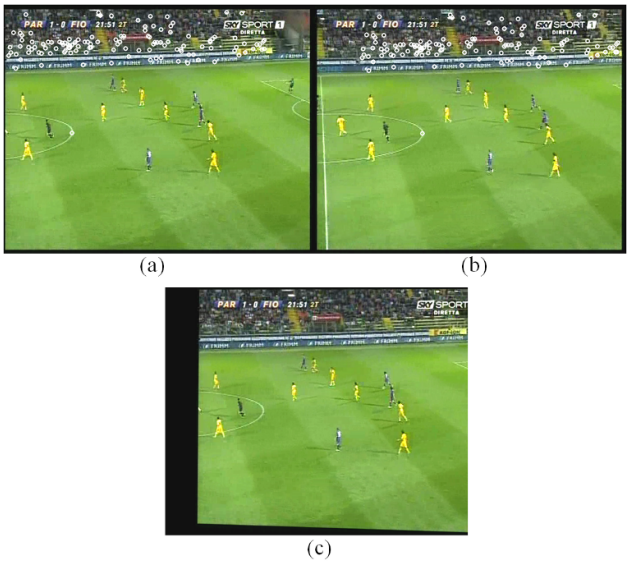


Figure 2 – *Exemple d'estimation du mouvement affine entre deux images : (a,b) - Inliers appariés entre les deux images (c) - Compensation du mouvement sur la première image.*

chaque image auront donc une taille de 6 valeurs. Enfin nous proposons de fusionner les deux primitives précédemment décrite pour alimenter le clasifieur avec des vecteurs de taille 36 (qui correspond à la concaténation entre les 6 coefficients précédents et le nombre de mots du dictionnaire).

4 Réseaux de neurones récurrents à base de LSTM pour la classification de séquences

Nous allons nous baser sur les descripteurs extraits pour chaque image représentant les primitives décrites dans la section précédente pour classer les actions de la base *MICC-Soccer-Actions-4*. Le choix de la méthode de classification est alors primordial. Dans [6], deux schémas de classification ont été testés, l'un basé sur les K-NN et l'autre sur les SVM. Dans cet article, nous avons opté pour une approche neuronale récurrente qui permet, comme présenté dans la section 2, d'analyser l'évolution temporelle de ces primitives.
Les réseaux de neurones récurrents (RNN : *Recurrent Neural Networks*) sont des réseaux possédant des connexions récurrentes qui permettent de prendre en compte à un instant t un certain nombre d'états passés. On parle alors de « mémoire à court-terme ». De ce fait, les RNN sont particulièrement adaptés aux applications faisant intervenir le contexte, et plus particulièrement au traitement des séquences.
Néanmoins, pour les applications faisant intervenir de longs écarts temporels (typiquement la classification de séquences vidéos), cette « mémoire à court-terme » n'est pas suffisante. En effet, les RNN « classiques » ne sont capables

de mémoriser que le passé dit « proche », et commencent à « oublier » au bout d'une cinquantaine d'itérations environ. Ce phénomène a été mis en évidence par Hochreiter et al. dans [10]. Les auteurs ont étudié plusieurs algorithmes d'apprentissage (BPTT, RTRL...) pour les RNN et montrent que l'erreur rétro-propagée liée à une entrée du réseau à l'instant t décroit de manière exponentielle après un certain nombre d'itérations. Ce phénomène est particulièrement problématique pour la classification de vidéos (contrairement à la classification de chaque image), puisque le réseau doit attendre la fin de la séquence avant de lui attribuer un label, et par conséquent, si la « mémoire à court-terme » est négligeable devant la taille de la séquence, la mise-à-jour des poids pendant l'apprentissage par rétro-propagation ne prendra en compte que les premiers instants.
Pour remédier à ce problème, Hochreiter et al. [10] ont mis au point des neurones particuliers, à large mémoire court-terme : les LSTM (*Long Short-Term Memory*). Intuitivement, un LSTM peut être vu comme un neurone ayant, en plus des connexions externes, une connexion auto-récurrente de coefficient constant égal à 1. Ceci permet de sauvegarder d'une itération à l'autre les états successifs du neurone. Des portes multiplicatives au niveau de l'entrée (*input gate*) et de la sortie (*output gate*) permettent de protéger respectivement l'état actuel de la mémoire et celui du reste du réseau. Dans cet article, nous allons nous appuyer sur l'architecture proposée par Gers et al. dans [11] (cf. figure 3), dans laquelle une nouvelle porte, dite « *forget gate* », permet de réinitialiser l'état de la mémoire au cours de la séquence.

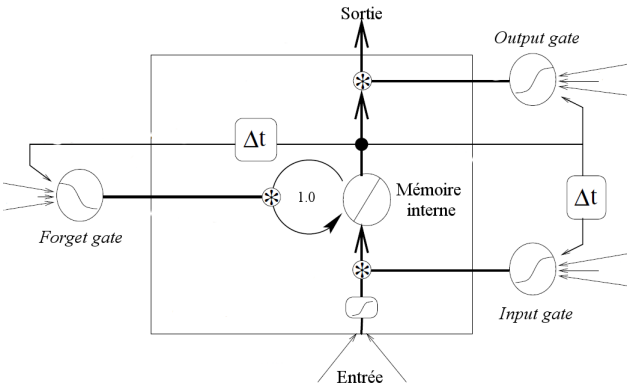


Figure 3 – *Exemple d'un neurone LSTM : architecture proposée dans [11].*

Les LSTM et ont été testés sur différentes applications (apprentissage de CSL, improvisation automatique de musique, classification de phonèmes, reconnaissance d'écriture manuscrite...) avec à chaque fois des résultats au moins aussi bons que ceux de l'état de l'art. Ils ont aussi été utilisés avec succès dans des applications de structuration de vidéos de tennis [12], pour modéliser l'évolution temporelle des transitions entre les plans de la vidéo sans en analyser le contenu. Dans cet article, nous proposons d'utiliser les LSTM pour analyser directement le contenu des vidéos.

5 Résultats expérimentaux

5.1 Données utilisées

Afin d'évaluer notre méthode de classification, nous avons effectué plusieurs tests sur la base publique [1] *MICC-Soccer-Actions-4* [6]. Cette base comprend 100 séquences vidéo au format MPEG-2 pleine résolution PAL (720×576 pixels, 25 images/seconde). La base contient quatre actions : *Shot-on-goal*, *placed-kick*, *throw-in* et *goal-kick* (cf. figure 4).

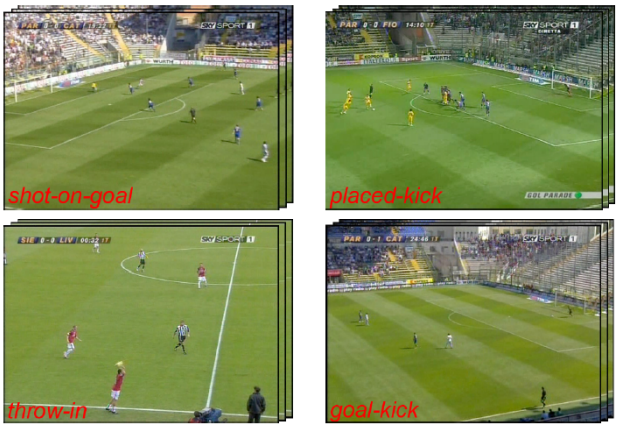


Figure 4 – *Les quatre actions constituant la base MICC-Soccer-Actions-4 [6] : Shot-on-goal, placed-kick, throw-in et goal-kick.*

Les séquences ont été générées par une détection de plans à partir de 5 vidéos de matchs, faisant intervenir 7 équipes et 5 stades différents, ainsi que des conditions d'éclairage assez variables (notamment 4 matchs se jouant sous un éclairage naturel à différentes heures de la journée, et le dernier sous un éclairage artificiel). Chacune des quatre classes est représentée par 25 séquences de longueurs variables, allant de 100 images à 2500 images. Même si toutes les séquences représentent des vues globales du terrain, la variabilité intra-classe reste très importante puisque les actions se déroulent suivant plusieurs scénarios différents.
Nous présentons dans un premier temps l'évaluation de la classification neuronale en comparant, dans les mêmes conditions, les résultats avec ceux présentés dans [6]. Puis nous étudierons l'apport du mouvement dominant. Enfin nous évaluerons notre proposition de fusionner les deux types de primitives. Toutes les expérimentations ont été effectuées par une validation croisée avec un partitionnement de la base en 3 parties (*3-fold cross validation*), selon la répartition décrite dans le tableau 1, de manière à ce que toutes les séquences soient testées une seule fois. Dans

1. Disponible sur : *www.micc.unifi.it/vim*

ce qui suit, les résultats présentés sont calculés en prenant compte les trois configurations.

	Apprentissage	**Test**
Config. 1	68 (17/classe)	32 (8/classe)
Config. 2	68 (17/classe)	32 (8/classe)
Config. 3	64 (16/classe)	36 (9/classe)

Tableau 1 – *Répartition du nombre de séquences entre « apprentissage » et « test » pour la validation croisée.*

5.2 Résultats

Evaluation de la classification neuronale basée sur les LSTM. Afin d'évaluer la classification neuronale, nous nous sommes placés dans les mêmes conditions que dans [6] pour pouvoir comparer les résultats. Ainsi, nous avons généré un dictionnaire de 30 mots visuels à partir d'une partie de la base (5 images extraites aléatoirement de chacune des 100 vidéos) avec une classification k-moyennes. Nous avons vérifié que l'augmentation de la taille du dictionnaire n'améliorait pas les résultats, mais augmentait considérablement la complexité, ce qui est en conformité avec les observations présentées dans [6]. Le dictionnaire est ensuite utilisé pour générer des histogrammes de mots visuels comme décrit dans la sous-section 3.1.
Pour le réseau, nous avons utilisé un RNN avec une couche en entrée de taille 30 (une entrée par mot visuel à chaque intervalle de temps de la stimulation du neurone), une couche de sortie de taille 4 (une sortie par classe) et une couche cachée comportant des neurones LSTM unidirectionnels totalement inter-connectés et connectés au reste du réseau. Nous avons noté qu'une augmentation importante du nombre de neurones LSTM conduisait à un sur-apprentissage du réseau (et augmentait considérablement la complexité). De même, un réseau de taille réduite conduit à une divergence de l'apprentissage. Nos expérimentations on montré que 150 LSTM pour la couche cachée est un bon compromis. Ainsi le nombre de poids à optimiser est de 109 654. Enfin, pour la couche de sortie, elle correspond à des fonctions d'activations de type *softmax*.
Le résultat de la classification est reporté sur le tableau 2. Pour comparaison, nous présentons aussi les résultats de Ballan et al. [6], qui correspondent respectivement à des classifications par *K-plus proches voisins* (K-NN) et *machine à vecteur de support* (SVM), les deux combinés avec une distance d'édition pour comparer des vecteurs de tailles différentes.
Le tableau 2 montre que l'approche neuronale est largement plus performante que les méthodes de type K-NN, et est comparable aux méthodes de type SVM. Même si le résultat est supérieur à celui obtenu par les méthodes basées sur les SVM, la différence n'est pas assez importante pour pouvoir généraliser. Néanmoins, les résultats permettent de valider cette approche.

Etude de l'apport du mouvement dominant. Nous allons évaluer l'apport du mouvement dominant pour la classification, dans un premier temps seul, puis en étudiant la possibilité d'une combinaison entre les deux informations. Pour ce faire, nous allons utiliser le même réseau que pour les tests précédents, en modifiant la taille de la couche d'entrée. Nous avons utilisé la répartition apprentissage / test décrite dans le tableau 1, en rajoutant des versions symétriques par rapport à la verticale pour les vidéos de la base d'apprentissage. Ceci permet au réseau d'apprendre pour chaque séquence, deux directions du mouvement. La figure 5-(b) montre la matrice de confusion relative à la classification basée sur le mouvement dominant.

	Taux de classification
k-NN [6]	52,75 %
SVM [6]	73,25 %
Méthode proposée	**76 %**

Tableau 2 – *Evaluation de la classification neuronale RNN-LSTM par rapport aux autres méthodes de classification utilisées dans [6].*

	Goal-kick	**Placed-kick**	**Shot-on-goal**	**Throw-in**
Goal-kick	**0,92**	0,08	0	0
Placed-kick	0,08	**0,8**	0	0,12
Shot-on-goal	0	0,2	**0,72**	0,08
Throw-in	0,12	0,12	0,16	**0.6**

(a)

	Goal-kick	**Placed-kick**	**Shot-on-goal**	**Throw-in**
Goal-kick	**0.64**	0.28	0.08	0
Placed-kick	0.08	**0.68**	0.08	0.16
Shot-on-goal	0.08	0	**0.88**	0.04
Throw-in	0.08	0	0.04	**0.88**

(b)

Figure 5 – *Matrices de confusion : (a) - Classification RNN-LSTM basée sur les sacs de mots visuels (b) - Classification RNN-LSTM basée sur le mouvement dominant.*

Les résultats de la classification sont comparables à ceux obtenus par les approches basées sur les sacs de mots (taux de classification de 77 %) et montrent que le mouvement de la caméra contient beaucoup d'informations discriminantes.
De plus, la comparaison avec la figure 5-(a) montre une complémentarité entre l'information visuelle et le mouvement dominant. En effet, les classes *throw-in* et *shot-on-goal* sont très adaptées à l'approche basée sur le mouvement de la caméra, vu que ce dernier est très caractéristique de l'une et de l'autre (mouvement quasi inexistant pour la première, et très caractéristique pour la deuxième, notamment à cause des zooms sur la cage de but). En revanche, les classes *goal-kick* et *placed-kick* sont caractérisées par des scénarios très variables (en terme de mouvement de la caméra) mais sont particulièrement adaptées à l'approche par le contenu visuel vu que dans les deux cas, l'ordre dans lequel apparaissent / disparaissent les mots est très caractéristique.

Nous proposons donc de concaténer les deux types de caractéristiques en entrée d'un réseau de neurones récurrent. Nous avons repris les expériences précédentes, dans les mêmes conditions, mais en concaténant les vecteurs d'entrée qui sont maintenant de taille 36 (Sacs de mots visuels + mouvement dominant). les résultats présentés dans la figure 6 montrent que ce système est capable de classer correctement 92 % des séquences.

	Goal-kick	Placed-kick	Shot-on-goal	Throw-in
Goal-kick	**1**	0	0	0
Placed-kick	0.04	**0.84**	0.08	0.04
Shot-on-goal	0	0.12	**0.88**	0
Throw-in	0.04	0	0	**0.96**

Figure 6 – *Matrice de confusion relative à la classification RNN-LSTM basée sur la combinaison entre les sacs de mots visuels et le mouvement dominant.*

6 Conclusion et discussion

	Taux de classification
Sacs de mots visuels + K-NN + Distance d'édition [6]	52,75 %
Sacs de mots visuels + SVM + Distance d'édition [6]	73,25 %
Sacs de mots visuels + RNN-LSTM	**76 %**
Mouvement dominant + RNN-LSTM	**77 %**
Sacs de mots visuels + Mouvement dominant + RNN-LSTM	**92 %**

Tableau 3 – *Récapitulatif des résultats obtenus.*

Dans cet article, nous nous sommes intéressés à la problématique de la classification des actions de sport. Pour ce faire, nous avons mis au point une méthode de classification neuronale prenant en compte aussi bien l'aspect visuel que le mouvement dominant. Les résultats de nos expérimentations sur la base « MICC-Soccer-Actions-4 », résumés sur le tableau 3, permettent de conclure que la classification neuronale à base de LSTM est aussi performante que celle basée sur les SVM, et dépasse clairement pour cette problématique les approches de type K-NN. De plus, nous avons démontré que le mouvement de la caméra est une information discriminante entre les classes, et permet à elle seule d'avoir des taux de classification équivalents à ceux utilisant les sacs de mots visuels ($\approx$ 77 %). Enfin, la complémentarité entre l'aspect visuel et le mouvement de la caméra a été prouvée, consuisant à un système de classification « hybride » des séquences vidéo capable d'obtenir un taux de classification de 92 %.
Plusieurs pistes peuvent être envisagées à l'issue de ce travail. Nous prévoyons d'abord de tester cette méthode sur d'autres sports que le football, afin de vérifier la généricité de l'approche. Une autre piste serait d'appliquer la méthode sur des données présentant des scénarios plus complexes, et en considérant plus de classes. Dans ce cas, d'autres informations de mouvement locaux pourraient être intégrées pour différentier entre les actions sémantiquement très proches (par exemple entre un « tir au but » et un « but »).

Références

[1] A. Ekin et A.M. Tekalp. Automatic soccer video analysis and summarization, Août 1 2003. US Patent App. 10/632,110.

[2] J. Assfalg, M. Bertini, C. Colombo, et A. Del Bimbo. Semantic annotation of sports videos. *IEEE MULTIMEDIA*, pages 52–60, 2002.

[3] Y. Gong, TS Lim, et HC Chua. Automatic Parsing of TV Soccer Programs. Dans *IEEE International Conference on Multimedia Computing and Systems*, pages 167–174, 1995.

[4] L.Y. Duan, M. Xu, et Q. Tian. Semantic Shot Classification in Sports Video. *Storage and retrieval for media databases 2003 : 22-23 January 2003, Santa Clara, California, USA*, 5021 :300, 2003.

[5] E. Kijak, G. Gravier, L. Oisel, et P. Gros. Audiovisual integration for tennis broadcast structuring. *Multimedia Tools and Applications*, 30(3) :289–311, 2006.

[6] L. Ballan, M. Bertini, A. Del Bimbo, et G. Serra. Action categorization in soccer videos using string kernels. Dans *Proc. of IEEE CBMI. Chania, Crete*, 2009.

[7] P. Dollár, V. Rabaud, G. Cottrell, et S. Belongie. Behavior recognition via sparse spatio-temporal features. *ICCV VS-PETS*, 2005.

[8] M. Fischler. RANSAC : A Paradigm for Model Fitting With Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6) :381–395, 1981.

[9] C. Wolf, J. Jolion, et F. Chassaing. Text Localization, Enhancement and Binarization in Multimedia Documents. Dans *Proc. of ICPR*, pages 1037–1040, 2002.

[10] S. Hochreiter et J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8) :1735–1780, 1997.

[11] F.A. Gers, N.N. Schraudolph, et J. Schmidhuber. Learning precise timing with LSTM recurrent networks. *The Journal of Machine Learning Research*, 3 :115–143, 2003.

[12] E. Delakis. *Structuration multimodale des vidéos de tennis en utilisant des modèles segmentaux*. Thèse de doctorat, Université de Rennes 1, 2006.

Analyse de Textures Dynamiques par décompositions spatio-temporelles : application à l'estimation du mouvement global

Sloven Dubois [1,2] Renaud Péteri [1] Michel Ménard[2]

[1] Laboratoire de **M**athématiques **I**mage et **A**pplications

[2] Laboratoire **I**nformatique **I**mage et **I**nteraction

Université de La Rochelle
Avenue Michel Crépeau, 17042 La Rochelle Cedex
FRANCE

{sloven.dubois01, renaud.peteri, michel.menard}@univ-lr.fr

Résumé

Le contexte de recherche de cet article est l'analyse et la caractérisation de textures dynamiques. Celles-ci peuvent souvent se modéliser comme la somme d'une onde porteuse se propageant à grand échelle et de phénomènes locaux oscillants. L'algorithme des Morphological Component Analysis (MCA), étendu au temps, est utilisé pour retrouver ces composantes. Nous définissons également une nouvelle stratégie de seuillage dans l'algorithme des MCA afin de réduire significativement les temps de calcul. Des résultats sur des vidéos réelles sont proposés. Cette méthode de décomposition est ensuite appliquée dans le cadre de l'estimation du flot optique à la recherche du mouvement dominant. Enfin des perspectives futures sont exposées.

Mots clefs

Analyse de textures dynamiques, décompositions spatio-temporelles, Morphological Component Analysis (MCA), extraction de mouvement.

1 Introduction

Un thème récent dans l'analyse de séquences d'images a pour objet l'extension des textures statiques au domaine temporel : les textures dynamiques. Celles-ci sont présentes couramment dans de nombreuses scènes naturelles : un drapeau dans le vent, des risées à la surface de l'eau, de la fumée, ou un escalator sont autant de textures dynamiques présentes dans les vidéos.
Sans être exhaustif, leur étude est une thématique active comportant de nombreux domaines comme la caractérisation [1, 2, 3], la synthèse [4] ou la segmentation [5].
Le contexte de nos travaux se situe dans le cadre de la caractérisation et l'analyse de ces textures dynamiques, dans un but d'indexation pour la recherche automatique dans des bases vidéos [6].
Une texture dynamique est composée de différents mouvements apparaissant à différentes échelles spatio-temporelles : par exemple, sur la figure 1.(a), un mouvement spatio-temporel lent du tronc et un mouvement spatio-temporel rapide des branches et du feuillage peuvent être observés. Caractériser efficacement les textures dynamiques implique de capturer ces comportements spatio-temporels.

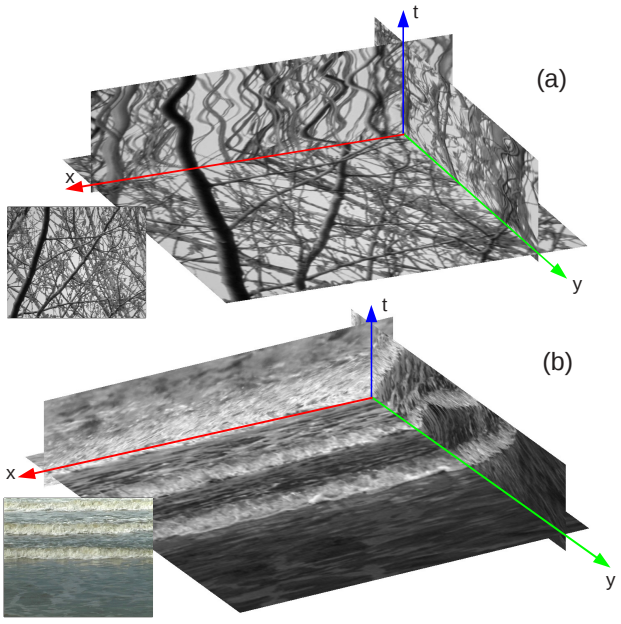


Figure 1 – *Sections 2D+T de textures dynamiques : on observe des mouvements à différentes échelles spatio-temporelles.*

Chaque texture dynamique possède donc ses propres caractéristiques, comme sa stationnarité, sa régularité, sa vitesse de propagation. Ces caractéristiques seront plus ou moins difficiles à extraire selon la complexité de la texture dynamique considérée. Par exemple, sur la figure 1.(b) représentant une séquence vidéo de la surface de la mer, deux mouvements peuvent être observés : un mou-

vement haute fréquence composé de l'écume porté par un mouvement d'ensemble, l'onde interne. Cette observation peut être étendue à de nombreuses textures dynamiques. On peut donc décomposer une texture dynamique, en une ou plusieurs ondes porteuses auxquelles s'ajoutent un ou plusieurs phénomènes localisés.

2 Décomposition de textures dynamiques

Devant la complexité des textures dynamiques, il est essentiel de décomposer ce phénomène afin d'en comprendre les différentes composantes pour ensuite les caractériser de manière pertinente. Les approches de décomposition d'images de la littérature [7, 8, 9] semblent donc adaptées pour l'extraction de ces composantes.
De part la richesse des bibliothèques d'analyse qu'elle permet d'utiliser, nous avons choisi l'approche Morphological Component Analysis (MCA). La diversité et l'intégration aisée de ces dernières, la souplesse de l'algorithme sont importantes au regard de la complexité des textures dynamiques. Ici, nous étendrons donc leur utilisation à la dimension temporelle.

L'approche MCA L'hypothèse de départ des Morphological Component Analysis est d'écrire un signal y comme une superposition linéaire de N composantes morphologiques perturbées par un bruit ε :

$$y = \sum_{i=1}^{N} y_i + \varepsilon \tag{1}$$

L'approche MCA permet donc de trouver une solution acceptable au problème inverse de la décomposition dans des bases, et d'extraire les composantes $(y_i)_{i=1,\dots,N}$ à partir de l'observation dégradée y selon un critère de parcimonie. Ceci suppose que chaque composante y_i est représentée de manière creuse dans une base associée Φ_i, c'est à dire $\forall i = 1,\dots,N, y_i = \Phi_i \alpha_i$. Ainsi, le dictionnaire représentant l'ensemble des bases est construit en associant plusieurs transformées $\Phi = [\Phi_1,\dots,\Phi_N]$ telles que pour chaque i, y_i est bien représenté dans Φ_i et n'est pas, ou alors très peu représenté dans Φ_j, $(j \neq i)$. Ceci se traduit par $\forall i, j \neq i, \|\Phi_i^T y_i\|_0 < \|\Phi_j^T y_i\|_0$ avec $\|\dots\|_0$ étant la pseudo-norme ℓ_0 (le nombre de coefficients non nuls). Résoudre cette équation revient à trouver une solution au système $y = \Phi\alpha$. Starck *et al.* proposent dans [10] et [9] de résoudre celui-ci, et de trouver les composantes morphologiques $(y_i)_{i=1,\dots,N}$ en résolvant le problème d'optimisation :

$$\min_{y_1,\dots,y_N} \sum_{i=1}^{N} \left\|\Phi_i^T y_i\right\|_p^p \quad \text{tel que} \quad \left\| y - \sum_{i=1}^{N} y_i \right\|_2 \leqslant \sigma \tag{2}$$

où le terme $\left\|\Phi_i^T y_i\right\|_p^p$. Ici nous prendrons $p = 0$ qui pénalise la non parcimonie de la solution. σ représente l'écart-type du bruit.
Ce problème d'optimisation (2) n'est pas facile à résoudre. Cependant, si toutes les composantes y_j sauf la i-ème sont fixées à l'itération $k-1$, alors il est prouvé que la solution $\alpha_i^{(k)}$ est donnée par un seuillage dur du résidu marginal $r_i^{(k)} = y - \sum_{j\neq i} y_j^{(k-1)}$:

$$\alpha_i^{(k)} = \delta_{\lambda^{(k)}}\left(\Phi_i^T\left(r_i^{(k)}\right)\right) \tag{3}$$

avec $\delta_{\lambda^{(k)}}$ l'opérateur de seuillage pour le seuil $\lambda^{(k)}$. Ces résidus marginaux r_i sont, par construction, susceptibles de contenir les informations marquantes de y_i. Cette idée dicte un algorithme itératif de seuillage sur les résidus marginaux (dont les opérations principales sont présentées dans l'algorithme 1).

Algorithme 1 Morphological Component Analysis

Boucle principale :
tant que $\left\|y - \sum_{j=1}^{N} \tilde{y}_j^{(k-1)}\right\|_2 \leqslant \sigma$ **faire**
// Pour chaque composante
pour $i = 1$ à N **faire**
// Calcul du résidu marginal
$\tilde{r}_i^{(k)} = y - \sum_{j\neq i} \tilde{y}_j^{(k-1)}$
// Projection du résidu marginal dans la base Φ_i
$\tilde{y}_i^{(k)} = \Phi_i\left(\delta_{\lambda^{(k)}}\left(\Phi_i^T\left(\tilde{r}_i^{(k)}\right)\right)\right)$
fin pour
// Incrémentation de l'itérateur k
$k = k + 1$
// Mise à jour du seuil λ
$\lambda^{(k+1)} =$ mise_à_jour($\lambda^{(k)}$,`stratégie`)
fin tant que

Choix du dictionnaire Le point crucial dans l'approche MCA est la définition du dictionnaire. Un choix non adapté des transformations par rapport à la dynamique des phénomènes présents dans la séquence est préjudiciable quant à la qualité du résultat : décomposition non pertinente, pseudo-norme ℓ_0 importante, coefficients non représentatifs. Comme nous l'avons observé dans la figure 1.(b), une texture dynamique peut se décomposer en deux phénomènes distincts. Il est donc nécessaire d'associer à chacun d'eux la base la plus représentative. Dans [11], les auteurs montrent que la transformée en curvelets [12] apporte une discrimination pertinente sur des phénomènes non locaux se propageant temporellement. Elle semble donc particulièrement intéressante pour modéliser les ondes porteuses présentes dans une texture dynamique. La deuxième partie d'une texture dynamique repose sur des phénomènes localement oscillants. Par conséquent, la deuxième base du dictionnaire est construite à partir d'une transformée locale adaptée aux oscillations. Nous avons proposé pour cette étude la transformée locale en cosinus qui semble la mieux adaptée.

Le dictionnaire Φ de décomposition des MCA est donc composé de la transformée en curvelets Φ_1 et de la transformée locale en cosinus Φ_2.

Stratégie de seuillage L'objectif de cette étude est la décomposition de textures dynamiques naturelles, et nous utiliserons dans nos expérimentations la base de données DynTex [13]. Les différentes séquences traitées ont une durée de 5 secondes (128 images) et une taille de 648 par 540 pixels [1]. Sur des volumes de cette taille, les transformées utilisées représentent un temps de calcul non négligeable. Certaines transformées nécessitent en effet plusieurs minutes de temps de calcul.
Dans [14], les auteurs s'accordent à dire qu'une centaine d'itérations est nécessaire à l'algorithme des MCA pour établir une bonne séparation des différentes composantes lorsqu'une stratégie de seuillage linéaire (SSL) est utilisée. Pour le dictionnaire choisi dans cette étude, ceci représente un temps de calcul pour une séquence d'images de : $100 * (T(\Phi_1^T) + T(\Phi_1) + T(\Phi_2^T) + T(\Phi_2))$, soit environ 21 heures, avec $T()$ mesurant le temps d'exécution d'une transformée Φ_i durant un cycle de l'algorithme (analyse via Φ_i^T et synthèse via Φ_i). Si nous étendons ce résultat à l'ensemble des séquences de la base de données DynTex, et toujours pour une durée de séquences de 5 secondes, nous obtenons environ 612 jours de calcul pour effectuer correctement la décomposition sur un ordinateur classique. Dernièrement, Bobin *et al.* ont proposé une stratégie de seuillage 'Mean of Max', SSMoM [14] qui conduit à des résultats équivalents mais avec un nombre d'itérations moindre (50 en moyenne au lieu de 100). Ceci représente un temps de calcul d'environ 10 heures 30 pour une de nos séquences vidéo, conduisant à environ 306 jours pour l'ensemble de la base.
Pour l'indexation d'une base comme DynTex [13], les temps de calcul de la stratégie SSMoM restent acceptables, puisqu'il est toujours possible de répartir la charge de calcul sur plusieurs unités. Dans le cadre de la recherche d'une texture particulière à l'aide d'une séquence requête, ces calculs ne peuvent se faire actuellement qu'à partir de séquences de durée limitée et de résolution faible. Un des objectifs de ce travail est donc de diminuer ces contraintes en proposant de nouvelles stratégies de seuillage.
La qualité des résultats de la décomposition d'un signal, à l'aide de l'algorithme des MCA, dépend fortement de l'évolution du seuil $\lambda^{(k)}$ au cours d'une itération de la boucle principale. Nous montrons sur la figure 2 deux évolutions différentes de $\lambda^{(k)}$ correspondant à deux stratégies (S1) et (S2). L'évolution de $\lambda^{(k)}$ est plus lente dans le cas (S1) que dans celui de (S2). Dans cet exemple, l'évolution (S1), respectivement (S2), conduit à répartir 5% de la plage des coefficients, respectivement 25%, sur les deux bases. Si on considère que l'évolution (S1) est dans cet exemple optimale en terme de seuillage, une évolution non maîtrisée de la valeur de $\lambda^{(k)}$ (cas (S2)) amènera à répartir trop rapidement un grand nombre de coefficients dans les bases, dégradant ainsi la décomposition.

1. soit plus de 44 millions de voxels

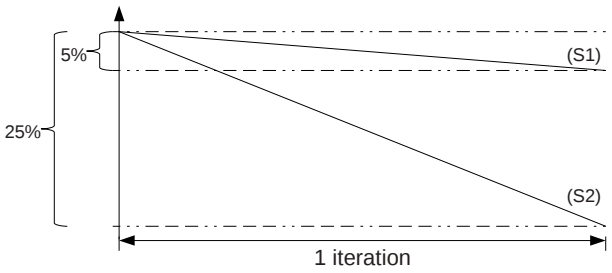


Figure 2 – *Deux stratégies de seuillage conduisant à des évolutions différentes de la valeur de seuil durant une itération de la boucle principale de l'algorithme des MCA.*

La stratégie de seuillage linéaire, SSL, conduit à une évolution optimale de $\lambda^{(k)}$ lorsqu'une centaine d'itérations est fixée [14]. Dans un grand nombre de textures naturelles, nous avons constaté que le nombre d'itérations peut être fortement diminué, et dépend de la texture elle-même, SSL n'est alors plus optimum. Cependant, l'évolution selon SSL peut être considérée comme une pente minimum en dessous de laquelle l'évolution de $\lambda^{(k)}$ sera sous-optimum. Une bonne stratégie pour le calcul de $\lambda^{(k)}$ doit donc conduire à une pente supérieure ou égale à celle engendrée par SSL. La stratégie 'Mean of Max', SSMoM, est très intéressante car elle permet de modifier l'évolution de $\lambda^{(k)}$ lorsque cela s'avère nécessaire. Sur les séquences de textures naturelles, cette stratégie a cependant souvent tendance à diminuer fortement la pente, voire presque à l'annuler.

Stratégie de seuillage adaptative à correction linéaire Nous proposons de rassembler les stratégies SSL et SSMoM en une nouvelle dite adaptative à correction linéaire, SSAcL, qui définit $\lambda^{(k)}$ comme le minimum des valeurs de $\lambda^{(k)}$ calculées par les stratégies SSL et SSMoM. SSAcL se formalise donc ainsi :

$$\lambda^{(k+1)} = \min\left(\frac{1}{2}(m_1 + m_2), \lambda^{(k)} - \frac{\lambda^{(1)} - \lambda_{\min}}{100}\right) \quad (4)$$

avec $m_1 = \max_{\forall i} \left\|\Phi_i^T r^{(k)}\right\|_\infty$, $m_2 = \max_{\forall j, j\neq i} \left\|\Phi_j^T r^{(k)}\right\|_\infty$ et $r^{(k)} = y - \sum_{j=1}^{K} \tilde{y}_j^{(k)}$ le résidu total. Ainsi, avec cette stratégie, nous sommes assurés de modifier la valeur de $\lambda^{(k)}$ selon la plus grande pente. Autrement dit, lorsque SSMoM conduit à des valeurs de $\lambda^{(k)}$ évoluant faiblement, les valeurs de $\lambda^{(k)}$ suivent la stratégie SSL, $\lambda^{(k)} - \frac{\lambda^{(1)} - \lambda_{\min}}{100}$. Sinon, $\lambda^{(k)}$ suit la stratégie SSMoM, $\frac{1}{2}(m_1 + m_2)$, et permet ainsi de diminuer le nombre d'itérations de l'algorithme 1.

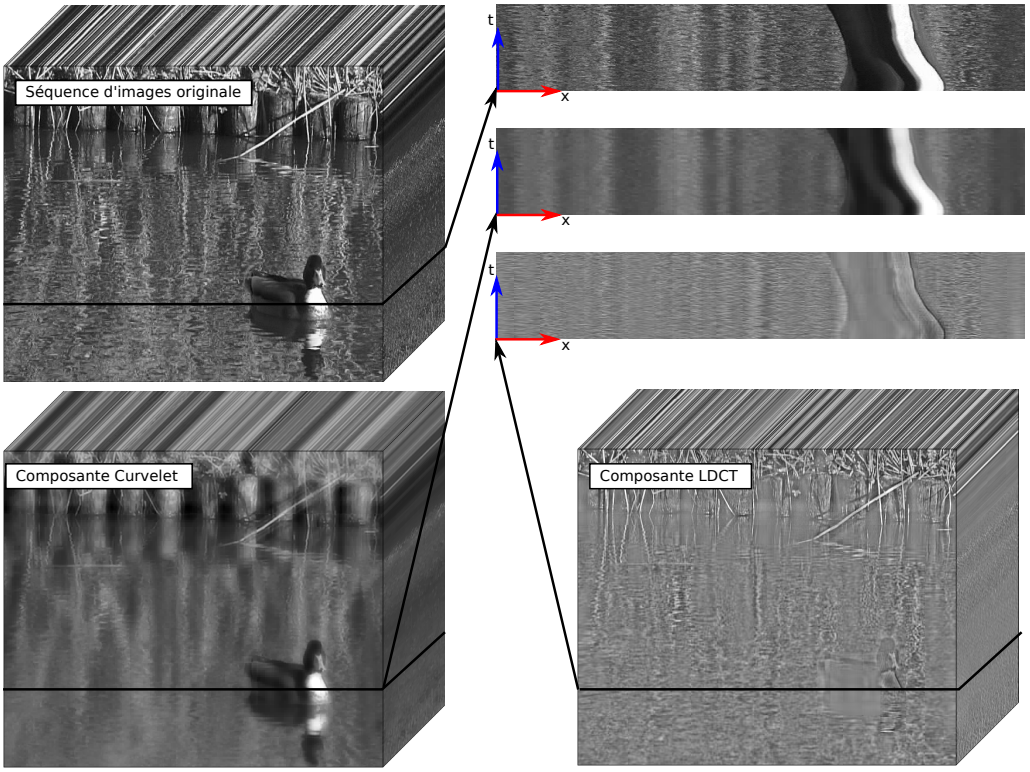


Figure 3 – *Résultat de la décomposition d'une vidéo à l'aide de l'algorithme des MCA et de la stratégie SSAcL. Une coupe spatio-temporelle xt est réalisée sur chacune des vidéos afin d'observer l'aspect temporel.*

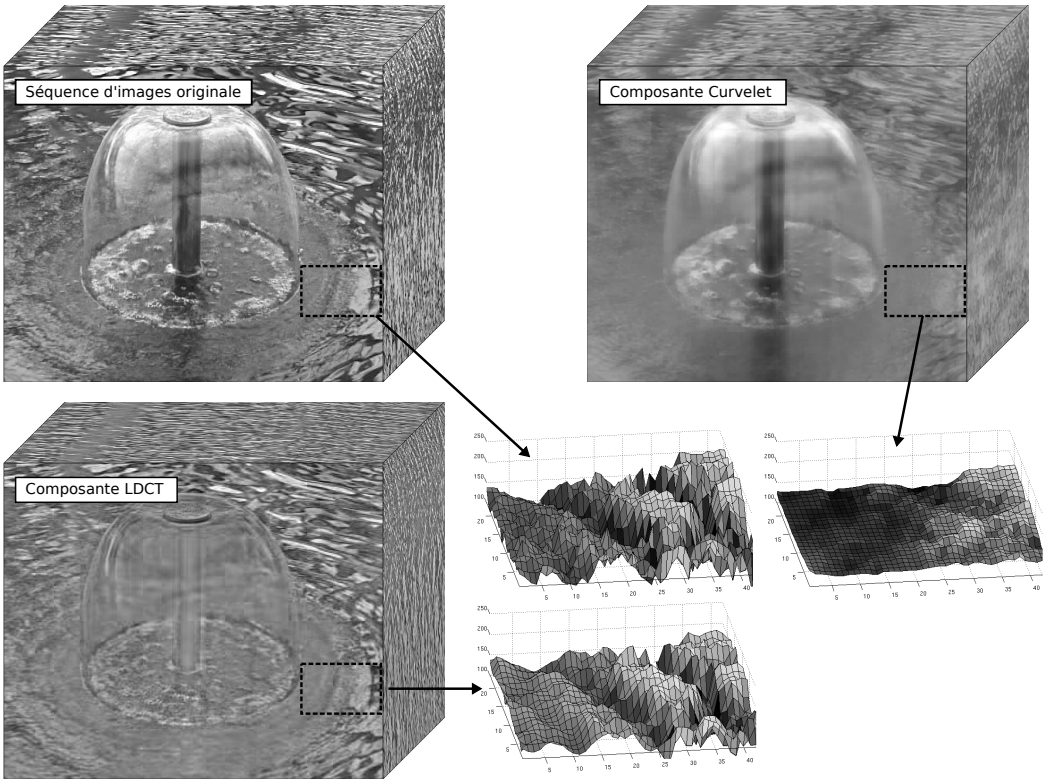


Figure 4 – *Résultat de la décomposition d'une vidéo à l'aide de l'algorithme des MCA et de la stratégie SSAcL. Des zones d'intérêt sont présentées sous forme de surface afin de mieux visualiser les résultats de décomposition.*

3 Résultats

La mise en place de la stratégie SSAcL a été appliquée aux séquences de la base DynTex et a permis de grandement réduire le temps de calcul (environ 2 heures par vidéo). Dans cette partie, deux d'entre elles seront détaillées [2].
La première vidéo est celle d'un lac sur lequel dérive lentement un canard (figure 3). Des reflets d'arbres dans l'eau ondulante et un fond statique texturé sont également observables. La figure présente le résultat de la décomposition obtenue sur cette vidéo à l'aide de l'algorithme des MCA et de la stratégie SSAcL. Nous retrouvons la composante géométrie grâce à la transformée en curvelet et la composante texture obtenue par la transformée locale en cosinus. Les vaguelettes, qui sont des phénomènes locaux, sont bien capturées par la composante texture, alors que nous retrouvons la surface de l'eau et les reflets dans la composante géométrie. Les coupes spatio-temporelles selon l'axe xt permettent de visualiser l'impact de la décomposition. Elles montrent que les différents objets de la scène (le canard, les reflets d'arbres,...) sont considérés comme de la géométrie. Les reflets des arbres ne sont pas présents dans la composante texture. On peut par contre les observer dans la composante géométrie : les oscillations qui se superposaient ont été grandement atténuées. La décomposition nous apporte également des informations qui n'étaient pas visibles sur la séquence d'images originale. Par exemple, dans la composante texture, on peut observer sous le cou du canard, la granularité du plumage invisible dans la vidéo originale.
La séquence d'images suivante représente une fontaine (figure 4). Celle-ci consiste en un jet d'eau qui une fois expulsé, vient agiter la surface de l'eau en créant des vaguelettes. Les résultats de la décomposition à l'aide de l'algorithme des MCA 2D+T sont également affichés sur la figure 4. Les deux composantes obtenues semblent pertinentes : dans la partie géométrie, la colonne centrale du jet et la forme en cloche engendrée par le jet sont visibles, alors que qu'elles sont quasiment absentes de la composante texture. On remarque également que toute la zone située devant le jet est dépourvue des vaguelettes, observables par contre dans l'autre composante. Cette observation est également très remarquable dans les zones misent en valeur par la représentation surfacique. En effet, la partie géométrie est privée des vaguelettes et seule une vague de faible amplitude est observable.
Dans les résultats que nous venons de présenter, la décomposition à l'aide de l'algorithme des MCA, étendue à la dimension temporelle, permet de bien extraire les différents phénomènes complexes présents dans une texture dynamique. Ces constatations s'observent sur d'autres vidéos de la base de données DynTex.

2. Ces vidéos et les résultats sont visibles à l'adresse : `http://mia.univ-larochelle.fr/demos/dynamic_textures/`

4 Application

La décomposition d'une texture dynamique en une composante géométrique et une composante texture nous apporte visuellement une meilleure compréhension des différents phénomènes. Afin d'aller plus loin dans l'analyse, nous allons appliquer la décomposition à la recherche du mouvement principal d'une texture dynamique. Le mouvement dans la vidéo est calculé à partir de l'algorithme de Horn et Schunk [15]. Le flot optique est estimé directement sur la vidéo originale et sur la composante géométrique.
Les résultats du flot optique estimé (présentés sur la figure 5) sont illustrés sur une séquence de mer sur laquelle des vagues et de l'écume sont observables. Afin de représenter le flot optique, nous utilisons deux systèmes de visualisation. Un champ de vecteurs coloré où la couleur (respectivement la saturation) indique la direction (respectivement l'intensité) du flot optique, et un critère d'homogénéité du champ de vecteurs, présenté dans [2]. Nous utiliserons un diagramme des orientations afin d'étudier le mouvement global. Dans le cas de la séquence originale, on remarque que tous les phénomènes dynamiques locaux sont extraits. En effet, aucune couleur ne se démarque des autres, les mouvements estimés sont principalement dus à l'écume, turbulente et sans direction privilégiée. Lorsque le calcul du flot optique s'effectue sur la composante géométrique de la vidéo, on observe une seule direction présente, celle des vagues principales (front d'onde). L'histogramme des orientations, bien plus isotrope dans le cas de la séquence originale que dans celui de la composante géométrique, indique bien cette direction.

5 Conclusion et perspectives

De nombreuses textures dynamiques peuvent se modéliser comme des ondes se propageant à grande échelle auxquelles s'ajoutent des phénomènes oscillants locaux. Nous montrons dans cet article que l'approche des MCA, étendue au temps, est très bien adaptée à la décomposition en ces différents phénomènes. Elle souffre cependant d'un temps de calcul important. Après avoir précisé les différents dictionnaires utilisés, nous proposons une nouvelle stratégie de seuillage adaptatif, SSAcL, conduisant à un gain important en terme de temps de calcul : par rapport à la stratégie originelle, nous réduisons d'environ un facteur 5 les temps de calculs nécessaires. Dans le cadre de l'indexation de textures dynamiques, ceci permet de relâcher quelque peu les contraintes de faible résolution et de durée lors de requêtes se présentant sous la forme d'une séquence d'images. Les résultats présentés dans la dernière section mettent en évidence les différents phénomènes complexes présents dans une texture dynamique. Enfin, cette décomposition est appliquée à l'estimation du mouvement global. Cette application montre que l'utilisation de la composante géométrique 2D+T permet une estimation plus robuste du mouvement dominant. On peut donc penser que les différentes composantes obtenues à l'aide de l'algorithme des MCA permettront d'ex-

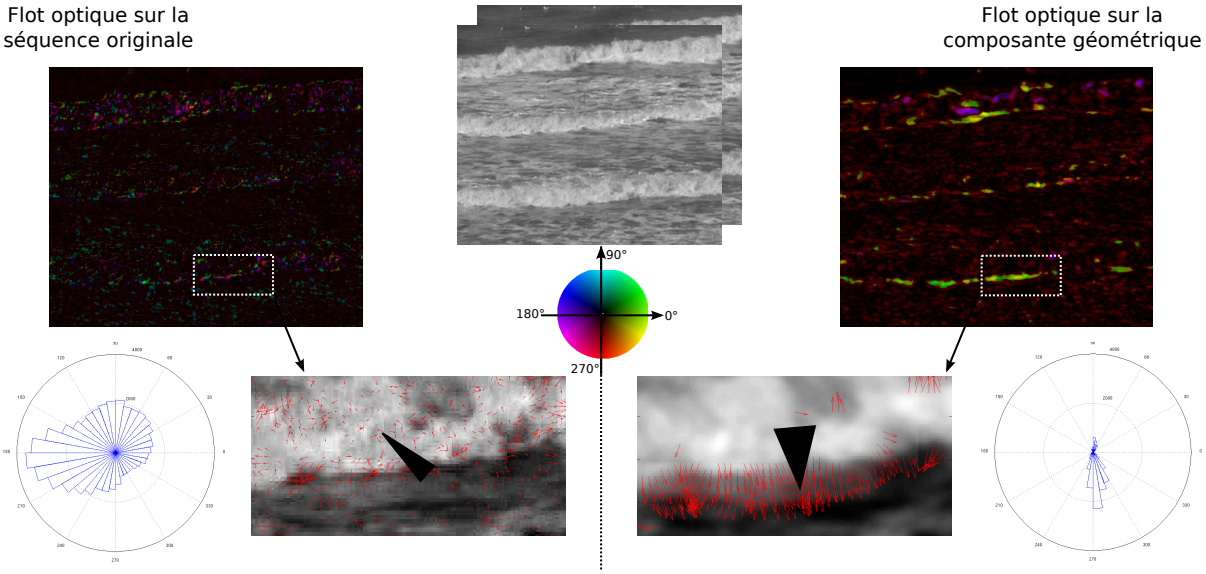


Figure 5 – *Résultats du flot optique Horn-Shunk sur une vidéo originale et sur la composante géométrique de cette vidéo. Les triangles noirs des images du dessous représentent la direction moyenne et l'écart-type du champ de vecteurs estimé.*

traire des signatures discriminantes : signatures liés à la géométrie de la texture dynamique (direction principale du mouvement, homogénéité du mouvement global, ...) et signatures caractérisant les phénomènes plus locaux (turbulence, mouvement de convection ...). Plusieurs autres applications de cette décomposition spatio-temporelle sont envisageables, comme annuler ou modifier la dynamique de certains phénomènes, ou bien masquer certains objets évoluant sur des fonds dynamiques (par exemple éliminer la présence du canard de la vidéo de la figure 3).

Références

[1] R.C. Nelson et R. Polana. Qualitative recognition of motion using temporal texture. *CVGIP : Image Underst.*, 56(1) :78–89, 1992.

[2] R. Péteri et D. Chetverikov. Qualitative characterization of dynamic textures for video retrieval. Dans *ICCVG 2004*, volume 32, pages 33–38, 2004.

[3] P. Saisan, G. Doretto, Y.N. Wu, et S. Soatto. Dynamic texture recognition. Dans *CVPR'01*, volume 2, pages 58–63, Kauai, Hawaii, December 2001.

[4] G. Doretto, A. Chiuso, Y.N. Wu, et S. Soatto. Dynamic textures. *Int. Journal of Computer Vision*, 51(2) :91–109, February 2003.

[5] G. Doretto, D. Cremers, P. Favaro, et S. Soatto. Dynamic texture segmentation. Dans *ICCV'03*, volume 2, pages 1236–1242, 2003.

[6] S. Dubois, R. Péteri, et M. Ménard. A comparison of wavelet based spatio-temporal decomposition methods for dynamic texture recognition. Dans *IbPRIA'09*, volume 5524, pages 314–321, Povoa de Varzim, Portugal, 2009.

[7] T.F. Chan, S. Osher, et J. Shen. The digital tv filter and nonlinear denoising. *IEEE Transactions on Image Processing*, 10(2) :231–241, 2001.

[8] J.F Aujol et A. Chambolle. Dual norms and image decomposition models. *Int. J. Comput. Vision*, 63(1) :85–104, 2005.

[9] J.L. Starck, M. Elad, et D.L. Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE Trans. on Image Processing*, 14 :1570–1582, 2005.

[10] J-L. Starck, M. Elad, et D.L. Donoho. Redundant multiscale transforms and their application for morphological component analysis. *Advances in Imaging and Electron Physics*, 132, 2004.

[11] S. Dubois, R. Péteri, et M. Ménard. A 3D discrete curvelet based method for segmenting dynamic textures. Dans *ICIP'09*, pages 1373–1376, Cairo, Egypt, November 2009.

[12] E. Candès, L. Demanet, D.L. Donoho, et L. Ying. Fast discrete curvelet transforms. Rapport technique, California Institute of Technology, 2005.

[13] Renaud Péteri, Sándor Fazekas, et Mark J. Huiskes. DynTex : a Comprehensive Database of Dynamic Textures. *Pattern Recognition Letters*, doi : 10.1016/j.patrec.2010.05.009. http ://projects.cwi.nl/dyntex/.

[14] J. Bobin, J-L. Starck, J.M. Fadili, Y. Moudden, et D.L. Donoho. Morphological component analysis : An adaptive thresholding strategy. Dans *IEEE Trans. on image processing*, pages 2675–2681. IEEE, 2007.

[15] B.K.P. Horn et B.G. Schunck. Determining optical flow. *Journal AI*, 17 :185–203, 1981.

Bayesian Fusion of Visible Cameras for Behaviour Recognition

Julien Ros[1] Kamel Mekhnacha[1]

[1] Probayes SAS

345, rue Lavoisier - Inovallée
38330 Montbonnot – FRANCE

{julien.ros, kamel.mekhnacha}@probayes.com

Abstract

The utilisation of several cameras to monitor human activity in a large space is essential due to the important field of view to be covered and the possible cluttered environment. The interpretation of this high number of data requires fast and powerful fusion algorithms in order to make easier the next human or computer work. In this paper the utilisation of a probabilistic occupancy map is proposed to fuse videos coming from different cameras. By estimating the occupancy and the velocity of each spatial cell representing the environment and obtained thanks to a background substraction algorithm, it is shown that human can be efficiently tracked. The tracking information is finally successfully used by a bayesian filter to recognise low level pedestrian behaviour such as standing, walking and running.

Keywords

Bayesian Occupancy Filter, visible camera fusion, behaviour recognition.

1 Introduction

Nowadays, employing video cameras to monitor a place has become very popular. Cameras can be used in home applications for power saving while facilitating the user everyday life; in supermarkets, to increase the average individual sales by bringing some interactivity and in the security domain to detect abnormal situations. Considering the security market in the United Kingdom and according to a study of urbaneye[1], there were approximately 4,200,000 cameras in 2002 in UK which represents one camera for every 14 people. Efficient systems used to facilitate the interpretation of this high quantity of information should thus be found. Indeed, multiple sensors could provide more reliable and robust information about the environment. As a consequence, these applications should be able to fuse different inputs provided by different sensors in order to make decisions about the current situation monitored. In practice, a videosurveillance application often requires four main steps (see Figure 1): human detection, sensor fusion, human tracking, and human behaviour recognition. This

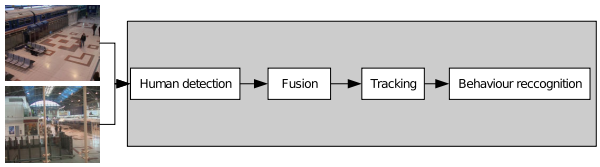


Figure 1 – *Multi-camera Human Behaviour Recognition System Architecture*

paper deals with all steps. Due to the sensor measurements which are noisy, it's seldom possible to construct an exact representation of the environment monitored by the different sensors. Thus, probabilistic approaches are usually used to cope with this problem and especially those using occupancy grids [1, 2, 3, 4, 5].

Among them, the Bayesian Occupancy Filter (BOF) introduced in [6], improved in [7], has already been used to perform videosurveillance task by fusing visible and infrared cameras in [8]. For this purpose, the BOF employs deeply the Bayesian approach in order to perform sensor fusion for the occupancy map estimation. It proposes to use a global filtering equation to estimate both the occupancy and the velocity of a given grid cell. The occupancy map is then given as input to a clustering algorithm to extract objects tracked in the next step.

Once human are correctly tracked, it is possible to perform human behaviour recognition and for this purpose, Bayesian approach are generally employed [9, 10, 11]. However, they often require a precise limb description of the human tracked difficult to obtain in large area surveillance or they are based on a training phase difficult to handle when the training set is too small. In this paper, we show that the position and velocity tracking information provided by the BOF are enough to employ a Bayesian filter to detect basic behaviour such as standing, walking and running.

The paper is organized as follows. Section 2 presents how the information provided by the human detection algorithm are fused into a single Bayesian occupancy map in order to

[1] http://www.urbaneye.net/results/ue_wp6.pdf

return track associated to each pedestrian monitored. Section 3 describes the Bayesian filtering process employed to detect low level behaviours. Experiments are presented in Section 4 and finally Section 5 concludes the paper.

2 Bayesian Occupancy Filter (BOF)

The Bayesian Occupancy Filter (BOF) is represented as a two-dimensional grid-based decomposition of the environment. Each cell of the grid contains two probability distributions: (i) the probability distribution over the occupancy of the cell (ii) and the probability distribution over its velocity. Given a set of input sensor readings, the BOF algorithm allows to update the occupancy/velocity estimates of each grid cell.
Figure 2 shows an example of BOF output using a computer vision car detector.

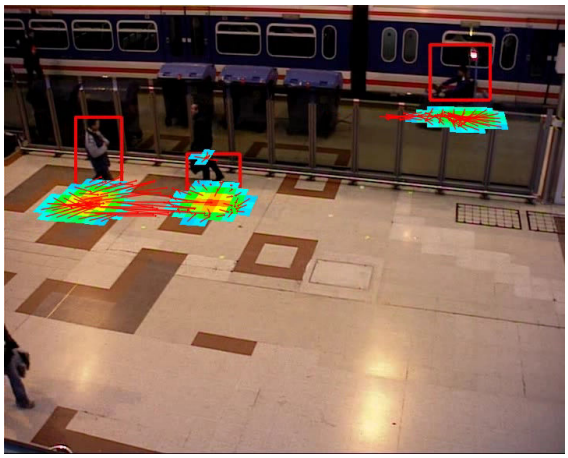


Figure 2 – *Example of BOF output using a computer vision pedestrian detector as input (red boxes). The BOF output is projected back on the image. It represents a grid of occupancy probability (blue-to-red mapped color) and the mean velocity (red arrows) estimates.*

A more detailed description of the BOF framework should be found in [12]. The BOF model is shown graphically in Fig. 3 and is described as follows:

2.1 Variables

For a given cell having $c \in \mathcal{Y}$ as index in the grid, let:

- $A_c^t \in \mathcal{A}_c \subset \mathcal{Y}$ represents each possible antecedent of cell c over all the cells in the grid domain $\mathcal{Y}$. The set of antecedent cells of cell c is denoted by $\mathcal{A}_c$ and is defined as a neighbourhood of the cell c.
- $A_c^{t-1} \in \mathcal{A}_c \subset \mathcal{Y}$ the same as A_c^t but for the previous time step.
- $O_c^t \in \mathcal{O} \equiv \{0, 1\}$ is a boolean variable representing the state of the cell in terms of occupancy at time t, either $[O_c = 1]$ if occupied, $[O_c = 0]$ if empty. Given the independency hypothesis, the occupancy of each cell at time t is considered apart from the occupancy of its neighbouring cells at time t.
- $Z_i^t \in \mathcal{Z}, 1 \leq i \leq S \in \mathbb{N}$, is a generic notation for measurements yielded by each sensor i, considering a total of S sensors yielding a measurement at the considered time instant.

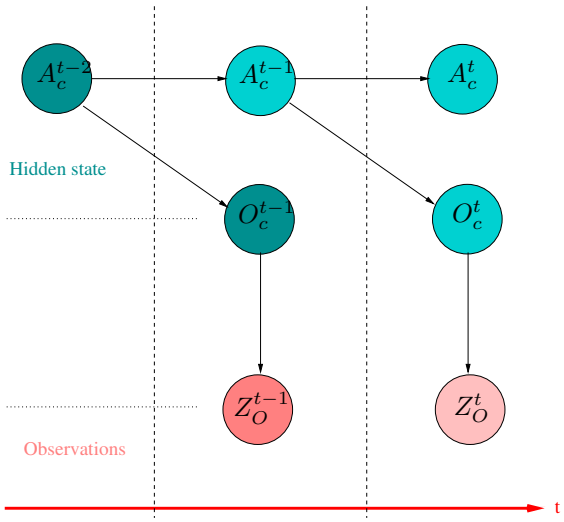


Figure 3 – *The Dynamic Bayesian Network corresponding to the BOF model for each grid cell c. Here, we suppose that only occupancy sensors are available.*

2.2 Joint distribution factors

The following expression gives the decomposition of the joint distribution of the relevant variables according to Bayes' rule and dependency assumptions:

$$P(A_c^{t-1}\, A_c^t\, O_c^t\, Z_1^t \cdots Z_S^t) =$$
$$P(A_c^{t-1}) P(A_c^t \mid A_c^{t-1}) P(O_c^t \mid A_c^{t-1}) \prod_{i=1}^{S} P(Z_i^t \mid A_c^t\, O_c^t) \quad (1)$$

The parametric form and semantics of each component of the joint decomposition are as follows:

- $P(A_c^{t-1})$ is the probability for a given neighbouring cell A_c to be the antecedent of c at time $t-1$. In order to represent the fact that cell c is *a priori* equally reachable from all possible antecedent cells in the considered neighbourhood, this probability table is initialized as uniform and is update in each time step.
- $P(A_c^t \mid A_c^{t-1})$ is the distribution over antecedents at time t given the antecedent of cell c at $t-1$. It represents the prediction (dynamic) model over velocity. If we assume a perfect *constant velocity hypothesis* between the two time frames $t-1$ and t, this distribution is simply:

$$P(A_c^t \mid A_c^{t-1}) = P(A^{t-1}_{A_c^{t-1}}).$$

In other words, the predicted probability is simply the probability at the preceding time instant for the antecedent at $t-1$.

Considering imperfect *constant velocity hypothesis* is possible by introducing the predicate $E \in \{0,1\} \equiv$ "There was an erroneous prediction", and assuming a probability $P(E) = \epsilon$. This value is a parameter of the system and corresponds of the probability of not respecting the *constant velocity hypothesis*. We have:

- $P(A_c^t \mid A_c^{t-1}\ \neg E) = P(A^{t-1}_{A_c^{t-1}})$,
- $P(A_c^t \mid A_c^{t-1}\ E) = \mathcal{U}(A_c^t)$,

where $\mathcal{U}(A_c^t)$ denotes a uniform distribution on A_c^t to say that all possible antecedents have the same probability when *constant velocity hypothesis* is not respected. Thus, $P(A_c^t \mid A_c^{t-1})$ may be written as a mixture:

$$P(A_c^t \mid A_c^{t-1}) = P(\neg E)P(A_c^t \mid A_c^{t-1}\ \neg E) + P(E)P(A_c^t \mid A_c^{t-1}\ E).$$

Which leads to:

$$\begin{aligned} P(A_c^t \mid A_c^{t-1}) &= (1-\epsilon)P(A^{t-1}_{A_c^{t-1}}) + \epsilon\,\mathcal{U}(A_c^t) \\ &= (1-\epsilon)P(A^{t-1}_{A_c^{t-1}}) + \epsilon/\|\mathcal{A}_c\|, \end{aligned}$$

where $\|\mathcal{A}_c\|$ is the cardinality of the considered antecedents set $\mathcal{A}_c$.

- $P(O_c^t \mid A_c^{t-1})$ is the distribution over occupancy given the antecedent of cell c at $t-1$. It represents the prediction (dynamic) model over occupancy. Similarly to $P(A_c^t \mid A_c^{t-1})$, the term $P(O_c^t \mid A_c^{t-1})$ may be written as a mixture:

$$\begin{aligned} P(O_c^t \mid A_c^{t-1}) &= (1-\epsilon)P(O^{t-1}_{A_c^{t-1}}) + \epsilon\,\mathcal{U}(O_c^t) \\ &= (1-\epsilon)P(O^{t-1}_{A_c^{t-1}}) + \epsilon/2. \end{aligned}$$

- $P(Z_i^t \mid A_c^t\, O_c^t)$ is the *direct model* for sensor i. It yields the probability of a measurement given the occupancy O_c^t and the antecedent (velocity) A_c^t of cell c. Measurements for all sensors are assumed to have been taken *independently from each other*. For sensors providing measurements depending exclusively of occupancy, this distribution can be written as $P(Z_i^t \mid O_c^t)$. In the same manner, for sensors providing measurements depending exclusively of velocity, this distribution can be written as $P(Z_i^t \mid A_c^t)$.

2.3 Occupancy and velocity estimation using the BOF model

At each time step, the estimation of the occupancy and velocity of a cell is answered through Bayesian inference on the model given in Equation (1). This inference leads to a Bayesian filtering process (Fig. 4).

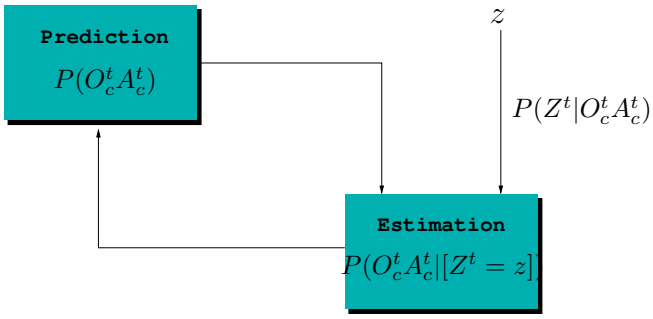


Figure 4 – *Bayesian filtering in the estimation of occupancy and velocity distribution in the BOF grids.*

2.4 A Gaussian Image Model Associated To Visible Cameras

It is generally assumed that the first step involved in a video surveillance application is the human detection. In this paper, we consider system where the cameras involved are static, humans are thus considered as moving regions in front of a relatively static background and a simple and inexpensive method to perform the detection is to use a background subtraction algorithm including foreground discrimination and blob segmentation.
To use these blobs as input to the BOF, a sensor model $P(Z_i^t \mid O_c^t)$ that takes as input the human bounding boxes is employed. This sensor model considers as sensor observations the bounding boxes and project them on the grid thanks to the calibration information in a Gaussian way as shown in Figure 5.

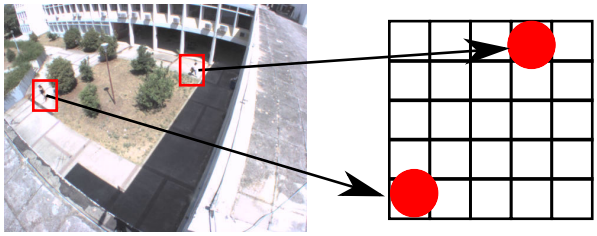


Figure 5 – *Insertion of the Blobs in the Bayesian Occupancy Grid.*

In fact this model considers that each detected bounding box represents an object on the floor (i.e. that the lower part of the bounding box corresponds to the floor). Thus by using the calibration matrix, it generates a Gaussian occupancy probability centred on this object and whose standard deviation is proportional to the bounding box width. The size (covariance matrix) of the Gaussian represents the localisation error.

2.5 Human Tracking From the BOF Output

As previously emphasized, the BOF allows to represent the surrounding environment by a map whose cells have an occupancy and velocity distribution. However, we are more interested, in this paper, in knowing the number of

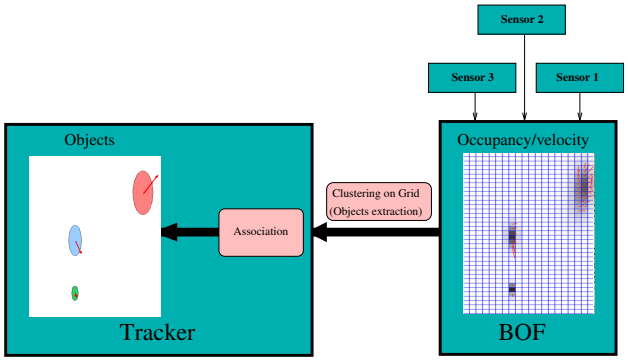


Figure 6 – *Sensing/Tracking System Architecture.*

people and their movement in the monitored area. People should thus be extracted from the BOF output and the people found need to be tracked.
For this purpose, the occupancy/velocity map is given as input to a clustering-tracking algorithm to extract objects tracked in the next step. The principle of this method is presented on Fig. 6 and a more detailed description of this clustering-tracking part could be found in [13].
It is important to note that thanks to the clustering procedure postponed at the end of the fusion process, the BOF has the advantage of not being based on an object-based representation. It allows a complexity reduction of the data association stage which could be encountered in a cluttered environment. The algorithm is highly parallelisable and can thus be used in real time to perform sensor fusion.

3 Behaviour Recognition From BOF Output

The objective of this section is to propose a filtering scheme for classifying the current motion mode of a given individual. The idea is to use the velocity information provided by the tracking system for estimating the current motion mode among three discrete hypotheses $\{Standing, Walking, Running\}$.
Our approach is based on a Bayesian filtering scheme in order to recursively update the belief about the motion mode. At each time index, the filter uses the speed observation (a Gaussian estimate) provided by the tracker for updating the belief table over the three possible hypotheses $\{Standing, Walking, Running\}$.
In order the describe the model, let's define the following variables associated to a track (a person):

- B^t: The current behaviour (at time index t), $B^t \in \{Standing, Walking, Running\}$.
- B^{t-1}: The previous behaviour (at time index $t-1$), $B^{t-1} \in \{Standing, Walking, Running\}$.
- S^t: The current speed (at time index t), $S^t \in \mathbb{R}$.
- S^t_{obs}: The observed speed (at time index t), $S^t_{obs} \in \mathbb{R}$.

The joint distribution corresponding to the proposed filter is:

$$P(B^{t-1}\ B^t\ S^t\ S_{obs}) = P(B^{t-1})P(B^t \mid B^{t-1})P(S^t \mid B^t)P(S^t_{obs} \mid S^t),$$

in which:

- $P(B^{t-1})$: The probability distribution corresponding to the estimation of the behaviour at the previous time index $t-1$.
- $P(B^t \mid B^{t-1})$: The prediction model providing the transition probabilities from a given mode to another. $P(B^t \mid B^{t-1})$ is the probability of switching to mode $P(B^t)$ (at time t) given the previous mode is B^{t-1} (at time $t-1$). This conditional distribution has been set by hand to the matrix in Tab. 1.

	Standing	Walking	Running
Standing	0.90	0.05	0.05
Walking	0.05	0.90	0.05
Running	0.05	0.05	0.90

Table 1 – *Transition matrix.*

- $P(S^t \mid B^t)$: The observation model providing the distribution over the speed given the actual mode. It's supposed to be Gaussian:

$$P(S^t \mid B^t) = \mathcal{N}(S^t; \mu(B^t), \sigma(B^t)),$$

in which the parameters have been set by hand to the matrix in Tab. 2. However, it is important to note that a learning scheme (E.M. algorithm) can be employed to set these parameters.

	Standing	Walking	Running
$\mu(B^t)$	0.0	0.5	3.0
$\sigma(B^t)$	0.1	0.2	0.5

Table 2 – $P(S^t \mid B^t)$ *parameters. The unit is meter per second.*

- $P(S^t_{obs} \mid S^t)$: The observation error model. It's assumed Gaussian:

$$P(S^t_{obs} \mid S^t) = \mathcal{N}(S^t_{obs}; S^t, \sigma^t_{obs}),$$

in which σ^t_{obs} is the standard deviation associated to the estimated speed S^t_{obs} returned by the tracker.

4 Experimental Results

The proposed fusion scheme has been applied to the Prometheus European project data set

(http://www.prometheus-fp7.eu/) and especially to the "ATM" scenario.
This scenario focuses on security around an automated teller machine. The corresponding data has been recorded in a wide outdoor area with two visible cameras (1024x768, 15fps).
All these sensors have been calibrated using the Camera Calibration Toolbox available at (http://www.vision.caltech.edu/bouguetj/calib_doc).
To detect moving blobs, a classical background subtraction algorithm is used [14]. For this purpose, a modified version of the efficient implementation provided by the OPENCV library using three adaptive Gaussian to model the background is employed. However, it has been improved to not update background components where blobs were detected at the previous step. It allows to continue to detect standing people even if they do not move for a long time.
The tracking and behaviour recognition results are shown in Figure 7 where people behaviours are displayed in red.
First, it can be seen that the tracking is very robust even when people are following each other resulting in a severe occlusion problem. The different points of view associated to the used sensors and the fusion process associated allow to reach this level of robustness.
Second, it is easy to see that the behaviours are correctly estimated because the system can efficiently discriminate standing, walking and running people thanks to the utilisation of the Bayesian filter.

5 Conclusions and Discussions

In this paper, we used the Bayesian Occupancy Filter framework to fuse the information provided by different cameras monitoring the same field of view. In the videos, pedestrians were detected thanks to a classical background subtraction algorithm whose results were given as input to the BOF. The BOF allows the computation of a grid storing both occupancy and velocity of each cell. This output was then clustered in order to obtain an object based representation of the environment allowing the pedestrian tracking and low level behaviour recognition thanks to a Bayesian Filtering process
The results show that this system can be efficiently used to resolve occlusions that often occur in a videosurveillance application. Moreover, the small computing times allow to envisage its integration in a commercial system.
However, the behaviours detected are always too simple and we are currently working on using a concurrent HMMs-based model in order to recognise high level behaviour using the motion information provided by the human tracks.

References

[1] Elfes A.. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22(6):46–57, 1989.

[2] Thrun S., Fox D., et Burgard W.. A probabilistic approach to concurrent mapping and localization for mobile robots. *Machine Learning*, 31:29–53, 1998.

[3] Fleuret F., Berclaz J., Lengagne R., et Fua P.. Multicamera people tracking with a probabilistic occupancy map. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 30(2):267–283, 2007.

[4] Ferreira J.F., Bessière P., Mekhnacha K., Lobo J., Dias J., et Laugier C.. Bayesian models for multimodal perception of 3d structure and motion. Dans *Proceedings of the International Conference on Cognitive Systems (CogSys 2008)*, April 2008.

[5] Beymer D.. Person counting using stereo. Dans *HUMO '00: Proceedings of the Workshop on Human Motion (HUMO'00)*, page 127, 2000.

[6] Coué C., Fraichard T., Bessière P., et Mazer E.. Multi-sensor data fusion using bayesian programming : an automotive application. Dans *Proceedings of the IEEE-RSJ International Conference on Intelligent Robots and Systems*, 2002.

[7] Tay C., Mekhnacha K., Chen C., Yguel M., et Laugier C.. An efficient formulation of the bayesian occupation filter for target tracking in dynamic environments. *International Journal Of Autonomous Vehicles*, 6(1/2):155–171, 2008.

[8] Ros J. et Mekhnacha K.. Multi-sensor human tracking with the bayesian occupancy filter. Dans *Proceedings of the 16th International Conference on Digital Signal Processing (DSP 2009)*, pages 1–8, Santorini, Greece, 2009.

[9] Oliver N.M., Rosario B., et Pentland A.P.. A bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):831–843, 2000.

[10] Nascimento J.C., Figueiredo M.A.T., et Marques J.S.. Segmentation and classification of human activities. Dans *Proceedings of HAREM International Workshop on Human Activity Recognition and Modelling*, 2005.

[11] Devlaeminck R.. *Human Motion Tracking With Multiple Cameras Using a Probabilistic Framework for Posture Estimation*. Thèse de doctorat, School of Electrical and Computer Engineering, Purdue University, 2006.

[12] Mekhnacha K., Mao Y., Raulo D., et Laugier C.. Bayesian occupancy filter based 'fast clustering-tracking' algorithm. Dans *Proc. of the IEEE Int. Conf. on Multisensor Fusion and Integration for Intelligent Systems*, 2008.

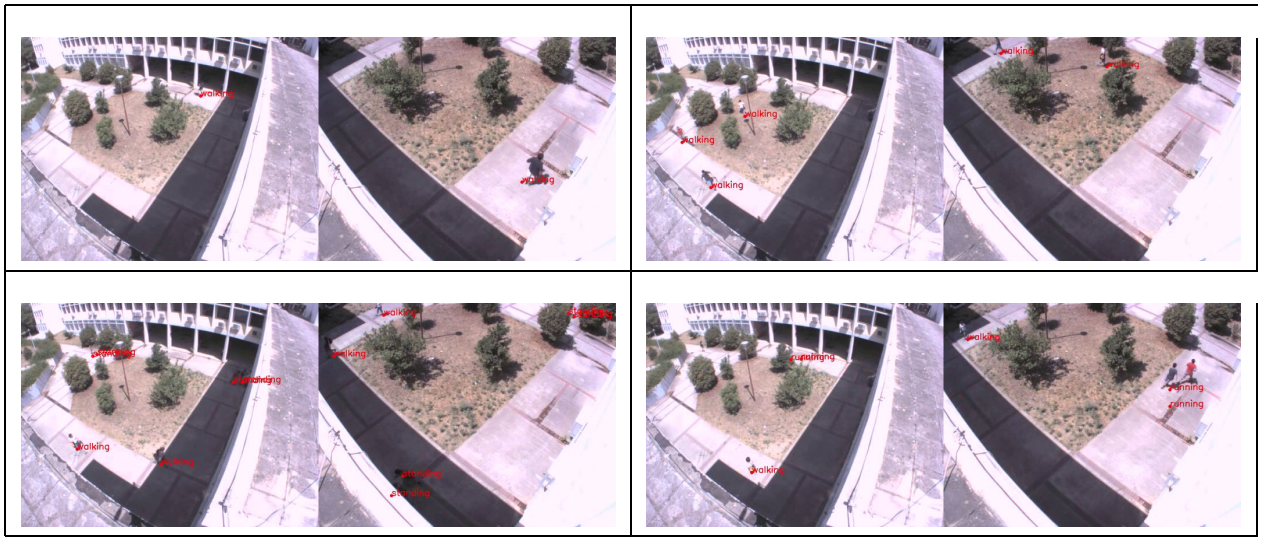


Figure 7 – *Tracking and Behaviour Recognition Results Using Two Visible Cameras.*

[13] Mekhnacha K., Mao Y., Raulo D., et Laugier C.. The 'Fast Clustering-Tracking' algorithm in the Bayesian Occupancy Filter framewor. Dans Springer Berlin Heidelberg, éditeur, *Multisensor Fusion and Integration for Intelligent Systems*, volume 35 de *Lecture Notes in Electrical Engineering*, pages 201–219. 2009.

[14] Kaewtrakulpong P. et Bowden R.. An improved adaptive background mixture model for real-time tracking with shadow detection. Dans *Proceedings of the 2nd European Workshop on Advanced Video Based Surveillance Systems (AVBS '01)*, Kingston, UK, 2001.

Une expérimentation subjective pour l'évaluation de segmentations de maillages 3D

H. Benhabiles[1] G. Lavoué[2] J-P. Vandeborre[1,3] M. Daoudi[1,3]

[1] LIFL (UMR USTL/CNRS 8022), Université de Lille, France

[2] Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France

[3] Institut TELECOM ; TELECOM Lille 1, France

{halim.benhabiles, jean-philippe.vandeborre, mohamed.daoudi}@lifl.fr
{glavoue}@liris.cnrs.fr

Résumé

Dans cet article, nous proposons une expérimentation subjective pour l'évaluation de la qualité de segmentations de maillages 3D. Dans ce but, nous avons conçu un protocole tout en respectant un certain nombre de facteurs : les conditions d'affichage, les interactions possibles, l'intervalle de notation, ainsi que le nombre d'opérateurs humains. Afin de mettre en œuvre cette expérimentation, plus de 30 opérateurs humains ont participé à la notation de 250 segmentations provenant de plusieurs algorithmes. Pour éviter l'effet du facteur de séquencement temporel, les segmentations ont été affichées aux opérateurs de manière aléatoire avec un biais pour obtenir suffisamment de notes pour chacune de ces segmentations. Le score moyen (Mean Opinion Score) est ensuite calculé pour chaque segmentation. Ce score reflète l'opinion des opérateurs vis-à-vis de la qualité de la segmentation.

Les résultats de l'expérimentation subjective sont utilisés pour évaluer la qualité des algorithmes automatiques utilisés ainsi que les métriques existantes de similarité entre segmentations.

Mots clefs

Maillage 3D, évaluation, segmentation, expérimentation subjective.

1 Introduction

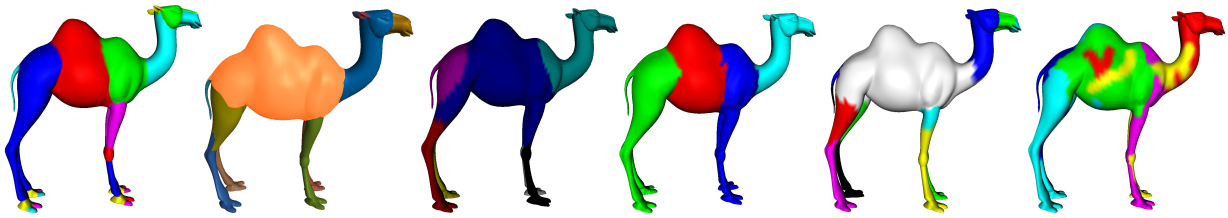
La segmentation de maillages 3D est un domaine de recherche très actif avec de nombreuses applications importantes telles que l'indexation, la compression, etc. La performance de ces applications dépend fortement de l'efficacité de l'algorithme de segmentation. L'*évaluation* de la qualité d'une segmentation de maillage 3D est donc une étape critique. Une approche naturelle pour atteindre cet objectif est d'effectuer des tests subjectifs basés sur le jugement humain quantitatif.

Dans ce contexte, l'objectif du présent travail est de mettre en œuvre une expérimentation subjective pour l'évaluation de la qualité des segmentations de maillages 3D. Pour cela, nous avons conçu un protocole tout en respectant un certain nombre de facteurs tels que l'intervalle de notation, les conditions d'affichage, etc. Ce protocole vise à standardiser l'évaluation subjective et à la rendre plus pertinente. Dans cette expérimentation, les opérateurs humains ont noté un ensemble de segmentations provenant de plusieurs algorithmes. Les résultats de cette expérimentation sont utilisés pour l'évaluation quantitative des algorithmes de segmentation automatique, ainsi que l'évaluation des métriques de similarités entre segmentations, utilisées dans les systèmes de benchmark récents [1, 2].

L'article est organisé comme suit. La section 2 fournit un court état de l'art sur les travaux existants concernant l'évaluation de la segmentation de maillage 3D. La section 3 décrit en détail notre expérimentation. La section 4 met en avant l'utilité des résultats de l'expérimentation subjective à travers l'évaluation quantitative de quatre algorithmes de segmentation récents, ainsi que l'évaluation objective des métriques de similarités proposées dans [1, 2]. La section 5 conclut l'article.

2 Etat-de-l'art sur l'évaluation de la segmentation de maillages 3D

Contrairement aux nombreuses propositions d'algorithmes adressant la segmentation de maillages 3D [7], moins d'attention a été accordée par la communauté graphique envers l'*évaluation* de la *qualité* de la segmentation produite par ces algorithmes. Deux travaux principaux ont été proposés récemment [1, 2] sur ce sujet de l'évaluation. Ils s'appuient sur un système de benchmarking incluant un corpus de vérités-terrains et un ensemble de métriques de similarité. Le corpus de vérités-terrains est composé d'un ensemble de modèles 3D faisant partie de différentes catégories (humain, animal, etc.). Chaque modèle 3D est associé à plusieurs vérités-terrains (segmentations manuelles) effectuées par des opérateurs humains. L'évaluation d'un algorithme de segmentation consiste à mesurer la simila-



(a) Vérité-terrain. (b) Shapira *et al.* [3]. (c) Tierny *et al.* [4]. (d) Attene *et al.* [5]. (e) Lavoué *et al.* [6]. (f) Aléatoire.

Figure 1 – *Segmentation du modèle* camel *en utilisant plusieurs algorithmes.*

rité, en utilisant des métriques de similarité, entre la segmentation automatique produite par cet algorithme pour un modèle donné, et ses vérités-terrains correspondantes. Plus la segmentation automatique est proche des vérités-terrains, meilleure est la qualité de l'algorithme.
Bien que ces solutions permettent une évaluation qui est à la fois objective et quantitative grâce aux vérités-terrains et aux métriques de similarités, le moyen idéal pour évaluer les algorithmes de segmentation reste une expérimentation subjective explicite, où des observateurs notent directement les segmentations résultantes. De plus, une telle expérimentation subjective permettra de quantifier l'efficacité des benchmarks existants ainsi que d'évaluer les métriques de similarité introduites.

3 L'expérimentation subjective

3.1 Corpus des segmentations

La conception du stimulus est une étape clé dans le protocole subjectif. Dans notre cas, nous avons besoin de sélectionner un ensemble de modèles 3D qui seront segmentés par plusieurs algorithmes puis notés par des opérateurs humains. A cette fin, nous utilisons notre corpus [1] de modèles 3D qui est disponible en-ligne [1], et est dédié à l'évaluation de la segmentation. La taille du corpus est raisonnable (28 modèles 3D), et son contenu est représentatif puisqu'il comprend différentes catégories communes de modèles 3D.
Dans notre expérimentation, nous avons demandé aux opérateurs humains de noter les segmentations de ces objets provenant de plusieurs algorithmes automatiques. Nous avons crée un ensemble de 250 segmentations basé sur les 28 modèles 3D du corpus. Pour cette tâche, nous avons pris en considération 4 algorithmes de segmentation automatique : Attene *et al.* [5], Lavoué *et al.* [6], Shapira *et al.* [3], et Tierny *et al.* [4]. Mis à part l'algorithme de Lavoué *et al.* [6], les autres sont hiérarchiques ; nous avons donc généré, pour chacun d'eux, deux niveaux de segmentations : grossière et fine, pour chaque modèle, ce qui donne au total 28×2 segmentations par algorithme hiérarchique et 28 segmentation pour l'algorithme de Lavoué *et al.* [6]. A ces 28×7 segmentations, nous avons ajouté 28 segmentations vérités-terrains provenant de notre corpus ainsi que 28 segmentations aléatoires générées par un algorithme simple basé sur un mécanisme de croissance de région aléatoire. La figure 1 illustre différentes segmentations du modèle *camel*. Le code source et/ou le binaire de chaque algorithme de segmentation automatique sont fournis par leurs auteurs. Ainsi, nous avons obtenu un corpus de 250 segmentations à noter.

1. `http://www-rech.telecom-lille1.eu/3dsegbenchmark/`

3.2 Protocole subjectif

Le protocole que nous proposons est inspiré de ceux qui existent déjà pour l'évaluation de la qualité de segmentation de vidéo [8], l'évaluation de la qualité de tatouage 3D [9], et l'évaluation de la qualité d'image [10]. Ces protocoles sont tous basés sur le *Single Stimulus Continuous Quality Scale* (SSCQS) qui est une technique standard de notation utilisée pour évaluer la qualité de vidéo et du contenu multimédia. Notre protocole comprend les étapes suivantes :

- ***Instructions orales.*** Nous expliquons à nos volontaires la tâche qu'il doivent compléter, et nous les familiarisons avec l'opération de notation, les modèles 3D, ainsi que les interactions possibles.
- ***Apprentissage.*** Nous montrons quelques segmentations vérités-terrains et aléatoires de différents modèles à l'utilisateur afin qu'il puisse comprendre le concept de bonne et de mauvaise segmentation, et établir un ordre de référence. Le but n'est pas d'apprendre à l'utilisateur les vérités-terrains de chaque modèle, mais plutôt de lui apprendre à distinguer entre les différentes segmentations pour qu'il soit en mesure de noter la qualité d'une segmentation donnée indépendamment des vérités-terrains.
- ***Tests expérimentaux.*** Pour chaque segmentation du corpus, nous avons demandé au volontaire de donner un score entre 1 et 10, indiquant sa qualité d'un point de vue sémantique, 10 pour une segmentation parfaite, et 1 pour une segmentation très mauvaise. Cet intervalle permet aux volontaires de distinguer plus facilement entre la qualité des segmentations.

Durant les tests expérimentaux, les segmentations sont affichées une par une au volontaire sur un écran LCD 22-pouces, sans les vérités-terrains. Pour éviter l'effet du facteur de séquencement temporel, la séquence de segmentations affichée à chaque participant est générée

aléatoirement. Les interactions de base sont autorisées (zoom, rotation, translation). Il est évident que la notation de 250 segmentations par un volontaire représente une tâche fastidieuse, ce qui nous a incité à restreindre ce nombre à 50 segmentations par volontaire de manière à obtenir suffisamment de notes pour les 250 segmentations. La figure 2 illustre l'interface développée pour l'opération de notation.

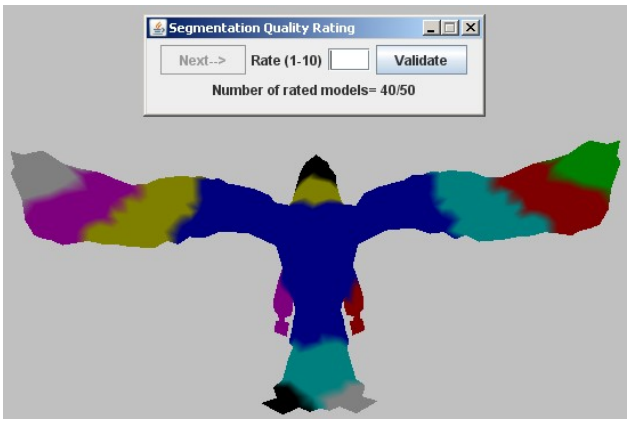


Figure 2 – *Interface utilisateur pour la notation des segmentations.*

Le *Mean Opinion Score* (MOS) est ensuite calculé pour chaque segmentation du corpus :

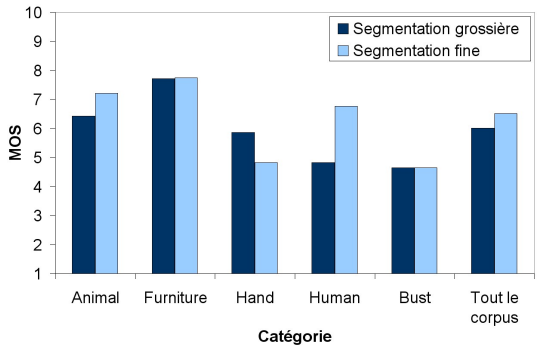
$$MOS_i = \frac{1}{n}\sum_{j=1}^{n} m_{ij} \quad (1)$$

MOS_i est le *mean opinion score* de la i^{eme} segmentation, n est le nombre de sujets, et m_{ij} est le score ($\in [1, 10]$) affecté par le j^{eme} sujet à la i^{eme} segmentation. Cette expérimentation subjective à été menée sur 35 personnes (étudiants et personnels) de l'Université de Lille 1, offrant un total de 7 scores par segmentation.

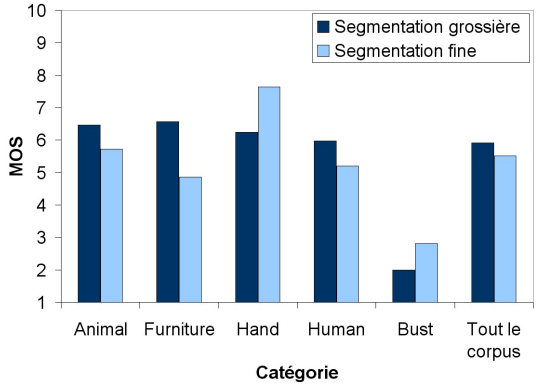
4 Résultats et analyse des données

4.1 Influence du raffinement sur la qualité de la segmentation

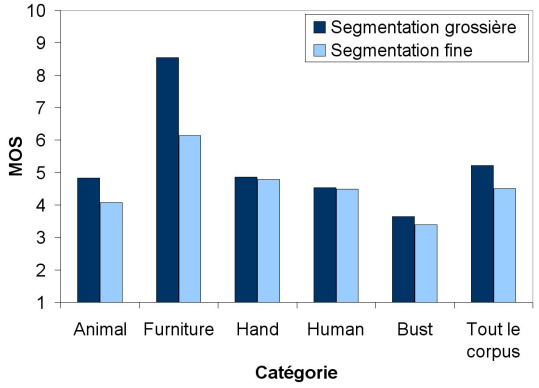
Certains algorithmes automatiques sont hiérarchiques, c'est-à-dire qu'ils sont capables de produire des segmentations avec différents niveaux de raffinement. Une expérience intéressante est d'étudier l'influence du niveau de granularité sur la qualité perçue par les observateurs. Dans ce but, nous avons calculé la moyenne du MOS des modèles de chaque catégorie, pour chaque algorithme, et pour les deux niveaux de segmentation (grossière et fine), ensuite nous avons comparé les résultats des deux niveaux. La figure 3 illustre les résultats obtenus pour les trois algorithmes hiérarchiques. On peut remarquer que les moyennes des deux niveaux de segmentation, pour une catégorie donnée ou bien pour tout le corpus, sont proches l'une de l'autre. Plus précisément, la variation moyenne entre les deux niveaux pour tout le corpus, et pour chacun des algorithmes : Shapira *et al.* [3], Tierny *et al.* [4], et Attene *et al.* [5], est respectivement de 7%, 10%, et 11%. Cela veut dire que les segmentations restent sémantiquement consistantes quelque soit leur niveau de raffinement.



(a) Shapira *et al.* [3].



(b) Tierny *et al.* [4].



(c) Attene *et al.* [5].

Figure 3 – *Moyennes des MOS des segmentations obtenues par des algorithmes hiérarchiques.*

4.2 Comparaison de la performance des algorithmes de segmentation

Le tableau 1 présente le classement, basé sur le MOS, de chaque algorithme (segmentation fine pour les algorithmes

Tableau 1 – *Classement des algorithmes accompagné des moyennes MOS pour chaque catégorie du corpus.*

	Vérités-terrains	**Shapira *et al.* [3]**	**Tierny *et al.* [4]**	**Attene *et al.* [5]**	**Lavoué *et al.* [6]**	**Aléatoire**
animal	1 / 8.26	2 / 7.20	3 / 5.72	5 / 4.83	4 / 5.01	6 / 2.37
bust	1 / 8.03	2 / 4.64	4 / 2.81	3 / 3.64	5 / 2.64	6 / 1.78
furniture	1 / 9.25	3 / 7.74	5 / 3.35	2 / 8.53	4 / 6.21	6 / 1.99
hand	1 / 8.68	5 / 4.82	2 / 7.64	4 / 4.85	3 / 5.53	6 / 1.60
human	1 / 7.77	2 / 6.77	3 / 5.20	5 / 4.54	4 / 4.62	6 / 2.28
tout	**1 / 8.36**	**2 / 6.51**	**3 / 5.27**	**4 / 5.21**	**5 / 4.92**	**6 / 2.10**

hiérarchiques) pour chaque catégorie de modèles ainsi que pour tout le corpus, incluant les segmentations aléatoires et les segmentations vérités-terrains. Les moyennes du MOS sont aussi affichées. Comme prévu, les vérités-terrains ont le meilleur classement pour chaque catégorie et pour tout le corpus, alors que les segmentations aléatoires sont les dernières en classement. Ceci valide la pertinence de notre corpus de vérités-terrains. Le tableau montre qu'il n y a aucun algorithme automatique qui atteint les meilleurs scores pour toute les catégories. Il montre aussi que la classe *bust* est la plus difficile à segmenter par les algorithmes automatiques, puisque la moyenne de son MOS est la plus faible en la comparant avec les autres classes. Ceci peut être dû à la complexité géométrique des modèles de cette classe, mais la raison principale est probablement le fait que ces modèles représentent des visages humains. Ce dernier type de modèles est connu dans les expérimentations subjectives comme étant un facteur de haut niveau qui attire l'attention humaine. En effet, certaines caractéristiques qui ne sont pas significatives d'un point de vue géométrique, peuvent être considérées manifestement significatives par les observateurs humains. Globalement, l'algorithme de Shapira *et al.* [3] semble être le meilleur après les vérités-terrains.

4.3 Evaluation des métriques de similarité

Une autre expérience intéressante est d'évaluer la qualité des métriques de similarité utilisées dans les systèmes de benchmark [1, 2]. Pour cela, nous utilisons le corpus présenté dans [1] qui est basé sur les mêmes 28 modèles que ceux utilisés dans l'expérimentation subjective, et comprend 4 vérités-terrains pour chaque modèle. Nous calculons la similarité entre les 250 segmentations et leurs vérités-terrains correspondantes à l'aide des métriques suivantes : Ecart de Frontières (EF), Erreur de Consistance Locale (ECL), Distance de Hamming (DH), et Indice de Rand (IR). Ensuite nous calculons la corrélation (Spearman rank correlation [11]) entre les 250 MOS et les 250 valeurs calculées par chacune des métriques. Si les métriques sont pertinentes, leurs valeurs devraient être corrélées avec les MOS données par les utilisateurs. Les résultats sont affichés dans le tableau 2. On peut distinguer dans ce tableau, 3 classes de corrélation : corrélation élevée (plus de 80%) pour la métrique IR, corrélation moyenne (entre 50% et 60%) pour les métriques ECL et DH, et corrélation faible (moins de 30%) pour la métrique EF. Le faible taux de corrélation de cette dernière métrique indique qu'elle échoue à clairement différencier entre une segmentation proche des vérités-terrains, et une mauvaise segmentation. Ainsi, un benchmark basé sur cette métrique donnera forcement des résultats qui ne sont pas pertinents. La métrique IR est la métrique qui donne les meilleurs résultats ; c'est donc celle qui devrait être utilisée en priorité dans les benchmarks existants [1, 2] puisqu'elle donne le meilleur taux de corrélation que se soit pour une catégorie donnée ou bien pour tout le corpus. Cette forte corrélation de 82% valide non seulement cette métrique mais atteste également de la qualité globale du benchmark présenté dans [1] (métrique et vérités-terrains).

Tableau 2 – *Corrélation de Spearman (%) entre les MOS et les valeurs des différentes métriques.*

	EF	**ECL**	**DH**	**IR**
animal	19.9	42	49.3	78.3
bust	8.9	73.5	68.1	81.6
furniture	21.6	52.9	69.2	85.3
hand	55.4	77.1	72.4	82.8
human	11.5	63.6	64.3	76.7
tout	**28.2**	**55.3**	**60.2**	**82.1**

5 Conclusion

Dans cet article, une expérimentation subjective pour la notation de segmentations de maillages 3D a été proposée. Dans ce but, un protocole a été soigneusement défini afin que les résultats obtenus soient pertinents. Ces résultats se sont avérés très utiles puisqu'ils nous ont permis d'effectuer une évaluation quantitative de la qualité des segmentations d'algorithmes automatiques récents ainsi que d'évaluer les métriques de similarité utilisées dans les systèmes de benchmark actuels. Dans le futur, nous visons à utiliser ces résultats avec notre corpus de vérités-terrains pour proposer un nouvel algorithme de segmentation.

Remerciements

Ce travail a bénéficié d'une aide de l'ANR (Agence Nationale de la Recherche) à travers le projet MADRAS (ANR-07-MDCO-015).

Références

[1] H. Benhabiles, J-P. Vandeborre, G. Lavoué, et M. Daoudi. A framework for the objective evaluation of segmentation algorithms using a ground-truth of human segmented 3d-models. Dans *IEEE International Conference On Shape Modeling And Application (SMI)*, 2009.

[2] X. Chen, A. Golovinskiy, et T. Funkhouser. A benchmark for 3d mesh segmentation. *ACM Transactions on Graphics (SIGGRAPH)*, 28(3), 2009.

[3] Lior Shapira, Ariel Shamir, et Daniel Cohen-Or. Consistent mesh partitioning and skeletonisation using the shape diameter function. *Vis. Comput.*, 24(4) :249–259, 2008.

[4] Julien Tierny, Jean-Philippe Vandeborre, et Mohamed Daoudi. Topology driven 3D mesh hierarchical segmentation. Dans *IEEE International Conference On Shape Modeling And Application (SMI)*, 2007.

[5] Marco Attene, Bianca Falcidieno, et Michela Spagnuolo. Hierarchical mesh segmentation based on fitting primitives. *Vis. Comput.*, 22(3) :181–193, 2006.

[6] G. Lavoué, F. Dupont, et A. Baskurt. A new cad mesh segmentation method, based on curvature tensor analysis. *Computer Aided Design*, 37(10) :975–987, 2005.

[7] A. Shamir. A survey on mesh segmentation techniques. *Computer Graphics Forum*, 27(6) :1539–1556, 2008.

[8] Elisa Drelie Gelasca, Touradj Ebrahimi, Mustafa Karaman, et Thomas Sikora. A framework for evaluating video object segmentation algorithms. Dans *CVPRW '06 : Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 198. IEEE Computer Society, 2006.

[9] Massimiliano Corsini, Elisa Drelie Gelasca, Touradj Ebrahimi, et Mauro Barni. Watermarked 3d mesh quality assessment. *IEEE Transaction on Multimedia*, 9(2) :247–256, February 2007.

[10] Bernice E. Rogowitz et Holly E. Rushmeier. Are image quality metrics adequate to evaluate the quality of geometric objects. Dans *in Human Vision and Electronic Imaging*, pages 340–348, 2001.

[11] W. W. Daniel. *A Foundation For Analysis In The Health Sciences Books*. 7th edition.John Wiley and sons., 1999.

Optimisation du rapport débit-distorsion de la compression progressive de maillages par adaptation de quantification

H. Lee[1] G. Lavoué[2] F. Dupont[1]

Université de Lyon, CNRS

[1]Université Lyon 1, LIRIS, UMR5205, F-69622, France
[2]INSA-Lyon, LIRIS, UMR5205, F-69621, France

{hlee, glavoue, fdupont}@liris.cnrs.fr

Résumé

Cet article présente une méthode de compression progressive de maillages triangulaires, centrée sur l'optimisation du rapport débit-distorsion. La précision de la quantification est adaptée à chaque maillage intermédiaire pour optimiser le compromis débit-distorsion. Cette adaptation peut être déterminée de manière optimale en calculant la mesure directe de la distance géométrique entre les maillages de différentes résolutions ; elle peut être également déterminée de manière quasi-optimale et rapide en utilisant une estimation de la quantification nécessaire pour chaque maillage intermédiaire grâce à un apprentissage. Les résultats montrent que ces deux méthodes d'adaptation de quantification produisent de meilleurs résultats que les algorithmes classiques en termes de débit-distorsion.

Mots clefs

Maillages 3D, compression progressive, optimisation débit-distorsion, quantification adaptative.

1 Introduction

Actuellement, l'utilisation des modèles géométriques tridimensionnels est très répandue dans de nombreuses applications telles que la conception assistée par ordinateur, la visualisation scientifique, la réalité virtuelle, l'imagerie médicale et les jeux-vidéo. Grâce aux avancées des techniques d'acquisition et de traitement, la complexité de ces modèles géométriques a augmenté afin de modéliser un objet ou une scène avec plus de réalisme ou un résultat d'analyse scientifique avec plus de précision. Ces modèles sont souvent représentés par des maillages triangulaires : la *géométrie* et la *connectivité* sont les deux composantes principales du maillage ; la première décrit les coordonnées euclidiennes des sommets et la dernière représente comment les sommets sont reliés entre eux pour former des triangles. En dépit de l'expansion des réseaux à haut-débit, l'augmentation de la taille des modèles s'accompagne d'une nécessité de techniques efficaces de compression afin de réduire le temps de transmission lors de l'échange de tels objets à travers les réseaux. Dans ce contexte, des techniques de compression progressive s'avèrent particulièrement efficaces puisqu'elles offrent la possibilité de visualiser progressivement l'objet sous différents niveaux de détails, de grossier vers fin, suivant la quantité de données transmises. De plus l'utilisateur peut arrêter la transmission à tout moment si la précision du modèle courant est suffisante pour l'application envisagée.

L'enjeu principal de la compression progressive consiste à reconstruire le modèle aussi fidèlement que possible à l'original pour une même quantité d'information reçue. Autrement dit, on cherche à optimiser le compromis débit-distorsion. Un des facteurs importants du débit-distorsion vient de la quantification de la géométrie du maillage. Nous observons dans la littérature que le rapport entre la précision de la quantification et la complexité des maillages intermédiaires n'est pas souvent optimal. Ainsi nous proposons dans ce travail une méthode d'adaptation de la précision de la quantification des maillages intermédiaires tout au long de la transmission afin d'optimiser le rapport débit-distorsion.

Dans le paragraphe suivant, nous présentons des travaux récents sur la compression progressive de maillage triangulaire. Ensuite nous proposons notre méthode d'amélioration du compromis débit-distorsion, et enfin nous montrons les résultats obtenus suivis de la conclusion.

2 Etat de l'art

La première approche de compression progressive est introduite par Hoppe [1]. Cette nouvelle représentation de maillage dite *maillages progressifs* consiste à appliquer itérativement des contractions d'arêtes en supprimant un sommet et deux faces incidentes à l'arête. La reconstruction, lors du décodage, est réalisée par l'opération inverse, séparation de sommets. A chaque étape, une arête à contracter est déterminée en utilisant la fonction d'énergie liée à la géométrie afin d'obtenir une meilleure approximation. Malgré son côté novateur, cet algorithme n'est pas très efficace à cause du coût de la localisation du sommet à séparer au moment du décodage. Cette méthode est étendue par

plusieurs chercheurs afin d'améliorer le taux de compression et aussi l'aspect débit-distorsion [2, 3, 4].

Cohen-Or et al. [5] présente une méthode basée sur la coloration de patch pour la transmission progressive. Leur algorithme supprime successivement un ensemble de sommets indépendants (les sommets de cet ensemble ne sont par relié entre eux par une arête). Ensuite, le trou généré est retriangulé de façon déterministe. Les triangles rebouchant un trou forment un patch. Ces patchs sont colorés en utilisant 2 ou 4 couleurs afin de permettre au décodeur de déterminer correctement chaque patch afin d'insérer proprement des sommets lors de la reconstruction. Cet algorithme compresse la connectivité avec un coût de 6 bits/sommet. Alliez et Desbrun [6] proposent une extension des algorithmes de compression mono-résolution [7, 8] qui exploitent la distribution des valences des sommets pour une compression compacte de la connectivité. Ils appliquent itérativement une conquête de décimation et une conquête de nettoyage en paire afin de générer différents niveaux de détails. La conservation de la régularité garantit le codage efficace de la connectivité et le coût de codage de la connectivité est en moyenne de 3,7 bits/sommet.

Notons que la géométrie est codée de la même manière dans toutes les approches décrites jusqu'ici : quantification suivi de prédiction. Lors de l'introduction d'un nouveau sommet par séparation de sommet ou par insertion de sommet, sa position est prédite en utilisant les informations de ses voisins et le résidu est codé. Pourtant, l'optimalité du codage de la géométrie est souvent limitée par le fait qu'il est guidé par le codage de la connectivité. Les approches citées ci-dessus sont dites guidées par la connectivité.

Gandoin et Devillers [9] proposent un premier algorithme qui est cette fois-ci dirigé par la géométrie, basé sur la subdivision en kd-tree. Pour le codage de la géométrie, l'espace est subdivisée récursivement en deux cellules jusqu'à ce que chaque cellule possède un ou zéro sommet et le nombre de sommets d'une de deux cellules est codé. La reconstruction de la connectivité est effectuée en codant le changement apparu lors de la subdivision en utilisant des opérations de séparations de sommets. En termes de taux de compression, cet algorithme présente des meilleurs résultats que les algorithmes guidés par la connectivité. Peng et Kuo [10] présentent un algorithme dirigé par la géométrie, basé sur la subdivision en octree. A chaque itération, une cellule est divisée en huit cellules-filles et au lieu du nombre de sommets dans chaque cellule-fille, la présence de sommet est codée. En optimisant le codage de la géométrie et de la connectivité grâce aux prédictions efficaces, cette approche donne un excellent résultat en termes de taux de compression. Cependant, les méthodes basées sur géométrie sont en général moins performantes que celles basées sur connectivité pour le compromis débit-distorsion.

Récemment, Valette et al. [11] proposent une nouvelle approche basée sur un schéma de raffinement. Contrairement aux méthodes précédentes, leur approche part d'un maillage grossier quelconque et une série de divisions d'arête est effectuée pour générer des maillages intermédiaires uniformes. En ne restituant la connectivité originale qu'à la fin de l'algorithme et en transmettant progressivement les coordonnées quantifiées des sommets, cette approche montre son efficacité en termes de débit-distorsion, particulièrement à bas débit.

L'allocation de bits est une technique qui permet d'optimiser le compromis débit-distorsion. Généralement, pour un débit imposé elle permet de minimiser la distorsion liée à la perte d'information géométrique. King et Rossignac [12] introduisent une approche qui détermine le nombre de sommets et le nombre de bit de quantification des coordonnées de sommets qui optimisent la distorsion pour un débit donné, en utilisant une mesure de complexité de la forme géométrique. Pour la compression progressive par ondelettes, Payan et Antonini [13] proposent une méthode d'allocation de bits qui optimise la quantification des coefficients d'ondelette des différentes sous-bandes de fréquence.

3 Optimisation du compromis débit-distorsion

Le rapport entre la précision de la quantification et la complexité des maillages intermédiaires est un facteur important pour la performance débit-distorsion [12]. L'optimisation de ce rapport est pourtant négligée dans la plupart des travaux sur la compression progressive. En effet, la précision de la quantification est grossière par rapport au nombre d'éléments de maillage pour les algorithmes dirigés par la géométrie. Au contraire, la précision des positions est plus grande que nécessaire pour des maillages intermédiaires de basse résolution pour les algorithmes guidés par la connectivité qui conservent la précision fine de la quantification initiale tout au long de simplification. Dans ce paragraphe, nous proposons une méthode qui permet d'adapter la précision de la quantification à la complexité des maillages intermédiaires en optimisant ainsi le compromis débit-distorsion.

3.1 Algorithme de base

En général, les algorithmes dirigés par la connectivité produisent des maillages intermédiaires de meilleure qualité que les algorithmes guidés par la géométrie. Pour cette raison, notre méthode est basée sur l'algorithme d'Alliez et Desbrun [6] qui est l'un des plus performants. Pour le codage de la connectivité, cet algorithme utilise la bonne propriété statistique de la distribution des valences de sommets. Un ensemble de sommets indépendants est décimé itérativement en combinant une conquête de décimation et une conquête de nettoyage qui génère différents niveaux de résolution. La conquête de décimation consiste à parcourir le maillage patch par patch en enlevant des sommets de valence inférieure ou égale à 6. Grâce à la re-triangulation déterministe et la conquête de nettoyage, la dispersion de la valence des sommets est minimisée. Pour la reconstruction de la connectivité, la valence des sommets enlevés est

codée.
Pour la géométrie, les auteurs appliquent d'abord une quantification initiale utilisant entre 8 et 12 bits suivant la complexité du maillage d'entrée. Pour coder la position d'un sommet enlevé, ils utilisent un repère local (repère de Frenet) obtenu grâce à l'estimation de la normale à la surface afin de séparer les composantes tangentielles et normales pour améliorer le taux de compression. Dans ce repère local, la différence entre la position du sommet et celle du barycentre de sommets voisins est codée.
Récemment, Lee et al. [14] proposent une amélioration de ce codeur géométrique basé sur une bijection. Ce codeur géométrique est utilisé pour la suite du papier.

3.2 Quantification adaptative

Dans la figure 1, la version initiale composée de 11362 sommets et la version simplifiée composée de 111 sommets de Venusbody sont quantifiées en utilisant respectivement 10 bits et 4 bits. Pour la version initiale (1.a et 1.b), la quantification sur 4 bits introduit une dégradation beaucoup plus importante au niveau de la qualité visuelle et de la distance géométrique que la quantification sur 10 bits. Au contraire, la quantification sur 10 bits et 4 bits de la version simplifiée (1.c et 1.d) génère une distorsion géométrique très similaire. Cette observation montre que chaque maillage intermédiaire peut être quantifié avec un nombre de bits plus adapté en fonction du nombre d'éléments, sans augmenter significativement la distorsion. Comme la quantification influence également la taille du fichier compressé, l'adaptation de la quantification à la complexité des maillages intermédiaires peut donc améliorer le compromis débit-distorsion.

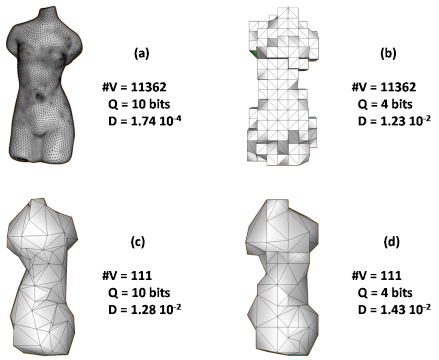


Figure 1 – *Comparaison de distorsion de la version initiale et de la version simplifiée de Venusbody en utilisant différentes précisions de quantification.*

La figure 2 décrit notre algorithme comparé aux algorithmes classiques dirigés par la connectivité. Traditionnellement, l'algorithme de compression simplifie le maillage initial, M_n^Q, dont les sommets sont quantifiés en utilisant Q bits. Après n itérations, le maillage de base M_0^Q est obtenu. Notre algorithme diminue aussi la précision de quantification au cours de l'encodage. Ainsi, une amélioration de la performance débit-distorsion est possible en trouvant un chemin optimal par combinaison d'une série de décimations et d'une série de diminutions de la précision de quantification.

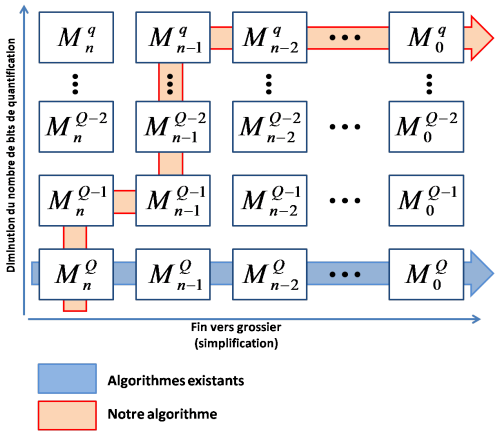


Figure 2 – *Notre algorithme (flèche rouge) et les algorithmes classiques (flèche bleue).*

Deux difficultés principales se présentent pour l'adaptation de la quantification :
– La détermination automatique de l'étape suivante à chaque itération qui conduit au meilleur compromis débit-distorsion.
– La diminution de la précision de la quantification et le codage efficace de l'opération inverse.

Notons que toutes ces opérations sont appliquées pendant l'encodage.

3.3 Diminution de la précision de la quantification

La quantification initiale utilisant b bits consiste à diviser la boîte englobante du maillage d'entrée en $2^b * 2^b * 2^b$ cubes puis à déplacer chaque sommet à l'intérieur d'un cube au centre. Si b est réduit à $b - 1$, la dimension de cube doit s'agrandir deux fois plus le long des trois axes, et chaque sommet doit être déplacé vers le centre du nouveau cube. Ainsi, la diminution de la précision de quantification peut être considérée comme une structure d'octree ; les cellules-filles (cubes initiaux) sont fusionnées en cellules-mères (nouveaux cubes).Pour l'opération inverse, il suffit de coder l'indice de cellule-fille d'origine de chaque sommet. Sans aucune méthode de prédiction, chaque déplacement de sommet coûte 3 bits car un indice parmi huit est à coder (huit cellules-filles sont fusionnées en une cellule-mère). Pour réduire ce coût, nous adoptons la méthode de prédiction de l'algorithme de Peng et Kuo [10]. Pour chaque cellule-fille, une valeur de priorité est calculée en utilisant la position des sommets voisins par rapport au sommet en cours de traitement. Ensuite, les indices de cellules-filles sont réordonnés en utilisant ces valeurs de priorité et le nouvel indice correspondant à la cellule-fille initiale est codé. Cette opération de diminution de quantification est illustrée en 2D dans la figure 3.

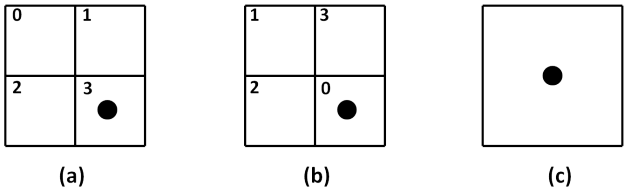


Figure 3 – *Les indices initiaux (a) des cellules-filles sont réordonnées (b). Ensuite, le sommet est déplacé vers le centre de cellule-mère (c) et l'indice 0 qui correspond à l'indice de la cellule-fille d'origine est codé.*

3.4 Détermination optimale de la quantification

Comme nous le montre la figure 2, un maillage intermédiaire, M_i^j, peut subir une décimation qui conduit à M_{i-1}^j, ou une diminution de la quantification qui conduit à M_i^{j-1}, durant l'encodage. Entre ces deux possibilités, nous choisissons celle qui améliore le plus la performance débit-distorsion. Pour effectuer de manière automatique et optimale cette sélection, nous calculons d'abord la différence de l'erreur géométrique ΔD par rapport à l'erreur de M_i^j et les bits nécessaires pour le codage de l'opération inverse, ΔB pour les deux cas. Pour le maillage décimé M_{i-1}^j, nous calculons ΔB_{Deci} en évaluant l'entropie des informations de connectivité et de géométrie. L'erreur géométrique ΔD_{Deci} est obtenue en mesurant la distance entre le maillage décimé et le maillage original. Similairement, ΔB_{quan} et ΔD_{quan} sont obtenus pour l'étape de diminution de quantification.
La prochaine opération qui optimise localement le compromis débit-distorsion est déterminée en comparant la pente $\Delta D_{Deci}/\Delta B_{Deci}$ et la pente $\Delta D_{quan}/\Delta B_{quan}$. La figure 4 décrit cette sélection automatique. Dans cet exemple, la décimation est choisie comme la prochaine étape car sa valeur de pente est inférieure à celle de la diminution de quantification.

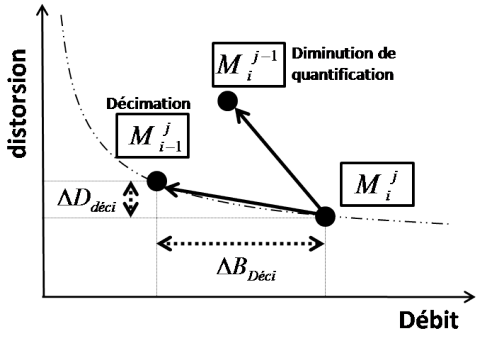


Figure 4 – *Choix de la meilleure opération à effectuer : décimation ou diminution de quantification.*

Notons que notre méthode de détermination optimale de la quantification optimise localement la performance débit-distorsion.

3.5 Détermination quasi-optimale de la quantification basée sur une méthode d'estimation

Notre méthode optimale (section 3.4) permet d'améliorer le compromis débit-distorsion comme nous le montre la figure 5. Pour un débit donné, notre algorithme peut trouver un meilleur compromis entre le nombre de sommets et la précision de la quantification en améliorant la qualité géométrique.
Cependant, cette approche possède un inconvénient qui empêche son application effective : le temps de calcul. En effet, à chaque itération notre approche a besoin de calculer des erreurs de distorsion du modèle issu de la décimation et du modèle issu de la diminution de la précision de quantification en utilisant un outil de mesure de distance géométrique entre deux maillages [15]. Pour le modèle Rabbit qui possède 67039 sommets, le temps de l'encodage est de 660 secondes alors que 2 secondes suffisent sans adaptation de quantification. Afin de réduire le temps de calcul lors de l'encodage, nous proposons de déterminer le meilleur chemin (la figure 2) basé sur une analyse de la complexité du modèle et sur un apprentissage afin de trouver une quantification adaptée sans calcul de distance.

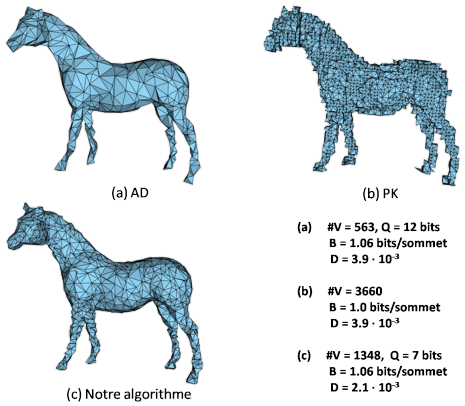


Figure 5 – *Résultats du modèle Horse obtenus par différentes approches à des débits similaires : Alliez et Desbrun [6], Peng et Kuo [10] et notre algorithme.*

Bien que la décision de l'opération suivante (décimation ou diminution de quantification) soit prise localement à chaque itération, nous observons que notre approche basée sur la détermination optimale adapte globalement la précision de quantification à la complexité des maillages intermédiaires.
En appliquant notre approche de la section 3.4 avec des différentes valeurs de quantification initiale, nous constatons qu'il existe un chemin unique qui ne dépend que du modèle d'entrée. Dans la figure 6, le chemin coloré en bleu est obtenu en utilisant le modèle Bimba avec une quantification initiale de 9 bits. Si une quantification est appliquée en utilisant plus de 9 bits, une série de diminutions de quantification est effectuée (flèche rouge de la figure 6), ensuite le

chemin restant est identique au chemin bleu. Au contraire, si une quantification utilisant moins de 9 bits est appliquée, une série de décimations est réalisée (flèche rouge de la figure 6) afin d'arriver au même point que le chemin bleu ; le chemin restant est identique. Pour le modèle Bimba, le nombre de bits de la quantification adaptée au niveau de résolution n est 9, il est de 8 pour le niveau $n-1$, etc. Cette observation prouve la possibilité de déterminer la quantification adaptée à chaque modèle en analysant les caractéristiques géométriques.

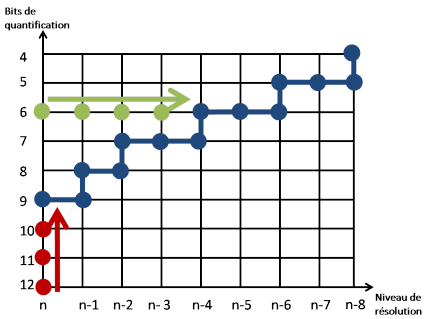


Figure 6 – *Le chemin parcouru en utilisant différentes quantifications sur le modèle Bimba.*

Nous avons choisi d'utiliser le rapport entre le volume de la boite englobante et l'aire de la surface du modèle qui permet de mesurer convenablement la compacité et la densité d'échantillonnage nécessaire du modèle.

$$C = \frac{volume\ de\ la\ boite\ englobante}{aire\ de\ surface\ \otimes\ nombre\ de\ sommets} \quad (1)$$

C est un estimateur adéquat pour déterminer automatiquement une quantification adaptée pour notre approche. Dans cette formule, l'aire de la surface est la somme des aires des triangles.
Pour 7 objets, nous avons déterminé la précision optimale de quantification pour les maillages intermédiaires. La figure 7 présente ces valeurs de quantification en fonction de C (105 exemples au total). Ces 105 exemples nous servent de base d'apprentissage afin d'établir une relation entre la complexité du maillage et la précision de la quantification nécessaire.
A partir de cette base d'apprentissage, nous utilisons la loi logarithmique qui génère une courbe de tendance mieux adaptée aux points de la figure 7 que des autres lois d'approximation. Les paramètres de cette courbe sont calculés par une méthode des moindres carrés afin d'estimer la valeur de quantification optimale :

$$Q = a * log(C) + b \quad (2)$$

avec $a = -1.248$ et $b = -0.954$.
Ainsi à chaque itération, nous calculons la valeur de Q en utilisant l'équation (2). Si la quantification actuelle est supérieure à Q, une diminution de précision de quantification est appliquée, sinon, la décimation est effectuée. Les résultats (voir la section suivante) montrent que cette approximation permet de déterminer convenablement la quantification pour tous les niveaux de détails des différents modèles.

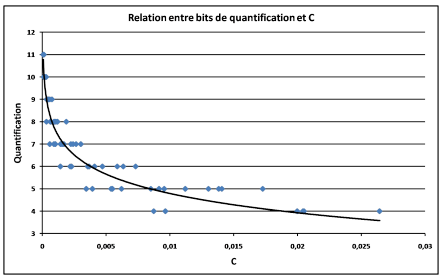


Figure 7 – *La précision optimale de quantification en fonction de C pour les différents maillages intermédiaires de notre base d'apprentissage.*

4 Résultats

Les figures 8 et 9 montrent respectivement les résultats de débit-distorsion de Venusbody et Venushead, qui sont quantifiés en utilisant 10 bits. Dans ces figures, on compare notre algorithme optimal (Section 3.4) et notre algorithme quasi-optimal (Section 3.5) avec celui d'Alliez et Desbrun [6]. Notre algorithme optimal améliore globalement le débit-distorsion par rapport à [6] et notre algorithme quasi-optimal donne un résultat très similaire. Notons que les modèles Venusbody et Venushead ne sont pas utilisés dans la base d'apprentissage ce qui vérifie la validité de notre méthode quasi-optimale. Dans ces figures, l'ordonnée indique la valeur de la distorsion qui est le maximum de distances RMS.

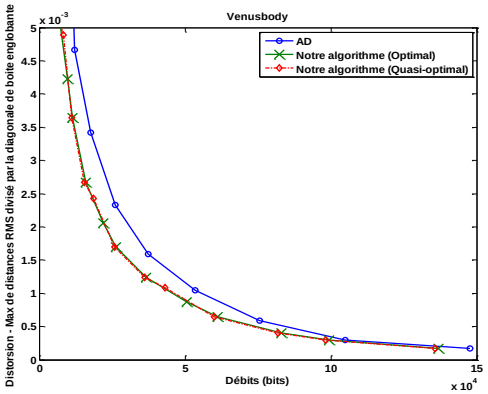


Figure 8 – *Courbe de débit-distorsion du modèle Venusbody.*

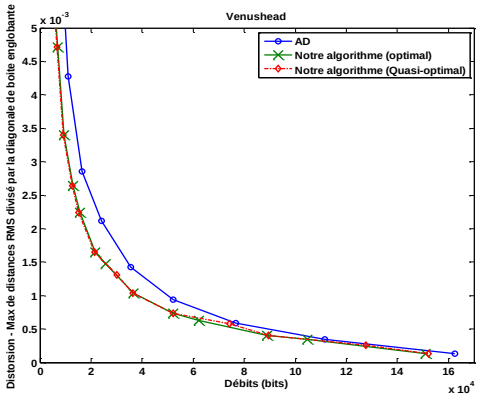


Figure 9 – *Courbe de débit-distorsion du modèle Venushead.*

Le tableau 1 nous donne le temps de calcul en secondes de l'algorithme d'Alliez et Desbrun [6] (AD), notre algorithme optimal et notre algorithme quasi-optimal. L'utilisation de l'estimateur de quantification permet de réduire significativement le temps de calcul avec un résultat similaire.
Le temps supplémentaire par rapport à l'algorithme de (AD) vient du temps nécessaire pour les étapes de diminution de quantification.

Tableau 1 – *Comparaison de temps de calcul.*

Modèle	# sommets	Q	Temps de calcul (s)		
			AD	Optimal	Quasi-optimal
Horse	19851	12	0.46	131.81	1.63
Mannequin	11703	10	0.23	49.79	0.49
Fandisk	6475	10	0.12	38.81	0.34
Venusbody	11362	10	0.22	49.61	0.44
Venushead	8268	10	0.17	32.68	0.46
Rabbit	67039	12	1.72	662.1	3.80

5 Conclusion et perspectives

Nous avons proposé une nouvelle méthode de compression progressive de maillages qui repose sur l'adaptation de la précision de quantification. Le niveau de quantification est optimisé en fonction du nombre d'éléments en permettant donc d'améliorer le débit-distorsion, particulièrement à bas débit. Afin de réduire le temps nécessaire pour l'encodage, nous avons aussi proposé une méthode quasi-optimale qui détermine la quantification adaptée à chaque maillage intermédiaire en utilisant une simple estimation.
Une des perspectives est d'étendre ce travail pour compresser les attributs associés, tels que la couleur et la normale, en adaptant également la précision de quantification de ces attributs afin d'améliorer le débit-distorsion.

Remerciements

Ce travail est soutenu financièrement par l'ANR grâce au programme COSINUS (projet COLLAVIZ n°ANR-08-COSI-003).

Références

[1] H. Hoppe. Progressive meshes. In *ACM SIGGRAPH,* 99–108, 1996.

[2] G. Taubin, A. Guéziec, W. Horn et F. Lazarus. Progressive forest split compression. In *ACM SIGGRAPH,* 123–132, 1998.

[3] R. Pajarola et J. Rossignac. Compressed progressive meshes. *IEEE Transactions on Visualization and Computer Graphics,* 6(1) :79–93, 2000.

[4] Z. Karni, A. Bogomjakov et C. Gotsman. Efficient compression and rendering of multi-resolution meshes. In *IEEE Visualization Conference Proceedings,* 347–354, 2002.

[5] D. Cohen-Or, D. Levin et O. Remez. Progressive compression of arbitrary triangular meshes. In *IEEE Visualization Conference Proceedings,* 67–72, 1999.

[6] P. Alliez et M. Desbrun. Progressive compression for lossless transmission of triangle meshes. In *ACM SIGGRAPH,* 198–205, 2001.

[7] C. Touma et C. Gotsman. Triangle mesh compression. In *Proceedings of Graphics Interface,* 26–34, 1998

[8] P. Alliez et M. Desbrun. Valence-driven connectivity encoding for 3D meshes. In *Eurographics,* 480–489, 2001.

[9] P.-M. Gandoin et O. Devillers. Progressive lossless compression of arbitrary simplicial complexes. *ACM Transactions on Graphics,* 21(3) :372–379, 2002.

[10] J. Peng et C.-C.J. Kuo. Geometry-guided progressive lossless 3D mesh coding with octree (OT) decomposition. In *ACM SIGGRAPH,* 609–616, 2005.

[11] S. Valette, R. Chaine et R. Prost. Progressive Lossless Mesh Compression Via Incremental Parametric Refinement. *Computer Graphics Forum (Proceedings of Symposium on Geometry Processing 2009),* 28(5) :1301–1310, 2009.

[12] D. King et J. Rossignac. Optimal bit allocation in 3D compression. *Journal of Computational Geometry, Theory and Applications,* 14 :91–118, 1999.

[13] F. Payan et M. Antonini. An efficient bit allocation for compressing normal meshes with an error-driven quantization. *Computer Aided Geometric Design,* 22 :466–486.

[14] H. Lee, G. Lavoué et F. Dupont. Adaptive coarse-to-fine quantization for optimizing rate-distortion of progressive mesh compression. In *VMV,* 73–81, 2009.

[15] P. Cignoni, C. Rocchini et R. Scopigno. Metro : Measuring error on simplified surfaces. *Computer Graphics Forum,* 17(2) :167–174, 1998.

Estimation de l'attitude d'un satellite par recalage d'images

R. Perrier[1] E. Arnaud[2] P. Sturm[1] M. Ortner[3]

[1] INRIA Rhône Alpes, [2] Université Joseph Fourier LJK, [3] EADS Astrium

[1] 655 avenue de l'Europe, 38330 Montbonnot, France

{regis.perrier, elise.arnaud, peter.sturm}@inrialpes.fr

Résumé

La plupart des applications d'imagerie satellitaire utilisent des cameras de type pushbroom pour imager une surface observée. Ce capteur linéaire est toujours fixé sur une plateforme mobile, il acquiert une image 1-D à chaque instant de temps et utilise le mouvement rectiligne de la plateforme pour balayer une zone à photographier. lorsqu'elles sont successivement collées entre elles, l'ensemble des images 1-D forment une image 2-D complète. La stabilité du satellite est un point crucial lors de l'aquisition d'une image 2-D complète. Cette acquisition peut prendre plusieurs secondes, et de légères variations de l'attitude [1] du satellite peuvent provoquer des déformations géometriques importantes dans l'image acquise. Dans cet article, nous proposons une méthode de recalage multi-image permettant d'estimer toute variation d'attitude du satellite afin de rectifier les images acquises. Nous utilisons une méthode de type Lucas Kanade associée à un modèle polynomial par morceaux des variations d'attitude. Nous démontrons les perfomances de notre approche sur deux jeux de données satellitaires.

Mots clefs

Camera pushbroom, recalage d'image, modèle polynômial par morceaux, satellite.

1 Introduction

L'imagerie aérienne et satellitaire est un domaine de recherche très actif pour plusieurs applications : reconstruction 3-D, super-resolution, fusion d'information pour ne citer que ceux-ci [1]. Dans la plupart des cas, les images sont supposées être prises par des caméras de type sténopé, alors que les applications d'observation de la terre utilisent majoritairement des capteurs de type pushbroom. Ce dernier est un CCD (Charge-Coupled Device) linéaire qui capture une image 1-D par instant de temps. Il est généralement embarqué sur une plateforme qui se déplace perpendiculairement à son axe au cours du temps, et l'ensemble des images 1-D accolées forment une image 2-D. Son principe d'acquisition peut se résumer à la figure 1. Les raisons de

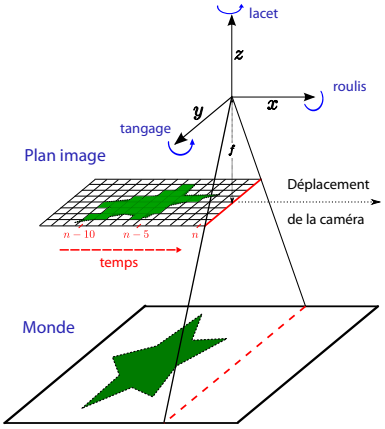


Figure 1 – *Principe de l'acquisition pushbroom : la caméra se déplace le long de l'axe x et enregistre une image 1-D à chaque pas de temps n ; l'ensemble des images 1-D accolées forment une image 2-D. y est l'axe de la caméra et z l'axe orthogonal au plan image. L'orientation de la caméra est défini par le lacet (rotation autour de l'axe z), le roulis (rotation autour de l'axe x) et le tangage (rotation autour de l'axe y)*

son utilisation très courante pour les applications d'observation sont les suivantes [2] :

– une caméra pushbroom peut enregistrer plus de 25000 pixels sur une seule ligne (image 1-D), en se déplaçant elle forme une bande d'image de résolution très importante ;
– à résolution équivalente, un capteur pushbroom coûte nettement moins cher qu'un capteur sténopé (qui correspond à un CCD 2-D).

Afin de couvrir différentes bandes spectrales telles que le rouge, le vert et le bleu, plusieurs capteurs pushbroom sont montés en parallèle sur un même plan focal, la restitution d'un image colorée se faisant par superposition des images des trois bandes. La figure 2 montre un exemple typique de ce plan focal. Le capteur pushbroom est également présent dans des applications plus courantes : scanner de documents, imageurs médicaux à rayons X et scanner d'inspection pour la surveillance aux aéroports.

1. L'attitude (ou orientation) en aéronautique correspond à la direction des axes d'un engin spatial par rapport à un repère de référence. Elle se définit par le lacet, le roulis et le tangage (les rotations autour des 3 axes).

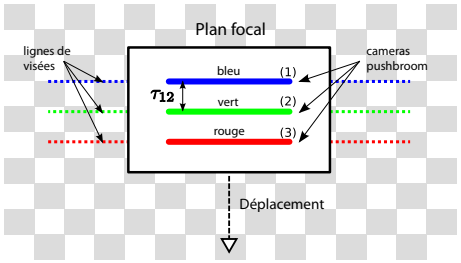


Figure 2 – *Géométrie standard d'un plan focal de satellite d'observation possédant 3 caméras pushbroom de radiométrie différente : rouge, verte et bleue (respectivement énumérées par 1, 2 et 3). Notons que ce qui est vu par la caméra 2 à l'instant n sera vu par la caméra 1 à l'instant $n + \tau_{12}$.*

Peu de travaux prennent en compte la spécificité de ce capteur, et jusqu'à maintenant il était supposé stable lors de l'acquisition d'une image 2-D [3, 4]. Cependant dans le contexte de l'imagerie aérienne et satellitaire, cette hypothèse est fragilisée par l'environnement dans lequel évolue l'imageur, ainsi que par sa structure mécanique. Des perturbations athmosphériques et des vibrations liées aux moteurs peuvent engendrer des variations de son attitude lors du processus d'acquisition. En conséquence, des déformations plus ou moins importantes peuvent apparaître dans l'image. La figure 3 presente un exemple synthétique de déformation suivant différentes variations d'attitude. Ce

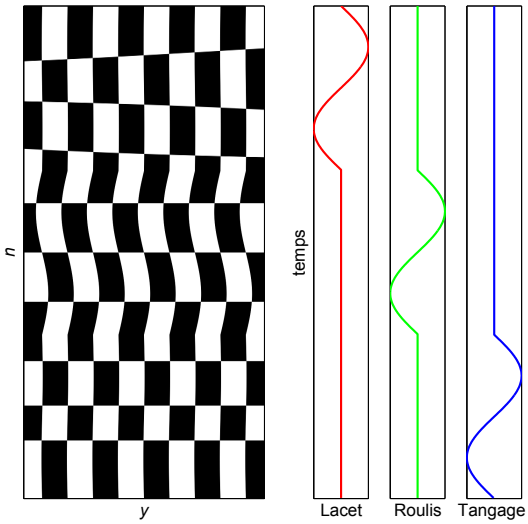


Figure 3 – *Exemple de déformations observées sur un damier régulier lorsque l'attitude de la caméra pushbroom varie lors de l'acquisition. Sur le damier, le lacet fait varier l'inclinaison des lignes horizontales, le roulis provoque l'oscillation des lignes verticales et le tangage augmente ou diminue la hauteur des cases.*

phénomène existe depuis plusieurs années dans le contexte de l'imagerie aérienne [5], et il pourrait apparaitre et s'amplifier dans le domaine de l'imagerie satellitaire avec le design de satellite plus petit et possédant une résolution d'image plus importante [6].

L'utilisation de gyroscopes, GPS et senseurs stellaires – dans le cas satellitaire pour ce dernier– permet de contourner ce problème ; plusieurs travaux font état de résultats satisfaisants dans la restitution de l'attitude à l'aide de ces capteurs [5] et, *in fine*, la correction des images. Cependant, la fréquence d'échantillonnage de ces capteurs comprise entre 4 et 16 Hz est nettement inférieure à la fréquence d'acquisition de chaque image 1-D faite par le capteur pushbroom (de l'ordre de 2500Hz). Par conséquent, une partie de l'information de mouvement de l'imageur ne peut pas être restituée par ces capteurs. De même dans le contexte satellitaire, les contraintes liées à l'environnement spatial rendent coûteuses l'utilisation de ce type de capteur ; pouvoir s'affranchir de ceux-ci serait un bénéfice non négligeable.

Comme nous l'avons remarqué précédemment, plusieurs capteurs pushbroom de modalités différentes sont généralement associés sur un même plan focal. De fait, une variation de l'attitude du satellite lors de l'acquisition perturbe l'ensemble des images. Par comparaison et recalage de l'ensemble de ces images, il serait possible de restituer les mouvements d'attitude inconnus du satellite. Cette idée a été suggérée par [6] ; leur approche utilise des correlations locales sur des points d'interêts extraits entre les différentes images. Cependant le caractère local de leur méthode d'estimation du mouvement contraint à procéder par blocs successifs pour restituer une variation d'attitude **absolue**. Dans ce cas, le problème de recalage multi-image ne peut pas être considéré globalement comme décrit dans [7], et l'estimation est sous optimale.

Dans ce papier, nous proposons de résoudre le problème d'estimation des variations d'attitude du satellite par un recalage multi-image et selon une méthode de type Lucas-Kanade [8, 7] ; ces techniques sont très bien adaptées à des problèmes de recalage sous-pixeliques. Nous exploitons la géométrie du plan focal ainsi que les origines mécaniques des variations d'attitude pour recaler les images dans un même système de coordonnées. Nous modélisons les variations d'attitude par une fonction polynomiale par morceaux sous contraintes. L'algorithme proposé est simple et ne nécessite pas de paramètres de régularisation, dont la détermination est souvent difficile.

2 Estimation de l'attitude

Par la suite, nous appelerons $\mathbf{I} = \{I_1, I_2, I_3\}$ l'ensemble des trois images rouge, verte et bleue, conformément au plan focal de la figure 2 ; les valeurs radiométriques de chaque image appartiennent à un même ensemble $\mathcal{V}$. Les pixels de chaque image sont référencés par leur coordonnées $[n, y]$ et appartiennent à l'ensemble $\mathcal{S}$; $n \in [1, N]$ correspond également aux instant de temps discrétisés de prise des images 1-D par la caméra pushbroom et N est le nombre total d'échantillons de temps (il est aussi équivalent au nombre de lignes de chaque image 2-D de $\mathbf{I}$).

Nous appelons τ_{ij} le temps qui espace l'observation d'une même scène entre les caméras pushbroom i et j. Nous supposerons qu'à l'échelle de l'acquisition d'une image, le satellite maintient une vitesse constante et effectue un mouvement de translation rectiligne ; ces hypothèses sont généralement très fiables [2]. En supposant que le satellite maintienne son attitude constante lors de la prise de vue, le cas idéal nous donnerait la relation suivante :

$$I_i(n, y) - f_{ij}(I_j(n + \tau_{ij}, y)) \sim \mathcal{N}(0, \sigma^2) \qquad (1)$$

où $f_{ij} : \mathcal{V} \to \mathcal{V}$ est une fonction linéaire chargée de compenser les différences radiométriques entre les images i et j, et σ^2 denote la variance d'un bruit gaussien centré i.i.d. sur tous les pixels de l'image. Afin d'obtenir une formulation plus reduite, nous appelons $I_j^{\tau_{ij}}$ l'image I_j décalée en temps telle que $I_j^{\tau_{ij}}(n, y) = I_j(n + \tau_{ij}, y)$, et $\mathbf{y} = [n, y]^T$ le vecteur des coordonnées de chaque pixel, l'équation (1) peut se réécrire de la façon suivante :

$$I_i(\mathbf{y}) - f_{ij}(I_j^{\tau_{ij}}(\mathbf{y})) \sim \mathcal{N}(0, \sigma^2). \qquad (2)$$

En réalité, les images sont déformées par les mouvements de rotation du plan focal autour de ses trois axes. Nous nommons $\theta(n) \in \Theta$ l'attitude inconnue du satellite pour l'instant de temps n ; c'est un vecteur (3×1) dont les composantes respectives sont le lacet $\theta_l(n)$, le roulis $\theta_r(n)$ et le tangage $\theta_t(n)$. Afin d'éviter toute redondance dans les équations en section 2.2, nous nous réfèrons à $\theta_\alpha(n)$ où α correspond séparemment au lacet, au roulis et au tangage. Nous appelons $\boldsymbol{\theta}$ le vecteur $(3N \times 1)$ qui rassemble l'attitude pour tous les instants de temps :

$$\boldsymbol{\theta} = [\theta_l(1) \dots \theta_l(N), \theta_r(1) \dots \theta_r(N), \theta_t(1) \dots \theta_t(N)]^T. \qquad (3)$$

Soit $W : \mathcal{S} \times \Theta \to \mathcal{S}$ la fonction de déformation qui déplace le pixel à une nouvelle coordonnée suivant l'attitude du satellite. A partir de l'équation (2), ce que nous observons réellement est :

$$I_i\big(W(\mathbf{y}; \theta(n))\big) - f_{ij}\big(I_j^{\tau_{ij}}\big(W(\mathbf{y}; \theta(n+\tau_{ij}))\big)\big) \sim \mathcal{N}(0, \sigma^2), \qquad (4)$$

il est important de noter ici que les deux images sont déformées par θ, mais pour des instants de temps différents.

2.1 Recalage multi-image global

Une technique courante de résolution de l'équation (4) est la méthode de Lucas Kanade qui utilise de façon optimale tous les pixels des images [8]. Cette méthode s'applique d'autant mieux à notre contexte puisque les déformations observées sont de quelques pixels et que nous souhaitons avoir une précision sous-pixelique dans le recalage d'images. Nous appelons $T_{\tau_{ij}}$ l'opérateur qui décale les échantillons de $\boldsymbol{\theta}$ d'un facteur τ_{ij}. Cet opérateur est une matrice creuse $(3N \times 3N)$ qui, pour chaque ligne $n \in [1, 3N]$ et colonne $m \in [1, 3N]$, vaut :

$$T_{\tau_{ij}}(n, m) = \begin{cases} 1 & \text{pour } n = m - \tau_{ij}, \text{et } n \notin]N - \tau_{ij}, N] \\ & \qquad \cup]2N - \tau_{ij}, 2N] \cup]3N - \tau_{ij}, 3N] \\ 0 & \text{ailleurs} \end{cases}$$

L'équation globale multi-image à minimiser est la suivante :

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i,j;i\neq j} \sum_{\mathbf{y}\in\mathcal{S}} \Big(I_i\big(W(\mathbf{y}; \boldsymbol{\theta})\big) - f_{ij}\big(I_j^{\tau_{ij}}\big(W(\mathbf{y}; T_{\tau_{ij}}\boldsymbol{\theta})\big)\big)\Big)^2. \qquad (5)$$

En opposition à la méthode de Lucas Kanade traditionnelle, notons qu'ici, il n'y a pas d'image référence à laquelle une image déformée vient se ratacher. Dans ce contexte, les deux images sont déformées. En choisissant l'une des deux images comme référence, ce que une technique par corrélation locale impose, la minimisation aboutirait à une estimation **relative** des déformations du type $\theta(n) - \theta(n + \tau_{ij})$. Retrouver la variation **absolue** $\theta(n)$ nécessiterait une étape de déconvolution pour retrouver la variation d'attitude **absolue** $\theta(n)$. Cette observation constitue un des défauts majeurs de [6] qui procède par corrélation locale sur des points d'intérêts. Dans notre cas, la méthode que nous définissons permet une estimation directe de l'attitude **absolue**.

Estimer directement $\boldsymbol{\theta}$ à partir de l'équation (5) est un problème mal posé si aucune contrainte n'est définie pour orienter la minimisation. Il serait possible d'ajouter un *a priori* sur $\boldsymbol{\theta}$ dans l'équation à minimiser, mais le choix de la fonction régularisante n'est pas toujours aisé. De plus, le calcul du coefficient de régularisation, qui équilibre le terme d'attache aux données (images) et le terme de régularisation, n'est pas trivial et est généralement coûteux en temps de calcul. En analysant de plus près les origines du problème, nous savons que les variations d'attitude du satellite sont principalement liées aux moteurs du satellite. Une propriété importante de ce type de mouvement est leur forme stationnaire dans le temps ; l'utilisation de modèles polynômiaux par morceaux est particulièrement appropriée pour approximer ce type de signaux très réguliers.

2.2 Modèle polynômial par morceaux

Pour modéliser les variations d'attitude, nous utilisons un modèle polynomial par morceaux où chaque polynôme est lié à ses voisins par des contraintes ; nous notons les polynômes $\Phi_{k,\alpha}$, $k \in [1, M]$. M est le nombre de polynômes nécessaire pour restituer les variations d'attitude. Il peut également être vu comme le nombre de fenêtres successives dans le temps qui englobent $\boldsymbol{\theta}$. Nous appelons n_k pour $k \in [1, M-1]$ les instants de temps où les contraintes entre deux polynômes successifs sont définies. Ceux-ci peuvent être mis en lien avec les points de contrôle des splines [9], à l'exception que nos points sont séparés par plusieurs échantillons de temps. Nous exposons le principe de ces polynômes par morceaux en figure 4. Nous pouvons écrire les variations d'attitude de la façon suivante :

$$\theta_\alpha(n) = \sum_{k=1}^{M} w_k(n) \Phi_{k,\alpha}(n), \qquad (6)$$

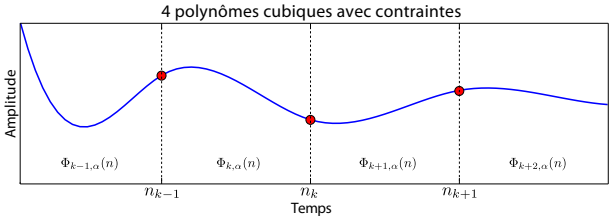


Figure 4 – *Exemple de fonction polynômiale par morceaux où 4 polynômes* $(\Phi_{k-1,\alpha}, \Phi_{k,\alpha}, \Phi_{k+1,\alpha}, \Phi_{k+2,\alpha})$ *sont contraints aux instants de temps* n_{k-1}, n_k *et* n_{k+1}. *Les points rouges montrent l'emplacement des contraintes et les lignes pointillées verticales délimitent chaque polynôme local. Dans cette exemple, nous avons choisi des polynômes de degré 3 et des contraintes jusqu'à leur dérivée seconde.*

où les $w_k(n)$ correspondent à des fonctions de pondération telles que :

$$w_k(n) = \begin{cases} 1 & \text{pour } n \in [n_{k-1}, n_k[\\ 0 & \text{ailleurs.} \end{cases} \quad (7)$$

Enfin les polynômes d'ordre P sont définis tels que :

$$\Phi_{k,\alpha}(n) = \sum_{p=0}^{P} a_{p,k,\alpha} n^p. \quad (8)$$

Les coefficients polynomiaux $a_{p,k,\alpha}$ caractérisent entièrement les variations d'attitude. Nous appelons $\mathbf{a}$ le vecteur $(3MP \times 1)$ qui contient l'ensemble des coefficients polynômiaux. Les contraintes entre chaque polynôme aux points n_k où $k \in [1, M-1]$ sont définies comme suit :

$$\Phi^{(p)}_{k,\alpha}(n_k) = \Phi^{(p)}_{k+1,\alpha}(n_k) \text{ pour } p \in [0, P-1]. \quad (9)$$

La relation (9) impose la continuité aux points n_k pour les dérivées jusqu'à l'ordre $P-1$. De façon similaire aux splines [9], il n'y a plus qu'un seul degré de liberté par segment de temps $[n_k, n_{k+1}[$. Il faut noter que l'équation (9) définit des contraintes d'égalité linéaires en $\mathbf{a}$, de sorte qu'on ait la relation matricielle suivante :

$$C\mathbf{a} = \mathbf{0}, \quad (10)$$

où C est une matrice creuse de taille $(3(M-1)P \times 3MP)$ et $\mathbf{0}$ est le vecteur nul de dimension $(3(M-1)P \times 1)$. Enfin, la combinaison des équations linéaires (6), (7) et (8) nous donne l'équation matricielle suivante :

$$\boldsymbol{\theta} = H\mathbf{a}, \quad (11)$$

où H est une matrice creuse de dimension $(3N \times 3MP)$ prenant en compte les $w_k(n)$ de l'équation (7) et les n^p de l'équation (8). Nous pouvons désormais reformuler le problème de recalage multi-image exposé en section 2.1 par la minimisation d'une équation non linéaire sous contraintes d'égalité linéaires :

$$\hat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \sum_{i,j;i\neq j} \sum_{\mathbf{y}\in\mathcal{S}} \Big(I_i\big(W(\mathbf{y}; H\mathbf{a})\big) - f_{ij}\big(I_j^{\tau_{ij}}\big(W(\mathbf{y}; T_{\tau_{ij}} H\mathbf{a})\big)\big)\Big)^2, \text{ tel que } C\mathbf{a} = \mathbf{0}. \quad (12)$$

Cette équation globale est minimisée itérativement ; les contraintes d'égalité sont utilisées pour réduire la dimension du problème avec une factorisation QR. Les variables restantes sont ensuite estimées avec une procédure des moindres carrés.
En pratique, la fonction de correction radiométrique f_{ij} définie au paragraphe 2 décrit une relation linéaire entre chaque paire d'image :

$$f_{ij}(I_j(\mathbf{y})) = r_{ij,1} + r_{ij,2} I_j(\mathbf{y}) \quad (13)$$

où les coefficients $[r_{ij,1}, r_{ij,2}]$ sont estimés au début de l'algorithme par une méthode des moindres carrées. La fonction de déformation W peut avoir une formulation analytique pour les trois angles de rotation dans le cas où la scène observée est plane. Dans les autres cas, il faut utiliser des dérivées numériques, corrélées à un modèle numérique d'élévation du terrain (Digital Elevation Model) pour estimer les variations d'attitude. Nous présentons les résultats suivants en choisissant la deuxième possibilité.

3 Expériences

Nous présentons des résultats obtenus sur deux jeux de données satellite qui ont été simulés par la société EADS Astrium. Dans ce contexte, la vérité terrain est connue, mais le processus de création des données nous est totalement inconnu. Ce dernier recrée des conditions réelles d'acquisition et le traitement de ces données peut être considéré comme aussi complexe que sur des données réelles. L'algorithme a été implémenté avec Matlab sur une machine possédant un Core2duo à 3GHz et 3.8GiB de mémoire RAM. Dans les deux jeux de données, les caméras 1-2 et 2-3 sont respectivement espacées de 40 et 20 échantillons de temps conformément à la figure 2 ($\tau_{12} = 40$ et $\tau_{23} = 20$). Toutes les images sont de dimension (2564×900) pixels.
Le premier jeu de données est composé de 3 images multipectrales (rouge, verte et bleue). Nous avons fixé le degré P à 3 et la taille des fenêtres polynômiales à 15 échantillons de temps. L'algorithme a convergé en 230 secondes sur 10 itérations. L'écart type de l'erreur sur l'estimation de l'attitude en roulis et en tangage est inférieur à $\frac{15}{100}$ (résultats détaillés en figure 5(a)). Le second jeu de données comporte 3 images monomodales. Ce second cas est complexe car le signal à estimer comporte plusieurs basses fréquences de forte amplitude. Nous avons choisi de fixer P à 3 et la taille des fenêtres polynomiales à 20 instants de temps. L'agorithme a convergé en 20 itérations et 410 secondes ; l'écart type observé sur l'erreur d'estimation de

l'attitude est inférieur à $\frac{20}{100}$ pour le roulis et le tangage (figure 5(b))). Il faut noter que nous ne donnons pas d'estimation du lacet dans ces résultats car les déformations induites par celui-ci sont inférieures à $\frac{2}{100}$ sur les extrémités des images. En pratique, le roulis et le tangage sont des transformations perspectives de la caméra, ils dominent les déformations géométriques observées.

Les résultats des figures 5(a) et 5(b) montrent les performances et les limites de notre algorithme. La plupart des basses fréquences sont restituées, et le recalage possède une bonne précision sous-pixelique. Le modèle polynomial par morceaux s'applique particulièrement bien dans notre contexte d'estimation d'un processus stationnaire. Cependant dans les deux cas, l'erreur résiduelle contient toujours des composantes hautes fréquences. Ceci est principalement lié au choix de la taille de fenêtre lors de l'estimation. Une petite fenêtre pourra plus facilement restituer des hautes fréquences car le modèle aura plus de degrés de liberté, mais le conditionnement du problème sera moins bon. A l'inverse une grande fenêtre restituera bien les basses fréquences.
Nous sommes actuellement en train de rechercher des solutions permettant de sélectionner automatiquement la taille de fenêtre la plus adaptée. De même, une analyse de la sensibilité de l'algorithme suivant les fréquences des variations d'attitude sera menée afin de confirmer sa robustesse et ses limites exactes. Enfin, notons que le modèle polynomial que nous avons adopté est très similaire aux méthodes de recalage par splines [10, 9], il serait intéressant d'étendre notre travail sur ces techniques sans introduire de fonction régularisante.

4 Conclusion

L'originalité de ce travail repose sur l'estimation de l'attitude d'un satellite par l'utilisation d'une méthode de recalage multi-image de type Lucas Kanade, et pour des images acquises par des caméras pushbroom. Les résultats montrent que la propriété stationnaire du mouvement d'attitude est bien prise en compte dans le modèle polynomial par morceaux. Ce problème est très mal conditionné puisque toutes les images acquises sont déformées, cependant notre algorithme est simple et ne nécessite pas de régularisation.
Nous cherchons maintenant à tester notre algorithme sur plusieurs données satellitaires simulées et réelles. Le problème ouvre plusieurs perspectives de travail intéressantes sur le recalage multi modal qui dans notre cas est approché par un modèle linéaire simple. Une analyse géométrique de l'acquisition pushbroom dans ce contexte devrait également permettre d'obtenir des fonctions de déformations analytiques simplifiées. Enfin, l'utilisation de techniques de sélection de modèle devrait nous permettre de déterminer une taille de fenêtre optimale pour ce problème de recalage.

Références

[1] Z. Liu R. S. Blum. *Multi-sensor image fusion and its applications*. Taylor and Francis, 2005.

[2] Gordon Petrie. Airborne pushbroom line scanners : An alternative to digital frame scanners. *Geoinformatics*, 8(1) :50–57, 2005.

[3] Rajiv Gupta et Richard I. Hartley. Linear pushbroom cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(9) :963–975, 1997.

[4] Jamil Drareni, Peter Sturm, et Sébastien Roy. Plane-Based Calibration for Linear Cameras. Dans *The 8th Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras - OMNIVIS*, 2008.

[5] Daniela Poli. General model for airborne and spaceborne linear array sensors. Dans *International Archives of Photogrammetry and Remote Sensing*, volume 34, 2002.

[6] F. de Lussy, D. Greslou, et L. Gross Colzy. Process line for geometrical image correction of disruptive microvibrations. Dans *International Society for Photogrammetry and Remote Sensing*, pages 27–35, 2008.

[7] S. Farsiu, M. Elad, et P. Milanfar. Constrained, globally optimal, multi-frame motion estimation. Dans *Proc. of the 2005 IEEE Workshop on Statistical Signal Processing*, pages 1396 – 1401, 2005.

[8] Richard Szeliski. Image alignment and stitching : a tutorial. *Found. Trends. Comput. Graph. Vis.*, 2(1) :1–104, 2006.

[9] M. Unser. Splines : A perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6) :22–38, November 1999.

[10] Adrien Bartoli, Mathieu Perriollat, et Sylvie Chambon. Generalized thin-plate spline warps. *Int. J. Comput. Vision*, 88(1) :85–110, 2010.

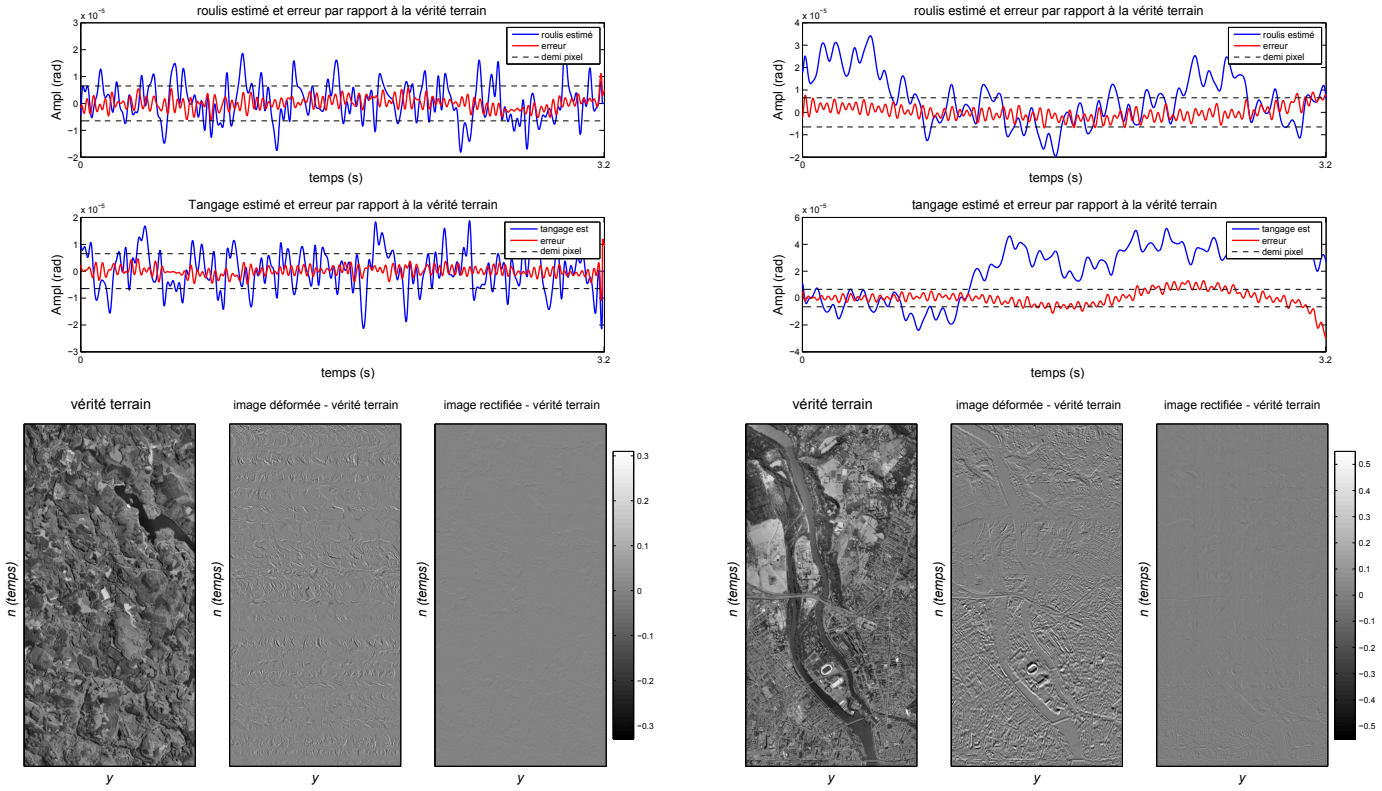


Figure 5 – *(a) Résultats sur le premier jeu de données (b) Résultats sur le second jeu de données : les deux figures du haut montrent l'estimation finale de l'attitude pour le roulis et le tangage, et l'erreur faite en comparaison à la vérité terrain. En dessous, la figure de gauche présente une image extraite de la vérité terrain de taille* $((1000 \times 500))$*. Les deux figures suivantes montrent l'erreur entre l'image acquise et la vérité terrain avant et après rectification.*

Utilisation de l'information photométrique pour la sélection des hyperparamètres en recalage géométrique d'images

F. Brunet[1,2] A. Bartoli[1] N. Navab[2] R. Malgouyres[3]

[1] ISIT - Université d'Auvergne, Clermont-Ferrand
[2] CAMPAR - Technische Universität München, Munich
[3] LIMOS - Université d'Auvergne, Clermont-Ferrand

florent.brunet@u-clermont1.fr

Résumé

Cet article traite du recalage paramétrique d'images à partir de correspondances de points en environnement déformable. Dans ce problème, il est essentiel de déterminer des valeurs correctes pour les hyperparamètres tels que le nombre de points de contrôle du modèle de déformation, un paramètre de régularisation ou l'échelle d'un M-estimateur. Cela est souvent réalisé à la main par tâtonnement ou en optimisant un critère générique comme la validation croisée. Dans cet article, nous proposons un nouveau critère pour sélectionner différents hyperparamètres en combinant les avantages de l'approche géométrique et de l'approche photométrique au recalage d'images. Plus précisément, nous proposons de considérer les correspondances de points comme un jeu d'entrainement et la photométrie comme un jeu d'essai. L'approche proposée est robuste dans la mesure où elle résiste à la fois aux correspondances de points erronées et aux défauts des images comme les occultations ou les spécularités.

Mots clefs

Hyperparamètre, recalage, image, déformation.

1 Introduction

Le recalage d'images consiste à déterminer les paramètres (naturels) d'une déformation de manière à aligner une image source et une image cible. Outre les paramètres naturels, des valeurs correctes doivent être affectées aux hyperparamètres du problème afin d'obtenir un recalage de bonne qualité. Les hyperparamètres sont soit des paramètres additionnels du modèle de déformation (*hyperparamètres de modèle*), soit des paramètres de la fonction de coût à optimiser (*hyperparamètres de coût*). La figure 1 montre à quel point le choix des hyperparamètres est crucial en recalage d'images. Il existe deux approches principales au recalage d'images [1] : l'approche géométrique et l'approche photométrique (ou approche directe). Chacune de ces méthodes a ses propres avantages mais aucune d'entre elles ne permet d'estimer directement des hyperparamètres appropriés. Nous proposons donc de combiner les forces respectives des approches géométriques et photométriques afin de construire une nouvelle méthode de sélection automatique des hyperparamètres en recalage d'images.

Soit $\mathcal{W} : \mathbb{R}^2 \times \mathbb{R}^l \rightarrow \mathbb{R}$ un modèle de déformation paramétré par un ensemble de l paramètres regroupés dans une matrice P ayant l coefficients. L'homographie [1, 4] est un exemple de déformation paramétré par les 8 coefficients indépendants de la matrice d'homographie. Les déformations libres [5] sont un autre exemple de modèle paramétré par $\frac{l}{2}$ points de contrôle 2D. Le nombre de points de contrôle d'une déformation libre ou la largeur de bande des fonctions à bases radiales [6] sont deux exemples parmi d'autres d'hyperparamètres de modèle.

Dans l'*approche géométrique* [1, 7], les images source et cible sont abstraites par un ensemble fini d'éléments saillants caractéristiques. Dans cet article, nous utilisons des points d'intérêt qui sont appariés pour former un ensemble de n correspondances de points $\{\mathbf{p}_i \leftrightarrow \mathbf{q}_i\}_{i=1}^n$. Le principe fondamental de l'approche géométrique consiste à minimiser l'erreur entre les points transformés de l'image source et les points de l'image cible. Une mesure robuste comme un M-estimateur peut être utilisée à cet effet. Des termes supplémentaires, comme un terme de régularisation [7], sont parfois ajoutés afin de faire face, par exemple, aux imprécisions des points ou à un manque de données dans certaines parties des images. Tous ces éléments sont regroupés dans la formulation suivante [7, 8] :

$$\min_{\mathsf{P}} \mathcal{E}(\mathsf{P};\boldsymbol{\theta}), \tag{1}$$

où $\boldsymbol{\theta}$ est un vecteur contenant les hyperparamètres et $\mathcal{E}$ est la fonction de coût ainsi définie :

$$\mathcal{E}(\mathsf{P};\boldsymbol{\theta}) = \sum_{i=1}^{n} \rho\big(\mathcal{W}(\mathbf{p}_i;\mathsf{P}) - \mathbf{q}_i;\gamma\big) + \lambda\mathcal{R}(\mathsf{P}), \tag{2}$$

avec ρ un M-estimateur, γ son facteur d'échelle, $\mathcal{R}$ un terme de régularisation [1] et λ un paramètre contrôlant le compromis entre attache aux données et régularisation. Les

1. Comme, par exemple, l'énergie de torsion (détaillée dans §4).

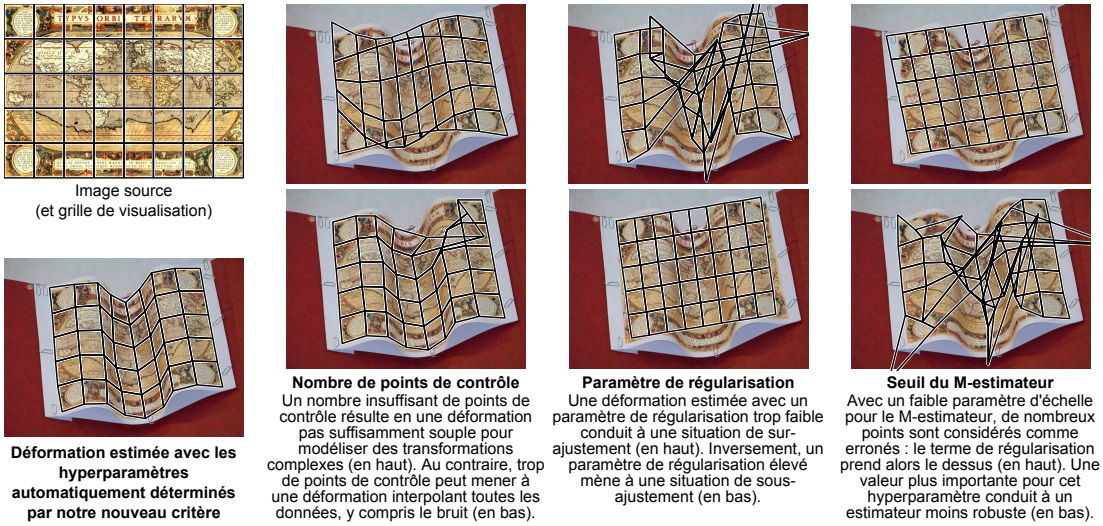


Figure 1 – *Illustration de l'influence de quelques hyperparamètres sur le recalage d'images. Dans cet exemple, les paramètres naturels de la déformation sont estimés à partir de correspondances de points extraites avec SIFT [2, 3]. L'approche proposée dans cet article permet de déterminer automatiquement des valeurs adéquates pour les hyperparamètres en combinant de manière judicieuse les approches géométriques et photométriques.*

valeurs γ et λ sont deux exemples d'hyperparamètres de coût. D'autres hyperparamètres apparaissent si l'on décide, par exemple, d'ajouter d'autres termes. Les atouts de l'approche géométrique sont qu'elle fonctionne lorsque la magnitude des déformations est importante et qu'elle est efficace en terme de temps de calcul [2]. Cependant, l'approche géométrique ne permet pas de déterminer les hyperparamètres. En particulier, comme expliqué en §2, il est *impossible* d'inclure les hyperparamètres directement dans le problème (1), c.-à-d. $\min_{\mathrm{P},\boldsymbol{\theta}} \mathcal{E}(\mathrm{P},\boldsymbol{\theta})$.

L'autre approche au recalage d'images est l'*approche photométrique* (ou *approche directe*) [10, 11] qui détermine les paramètres de la transformation en minimisant la différence de couleur entre les pixels de l'image cible transformée et ceux de l'image source. L'avantage principal de cette méthode réside dans l'importante densité des données (c.-à-d. les pixels) utilisées pour estimer les paramètres de la déformation. De la même manière qu'avec l'approche géométrique, l'approche photométrique à elle seule ne permet pas de déterminer les hyperparamètres.

Dans la mesure où les hyperparamètres ne peuvent pas être trivialement estimés, ils sont souvent fixés empiriquement et de manière définitive pour une application donnée. Il est aussi possible de les déterminer par tâtonnement pour chaque paire d'images à recaler. Bien entendu, cette technique n'est pas satisfaisante à cause de son manque d'automatisme et de son manque d'« objectivité ». Il existe des méthodes génériques permettant de sélectionner automatiquement certains hyperparamètres. Elles consistent généralement à minimiser un critère qui évalue la qualité de la transformation estimée en fonction des hyperparamètres. Par exemple, ces critères peuvent mesurer la capacité d'une déformation à prédire de nouvelles données. Le critère informatif d'Akaike (*Akaike Information Criterion*) [12]), le critère C_P de Mallow (*Mallow's* C_P) [13], le critère de longueur minimale de description (*Minimum Description Length*) ou les techniques de validation croisée [8, 14, 15] (détaillées en §2) sont des exemples de telles approches. Notons néanmoins qu'aucune de ces méthodes n'est spécifique au problème du recalage d'images et que, par conséquent, aucune n'exploite pleinement les particularités des données de ce problème.

Les approches existantes pour la sélection des hyperparamètres ont la caractéristique commune de n'employer uniquement que les correspondances de points. Or, en recalage d'images, une autre information est disponible : la photométrie. Nous proposons donc un nouveau critère qui utilise *toute* l'information disponible : les correspondances de points sont utilisées comme un jeu d'entrainement et l'information photométrique est utilisée comme un jeu de test. Autrement dit, nous proposons de combiner les avantages des deux approches classiques : l'approche géométrique est employée pour déterminer les paramètres naturels de la déformation tandis que les principes de l'approche photométrique servent à la sélection des hyperparamètres. Notre critère est plus flexible que les approches statistiques classiques dans la mesure où des hyperparamètres de types différents (entiers ou réels) peuvent être simultanément recherchés. Des expérimentations sur données synthétiques et réelles sont menées en §4 pour différents hyperparamètres avec comme modèle de déformation des B-splines bidimensionnelles.

Notations. Les scalaires sont notés en italique (x), les vecteurs en gras ($\mathbf{p}$) et les matrices en caractères sans-

2. Cela est particulièrement vrai lorsque l'on utilise un détecteur de points efficace (comme SIFT [2] ou SURF [9]) et un bon appariteur (comme celui implémenté dans [3]).

serif (M). La norme euclidienne d'un vecteur $\mathbf{v}$ est notée $\|\mathbf{v}\|$. Les images, notées en fonte calligraphique ($\mathcal{I}$), sont considérées comme des fonctions de $\mathbb{R}^2$ dans $\mathbb{R}^c$ avec c le nombre de canaux. Leur évaluation pour des coordonnées non entières est réalisée par interpolation bilinéaire.

2 Travaux antérieurs sur la sélection d'hyperparamètres

2.1 Sélection automatique

Différents hyperparamètres ont été présentés dans l'introduction. Il est important de comprendre que des résultats aberrants sont obtenus si l'on introduit les hyperparamètres directement dans le problème (1). Par exemple, si le terme de régularisation est un terme toujours positif, mettre λ à 0 est la manière la plus simple de diminuer sa contribution. De même, le coût est artificiellement diminué lorsque l'échelle γ du M-estimateur tend vers 0 puisque cela conduit à considérer comme erronés la quasi totalité des points (et que le coût assigné à de tels points tend vers 0 quand $\gamma \to 0$).
L'approche habituelle pour construire une méthode de sélection automatique des hyperparamètres consiste à créer un critère $\mathcal{C}$ qui évalue la qualité d'un jeu donné d'hyperparamètres [8, 16]. L'estimation conjointe des paramètres et des hyperparamètres est alors réalisée en optimisant le problème de minimisation imbriqué suivant :

$$\min_{\mathsf{P}} \mathcal{E}(\mathsf{P}; \arg\min_{\boldsymbol{\theta}} \mathcal{C}(\boldsymbol{\theta})). \tag{3}$$

Remarquons bien que l'introduction du critère $\mathcal{C}$ rend le problème (3) complètement différent du problème dégénéré $\min_{\mathsf{P},\boldsymbol{\theta}} \mathcal{E}(\mathsf{P}; \boldsymbol{\theta})$.

2.2 La validation croisée

La validation croisée (VC) est un principe général à fondements statistiques communément utilisée pour déterminer de manière automatique des hyperparamètres [16]. Dans le contexte du recalage géométrique d'images, une procédure de VC consiste à minimiser un critère (fonction des hyperparamètres) mesurant la capacité qu'a une déformation (estimée pour un jeu donné d'hyperparamètres) à prédire de nouveaux points. Pour cela, l'ensemble des correspondances est partitionné. Chacune des parties est ensuite utilisée alternativement comme jeu d'entrainement et comme jeu de test afin de construire le critère de la VC. Ce critère a été utilisé par [8] dans le cadre du recalage d'images. Nous présentons maintenant deux variantes de la VC : la VC Ordinaire et la *V-fold Cross-Validation*.

Validation Croisée Ordinaire (VCO). Étant donné un jeu d'hyperparamètres $\boldsymbol{\theta}$, soit $\mathsf{P}_{\boldsymbol{\theta}}^{(k)}$ la matrice des paramètres de la déformation estimée en laissant la k-ème correspondance de côté. Le critère de la VCO est défini ainsi :

$$\mathcal{C}_{VCO}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{k=1}^{n} \left\| \mathbf{q}_k - \mathcal{W}\left(\mathbf{p}_k; \mathsf{P}_{\boldsymbol{\theta}}^{(k)}\right)\right\|^2. \tag{4}$$

Sélectionner les hyperparamètres avec la VCO consiste à minimiser $\mathcal{C}_{VCO}$ par rapport à $\boldsymbol{\theta}$. La VCO présente différents inconvénients. Premièrement, son calcul est extrêmement coûteux : l'évaluation avec (4) de $\mathcal{C}_{VCO}$ pour une seule valeur de $\boldsymbol{\theta}$ nécessite l'estimation de chacune des n matrices $\{\mathsf{P}_{\boldsymbol{\theta}}^{(k)}\}_{k=1}^{n}$. Il existe des approximations de la formule (4) pour réduire les temps de calcul mais celles-ci ne sont valables que dans le cadre d'une estimation des paramètres par moindres carrés [8, 17]. Deuxièmement, le critère $\mathcal{C}_{VCO}$ n'est pas robuste aux correspondances de points erronées. Enfin, les valeurs calculées avec le critère $\mathcal{C}_{VCO}$ ne sont pas fiables lorsque le nombre de correspondances de points est faible [16].

V-fold Cross-Validation (VCV). Le principe de la VCV est de diviser l'ensemble des correspondances de points en V sous-ensembles disjoints de tailles à peu près identiques (avec V souvent choisi comme $V = \min(\sqrt{n}, 10)$). Une étude détaillée de la VCV est donnée dans [14]. Soit $\mathsf{P}_{\boldsymbol{\theta}}^{[v]}$ la matrice des paramètres de la déformation obtenue en laissant le v-ème sous-ensemble de côté et soit m_v le nombre de points de ce même sous-ensemble. Le critère de la VCV est donné par :

$$\mathcal{C}_{VCV}(\boldsymbol{\theta}) = \sum_{v=1}^{V} \frac{m_v}{n} \sum_{k=1}^{m_v} \frac{1}{m_v} \left\| \mathbf{q}_k - \mathcal{W}\left(\mathbf{p}_k; \mathsf{P}_{\boldsymbol{\theta}}^{[v]}\right)\right\|^2. \tag{5}$$

Comme pour la VCO, ce critère n'est pas robuste aux données erronées. Il est cependant possible de le rendre robuste [14] en remplaçant la moyenne $\sum_{k=1}^{m_v} \frac{1}{m_v} \|\mathbf{q}_k - \mathcal{W}(\mathbf{p}_k; \mathsf{P}_{\boldsymbol{\theta}}^{[v]})\|^2$ dans l'équation (5) par une mesure plus robuste comme la moyenne tronquée.

2.3 Autres approches

Il existe d'autres approches permettant de sélectionner automatiquement des hyperparamètres. En voici quelques exemples : le critère informatif d'Akaike (*Akaike Information Criterion*) [12]), le critère C_P de Mallow (*Mallow's* C_P) [13] ou le critère de longueur minimale de description (*Minimum Description Length*). Il existe aussi des variantes robustes de ces critères ; par exemple, une version robuste du critère C_P de Mallow est proposée dans [13]. Le principal inconvénient de ces approches est qu'elles ont été conçues pour sélectionner un modèle parmi un ensemble donné de modèles [8]. Elles sont donc mal adaptées pour le réglage d'hyperparamètres continus comme l'échelle d'un M-estimateur. De plus, ces méthodes sont peu fiables lorsque le nombre de correspondances de points est faible.

3 Notre nouveau critère

Les approches présentées en §2 ont la caractéristique commune de n'utiliser que les correspondances de points, aussi bien pour l'estimation des paramètres que pour l'estimation des hyperparamètres. Nous proposons ici une nouvelle approche qui utilise toute l'information disponible c.-à-d.

les correspondances de points mais aussi les données photométriques. Pour cela, nous combinons les approches géométriques et photométriques de la manière suivante :

– étant donné un jeu d'hyperparamètres $\boldsymbol{\theta}$, l'approche géométrique est utilisée pour estimer les paramètres naturels de la déformation ;
– la fonction de coût typiquement optimisée dans l'approche photométrique est utilisée pour évaluer la qualité des hyperparamètres $\boldsymbol{\theta}$ en se basant sur l'idée suivante : de bons hyperparamètres doivent conduire à une estimation des paramètres induisant une déformation qui minimise la différence photométrique entre l'image source et l'image cible transformée.

En d'autres termes, nous proposons d'utiliser les correspondances de points comme un jeu d'apprentissage et les pixels des images (c.-à-d. l'information photométrique) comme un jeu de test. Étant donnés des paramètres naturels $\mathsf{P}_{\boldsymbol{\theta}}$ issus de l'estimation de la déformation à partir de correspondances en utilisant un ensemble donné d'hyperparamètres $\boldsymbol{\theta}$, notre critère est ainsi défini :

$$\mathcal{C}_\star(\boldsymbol{\theta}) = \frac{1}{|\mathfrak{R}|} \sum_{\mathbf{p}\in\mathfrak{R}} \left\|\mathcal{S}(\mathbf{p}) - \mathcal{T}(\mathcal{W}(\mathbf{p};\mathsf{P}_{\boldsymbol{\theta}}))\right\|^2, \quad (6)$$

où $\mathcal{S}$ et $\mathcal{T}$ sont respectivement les images source et cible et où $|\mathfrak{R}|$ est la région d'intérêt (de taille $|\mathfrak{R}|$), souvent définie par un rectangle obtenu en supprimant une marge au domaine de l'image source.

Notons que le critère de l'équation (6) correspond à la fonction de coût typiquement minimisée dans l'approche photométrique [10, 1]. La différence avec l'approche photométrique est que notre critère est considéré comme une fonction des hyperparamètres $\boldsymbol{\theta}$ et non des paramètres naturels P de la déformation.

Robustesse. Les occultations et les spécularités sont deux exemples de phénomènes pouvant être considérés comme des données aberrantes pour notre critère. Les problèmes liés à ce genre d'erreurs peuvent être évités en utilisant une mesure plus robuste que celle utilisée dans l'équation (6), comme une moyenne tronquée :

$$\mathcal{C}'_\star(\boldsymbol{\theta}) = \frac{1}{\frac{100-\alpha}{100}|\mathfrak{R}|} \sum_{\mathbf{p}\in\mathfrak{R}_\alpha} \left\|\mathcal{S}(\mathbf{p}) - \mathcal{T}(\mathcal{W}(\mathbf{p};\mathsf{P}_{\boldsymbol{\theta}}))\right\|^2, \quad (7)$$

où $\mathfrak{R}_\alpha$ est un sous-ensemble de $\mathfrak{R}$ obtenu en enlevant les $\alpha\%$ de pixels qui produisent les valeurs les plus élevées de $\|\mathcal{S}(\mathbf{p}) - \mathcal{T}(\mathcal{W}(\mathbf{p};\mathsf{P}_{\boldsymbol{\theta}}))\|^2$.

4 Résultats expérimentaux

4.1 Détails techniques

Nous spécialisons ici notre contribution générique afin de mener quelques expériences.

Modèle de déformation. Comme dans [5], nous utilisons des déformations libres à base de B-splines. Cette déformation est paramétrée par un ensemble de $\frac{l}{2}$ points de contrôle 2D rangés dans une matrice $\mathsf{P} \in \mathbb{R}^{\frac{l}{2}\times 2}$. Elle est définie par :

$$\mathcal{W}(\mathbf{q};\mathsf{P}) = \mathbf{w}(\mathbf{p})^\mathsf{T}\mathsf{P}, \quad (8)$$

où $\mathbf{w} : \mathbb{R}^2 \to \mathbb{R}^l$ est la fonction définie par :

$$\mathbf{w}(\mathbf{p}) = \begin{bmatrix} N_1(x)N_1(y) & \dots & N_{l_x}(x)N_{l_y}(y)\end{bmatrix}^\mathsf{T}. \quad (9)$$

où $\mathbf{p} = (x, y)$ et où l_x et l_y sont deux hyperparamètres donnant le nombre de points de contrôle selon l'axe des abscisses et des ordonnées respectivement (avec $l = l_x l_y$). Les fonctions $N_i : \mathbb{R} \to \mathbb{R}$ constituent la base de l'espace vectoriel des B-splines à une dimension [18].

Terme de régularisation. Nous utilisons l'énergie de torsion comme terme de régularisation :

$$\mathcal{R}(\mathsf{P}) = \sum_{i=1}^{2} \int_\Omega \left\| \frac{\partial^2 \mathcal{W}^i}{\partial \mathbf{p}^2}(\mathbf{p};\mathsf{P}) \right\|_\mathcal{F}^2 \mathrm{d}\mathbf{p}, \quad (10)$$

où Ω est le domaine de définition de la déformation, $\mathcal{W}^i$ est la i-ème coordonnée de $\mathcal{W}$ et $\|\bullet\|_\mathcal{F}$ est la norme de Frobenius de la matrice hessienne de la déformation. Avec les déformations B-splines, il est possible de montrer qu'il existe une matrice B telle que $\mathcal{R}(\mathsf{P}) = \|\mathsf{BP}\|_\mathcal{F}^2$.

M-estimateur. Nous utilisons le M-estimateur de Cauchy défini par la fonction ρ suivante :

$$\rho(x;\gamma) = \log\left(1 + \frac{x^2}{\gamma^2}\right), \quad (11)$$

où $\gamma \in \mathbb{R}_+^*$ est un hyperparamètre contrôlant l'échelle du M-estimateur (c.-à-d. sa sensibilité aux données erronées). Nous omettons ici les détails mais il peut être montré qu'utiliser cet M-estimateur est raisonnable au vu des erreurs rencontrées lorsque l'on utilise SIFT ou SURF.

Optimisation. Tous les critères étudiés dans les expériences à venir sont optimisés en utilisant une recherche exhaustive. Cette approche est potentiellement coûteuse en temps de calcul mais a l'avantage de produire des résultats fiables car non soumis aux problèmes de minima locaux.

4.2 Données synthétiques

Génération des données. Une image source est générée en extrayant une zone rectangulaire d'une image texturée (choisie aléatoirement parmi 15 images). L'image cible est construite en déformant une autre partie de la même image texturée avec une déformation B-spline connue $\mathcal{W}^\star$ dont les 5×5 points de contrôle sont choisis aléatoirement de manière à ce que l'amplitude moyenne de la déformation soit de 20 pixels. Les tailles des images source et cible sont respectivement 160×160 et 320×240 pixels. Un bruit gaussien centré d'écart type égal à 5% de la valeur maximale d'un pixel est ajouté aux images. Un ensemble $\mathcal{P} = \{\mathbf{p}_i \leftrightarrow \mathbf{q}_i\}_{i=1}^n$ de correspondances de points est construit en choisissant aléatoirement des points dans l'image source et en calculant leurs correspondants dans l'image cible avec $\mathcal{W}^\star$. Ces points sont ensuite perturbés en ajoutant des erreurs qui suivent une distribution de Cauchy avec un paramètre d'échelle de 1 pixel.

Oracle. Nous appelons *oracle* la déformation estimée à partir des points de $\mathcal{P}$ la plus proche possible de la vérité terrain $\mathcal{W}^\star$ [3]. Il s'agit de la déformation induite par les paramètres P_o et les hyperparamètres $\boldsymbol{\theta}_o$ solutions de ce problème :

$$\min_{(\mathsf{P},\boldsymbol{\theta})} \iint_{\Omega_{\mathcal{W}^\star}} \|\mathcal{W}^\star(\mathbf{p}) - \mathcal{W}(\mathbf{p};\mathsf{P})\| \,\mathrm{d}\mathbf{p}. \qquad (12)$$

Erreur Géométrique Relative (EGR). L'EGR mesure la différence entre une déformation estimée et l'oracle. Soit $\boldsymbol{\theta}_\bullet$ les hyperparamètres minimisant le critère $\mathcal{C}_\bullet$ et soit $\mathsf{P}_{\boldsymbol{\theta}_\bullet}$ les paramètres estimés à partir des points pour les hyperparamètres $\boldsymbol{\theta}_\bullet$. L'EGR est définie par :

$$\iint_{\mathbf{p}\in\Omega_\mathcal{S}} \frac{\|\mathcal{W}(\mathbf{p};\mathsf{P}_o) - \mathcal{W}(\mathbf{p};\mathsf{P}_{\boldsymbol{\theta}_\bullet})\|}{\|\mathcal{W}(\mathbf{p};\mathsf{P}_o)\|} \mathrm{d}\mathbf{p}. \qquad (13)$$

Nous comparons dans la figure 2 les EGR obtenues avec différentes approches pour sélectionner l'échelle γ du M-estimateur et le paramètre de régularisation λ :
– notre critère (Photo) et ses variantes robustes pour des seuils α de 25% (Photo25) et de 50% (Photo50) ;
– la VCV (VFold) et ses variantes robustes pour des seuils α de 20% (VFold20) et de 40% (VFold40).

Le nombre de points de contrôle des déformations est fixé à 8×8 et 100 correspondances de points sont utilisées. Les valeurs reportées sont obtenues sur 100 essais différents.
Nous observons sur la figure 2 que les EGR les plus faibles sont obtenues avec notre critère photométrique. Nous remarquons aussi que la VCV non robuste est très sensible à la présence de correspondances de points erronées. Ces valeurs erronées n'ont que très peu d'impact sur la version non robuste de notre critère puisque celui-ce repose principalement sur les données photométriques.

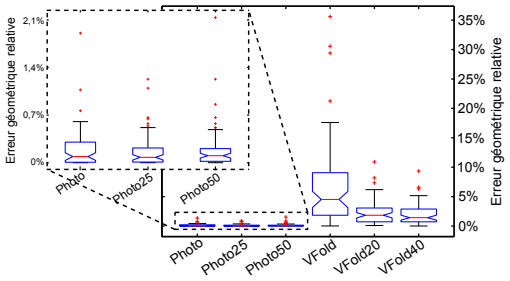


Figure 2 – *« Boîtes à moustaches » des ERG obtenues pour différents critères de sélection des hyperparamètres. De manière générale, notre critère (*Photo*) et ses variantes robustes (*Photo25, Photo50*) donnent de meilleurs résultats que ceux reposant sur la VCV (*VFold, VFold20 *et* VFold40*).*

4.3 Données réelles

Ici, les images sources sont des images numériques et les images cibles sont obtenues en imprimant les images sources puis en les photographiant. Une vérité terrain est établie en cliquant plusieurs centaines de points à la main. Notons que l'illustration de la figure 1 est un exemple de recalage sur de telles données.

Image cubiste. La figure 3 montre les recalages obtenus avec différents critères pour la sélection de ces hyperparamètres : paramètre de régularisation, échelle du M-estimateur et nombre de points de contrôle. L'algorithme SIFT [3] a été utilisé pour extraire 314 correspondances de point, dont approximativement 8% de fausses. De manière générale, nous observons sur la figure 3 que les résultats obtenus avec notre critère sont meilleurs que ceux obtenus avec le critère de la VCV. Le tableau 1 confirme cette constatation. Cela provient principalement du fait qu'il y a un manque de correspondances de points dans la partie inférieure droite des images (notre approche permet de combler ce manque).

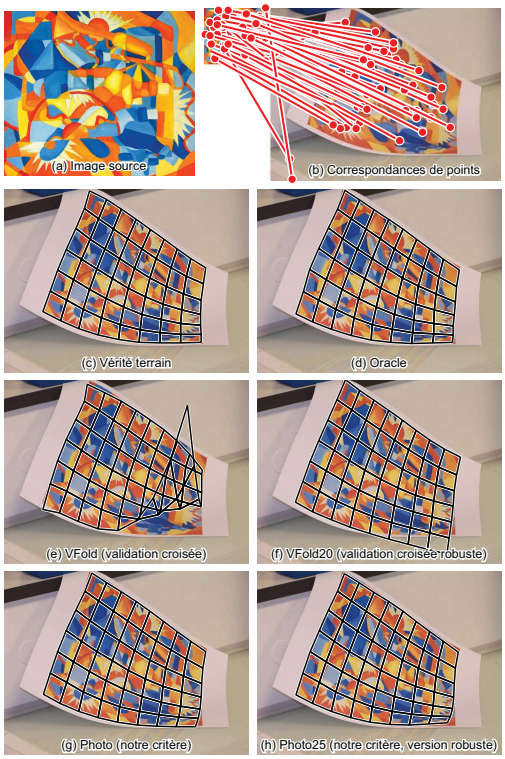


Figure 3 – *Recalage d'images avec 3 hyperparamètres réglés par des critères différents. Dans ce test, les 2 variantes de notre approche donnent les meilleurs résultats.*

Critère	EGR
VCV	1.852%
VCV (variante robuste)	0.675%
Notre critère	0.190%
Notre critère (variante robuste)	0.197%

Tableau 1 – *Erreur géométrique relative (EGR) pour l'expérience de la figure 3.*

3. En fonction des modèles de déformation et des correspondances de points, l'oracle et la vérité terrain ne sont pas nécessairement identiques.

« Waterfall » de Maurits Escher. Nous reportons dans la figure 4 les résultats obtenus pour des tests similaires au test précédent. Ne sont considérés ici que les hyperparamètres de régularisation et d'échelle du M-estimateur. Une occultation a été artificiellement ajoutée dans l'image cible. L'algorithme SURF [9] a été utilisé pour extraire 621 correspondances, dont approximativement 12% de fausses. Comme dans le cas précédent, les hyperparamètres estimés avec nos critères donnent de meilleurs résultats que ceux estimés avec le critère de la VCV. Pour les deux critères, les versions robustes sont meilleures que les versions non robustes. Bien que le critère de la VCV ne repose pas sur l'information photométrique, la présence d'une occultation dans l'image cible influence quand même ce critère puisqu'elle introduit de fausses correspondances de points.

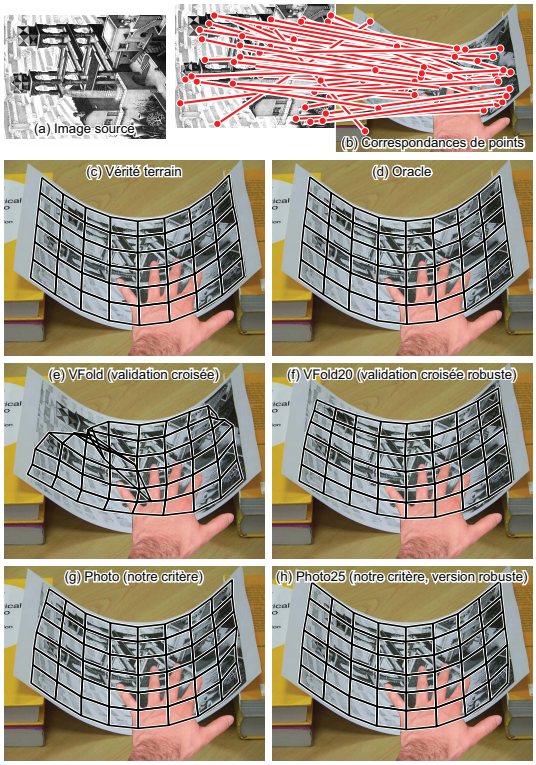


Figure 4 – *Recalage d'images avec 2 hyperparamètres sélectionnés par des critères différents. De manière générale, les variantes robustes aboutissent à de bons résultats. Les meilleurs résultats sont obtenus avec notre critère robuste.*

5 Conclusion

Nous avons proposé un nouveau critère permettant de sélectionner automatiquement des hyperparamètres en recalage d'images. Nous avons montré qu'en général notre critère aboutit à des hyperparamètres meilleurs que ceux obtenus avec d'autres méthodes. Cela est rendu possible par la combinaison des approches géométriques et photométriques au recalage d'images. Bien que pour des raisons pratiques nous ayons limité nos expériences à certains hyperparamètres particuliers, notre approche est générique et pourrait donc être utilisée avec d'autres hyperparamètres ou d'autres modèles de déformations.

Références

[1] R. Szeliski. Image alignment and stitching: A tutorial. *Foundations and Trends in Computer Graphics and Vision*, 2:1–104, 2006.

[2] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[3] A. Vedaldi et B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008.

[4] R. Hartley et A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[5] D. Rueckert, L. Sonoda, C. Hayes, D. Hill, M. Leach, et D. Hawkes. Nonrigid registration using free-form deformations: Application to breast MR images. *IEEE Transactions on Medical Imaging*, 18:712–721, 1999.

[6] F. Bookstein. Principal warps: Thin-Plate Splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:567–585, 1989.

[7] J. Modersitzki. *Numerical Methods for Image Registration*. Oxford Science, 2004.

[8] A. Bartoli. Maximizing the predictivity of smooth deformable image warps through cross-validation. *Journal of Mathematical Imaging and Vision*, 31(2-3):133–145, 2008.

[9] H. Bay, A. Ess, T. Tuytelaars, et L. Van Gool. SURF: Speeded up robust features. *Computer Vision and Image Understanding*, 110:346–359, 2008.

[10] M. Irani et P. Anandan. About direct methods. Dans *Workshop on Vision Algorithms*, 1999.

[11] S. Baker et I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56:221–255, 2004.

[12] M. Cetin et A. Erar. Variable selection with Akaike information criteria: a comparative study. *Hacettepe Journal of Mathematics and Statistics*, 31:89–97, 2002.

[13] E. Ronchetti et R. Staudte. A robust version of Mallow's C_P. *Journal of the American Statistical Association*, 89:550–559, 1994.

[14] J. De Brabanter, K. Pelckmans, J. Suykens, J. Vandewalle, et B. De Moor. Robust cross-validation score functions with application to weighted least squares support vector machine function estimation. Rapport technique, Katholieke Universiteit Leuven, 2003.

[15] G. Wahba et S. Wold. A completely automatic French curve: fitting spline functions by cross-validation. *Communications in Statistics*, 4:1–17, 1975.

[16] G. Wahba. *Spline Models for Observational Data*. Society for Industrial and Applied Mathematics, 1990.

[17] M. Farenzena, A. Bartoli, et Y. Mezouar. Efficient camera smoothing in sequential structure-from-motion using approximate cross-validation. Dans *European Conference on Computer Vision*, 2008.

[18] P. Dierckx. *Curve and Surface Fitting with Splines*. Oxford University Press, 1993.

Approche hiérarchique pour un alignement musique-sur-partition efficace

Cyril Joder Slim Essid Gaël Richard

Institut TELECOM – TELECOM ParisTech, CNRS/LTCI

37, rue Dareau, 75014 Paris – FRANCE

{joder, essid, grichard}@telecom-paristech.fr

Résumé

Dans le cadre du problème d'alignement audio-sur-partition, nous utilisons un modèle à états cachés pour modéliser l'évolution du contenu du signal sonore en rapport avec la partition. Nous proposons dans cet article une méthode hiérarchique de réduction de l'espace de recherche pour un tel modèle. Nos expériences menées sur une base de 94 morceaux de musique pop montrent qu'avec cet algorithme, l'utilisation d'un descripteur détectant les attaques de notes permet d'obtenir une précision d'alignement supérieure à celle de l'algorithme de programmation dynamique (DTW), avec une complexité significativement moindre.

Mots clefs

Musique, Alignement, Modèle à états cachés.

1 Introduction

Nous nous intéressons au problème de l'alignement d'une partition musicale polyphonique avec un enregistrement audio de la même pièce. Nous traitons cette tâche par une stratégie "hors ligne", qui permet d'utiliser l'enregistrement dans son ensemble. Nous nous intéressons à un alignement au niveau *symbolique*, dont le résultat sera l'ensemble des positions dans l'enregistrement de chaque note de la partition. L'alignement audio-partition peut être utilisé pour l'indexation d'un morceau de musique par sa partition, l'analyse d'interprétation ou encore comme guide pour une séparation de source informée.

Alors que la plupart des systèmes de suivi de partition en temps-réel utilisent des modèles statistiques qui peuvent être assez élaborés [1, 2, 3, 4], les méthodes "hors-ligne" se contentent souvent de l'algorithme de programmation dynamique DTW (Dynamic Time Warping) ou de variantes [5, 6, 7]. Ces approches sont en général plus simples, et peuvent aussi être appliquées au problème de synchronisation audio-audio.

Cependant, la complexité (en temps et en espace) de la DTW est quadratique en le nombre de trames audio. Ce problème a été étudié dans [8], où une DTW "à court terme" est proposée pour réduire la complexité en espace au prix d'une augmentation de la complexité en temps. Dans [9], Müller *et al* utilisent une DTW "multi-échelle", où certains chemins sont supprimés de façon hiérarchique. Ce procédé diminue la complexité de l'algorithme mais ne garantit plus d'obtenir le chemin d'alignement optimal.

Avec l'algorithme DTW ou ses variantes, l'utilisation de plusieurs descripteurs de nature différente peut être malaisé. Aussi ces systèmes se limitent généralement à l'emploi de vecteurs de chroma. Une exception notable est [7], qui propose une stratégie pour combiner les distances locales issues de descripteurs de chroma ainsi que de détecteurs d'attaque de note. Un modèle statistique à états cachés rend plus naturelle la fusion de ces informations de types différents. Cette structure est souvent utilisée dans des systèmes temps-réel [10, 11] qui modélisent chaque descripteur par un mélange de Gaussiennes.

Le système à états cachés présenté ici exploite un modèle différent pour chaque type de descripteur : un modèle "d'histogramme" (voir 3.1) pour les vecteurs de chroma, et un modèle logistique (voir 3.2) pour la fonction de détection de transitoire. Ce système obtient une très bonne précision d'alignement avec une complexité très inférieure à la DTW.

Nous exploitons en outre une approche hiérarchique de réduction de complexité, qui opère un élagage de l'arbre des états possibles, de manière adaptée aux données. Cette approche, plus souple que celle utilisée dans [9], s'avère bénéfique en terme de complexité globale, sans nuire aux performances d'alignement.

Dans les sections suivantes sont présentés le cadre statistique et les modèles d'observation utilisés. Les trois systèmes d'alignement testés sont comparés dans la section 4. Notre méthode d'élagage hiérarchique pour un décodage approché de ces modèles est proposée en section 5, avant de suggérer quelques conclusions (section 6).

2 Modèle à états cachés pour l'alignement audio-sur-partition

La partition musicale indique les temps d'attaque et durée de chacune des notes, dans une échelle temporelle – le *tempo*– qui est inconnue et variable. Néanmoins, en négligeant les possibles petites erreurs de synchronisation entre les musiciens, cela nous permet d'effectuer une seg-

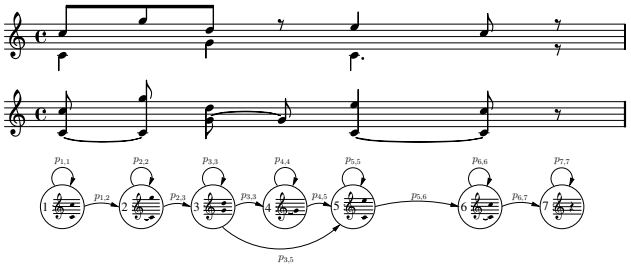


Figure 1 – *Représentations de la partition. Haut : forme graphique originale. Milieu : séquence d'accords. Bas : automate fini correspondant.*

mentation de la partition en *accords*, qui sont des ensembles de notes jouées simultanément (de façon similaire à [2]). À chaque fois qu'une note apparait ou s'éteint, un nouvel accord est créé. La partition est donc vue comme une séquence d'accords, caractérisés par les notes qu'ils contiennent. Cette segmentation en accord est représentée sur la figure 1 (haut et milieu).

Nous faisons ensuite l'hypothèse que les notes présentes à un instant de l'enregistrement dépendent uniquement de l'accord courant. Ainsi, il est possible d'utiliser un modèle à états cachés, dont les états sont les accords précédemment segmentés. Un automate fini est donc construit à partir de la partition, comme illustré sur la figure 1 (bas). La tâche d'alignement revient alors à trouver le chemin optimal (dans un sens explicité par la suite) dans l'automate correspondant à l'enregistrement.

Nous prenons le parti de ne pas utiliser les informations rythmiques de la partition, considérant que nous n'avons aucune connaissance *a priori* sur le tempo. Le critère d'optimalité utilisé est alors le maximum de vraisemblance. Soient $\mathbf{y}=y_1,\dots,y_N$ la séquence de descripteurs extraits du signal, et S_n la variable aléatoire décrivant l'état courant au temps n. Le chemin optimal $\hat{\mathbf{S}}$, calculé par l'algorithme de Viterbi, est :

$$\hat{\mathbf{S}} = \underset{\mathbf{S}\in\mathcal{S}}{\operatorname{argmax}}\, P\big(\mathbf{y}\big|\mathbf{S}\big) = \underset{\mathbf{S}\in\mathcal{S}}{\operatorname{argmax}} \prod_{n=1}^{N} P(y_n|S_n), \quad (1)$$

où $\mathcal{S}$ est l'ensemble des chemins acceptables. Nous considérons comme acceptables les chemins parcourant tous les accords dans le bon ordre.

3 Modèle d'observation

De façon similaire à [12], deux sortes d'information sont considérées ici : les hauteurs de notes et les informations d'attaque. Ainsi, deux types de descripteurs sont utilisés. Pour modéliser le contenu spectral du signal, nous utilisons des *vecteurs de chroma*, et le *flux spectral* est censé détecter les attaques de notes. Ces deux descripteurs sont extraits à une fréquence de 50 Hz.

3.1 Vecteurs de chroma

Un *vecteur de chroma* est un vecteur à douze dimensions, qui représentent la "puissance" de chaque classe chromatique de la gamme musicale tempérée (de do à si). Les vecteurs de chroma utilisés ici sont calculés par la méthode décrite dans [14]. Bien que ne prenant pas en compte l'information d'octave des note, les vecteurs de chroma fournissent une représentation compacte du contenu "harmonique" du signal, efficace pour l'alignement audio-sur-partition, comme observé dans [13].

Pour chaque accord, une loi de probabilité $\{\tilde{g}(i)\}_{i=1\dots12}$ sur les douze classes chromatiques est créée, par la superposition de lois élémentaires correspondant aux notes de cet état. Une loi élémentaire est une simple fonction de Kronecker $\{\delta(i,j)\}_{i=1\dots12}$ où j est la classe chromatique de la note considérée. Une composante constante est ajoutée pour modéliser le bruit, ce qui donne une loi g définie par $g(i) = (1-q)\tilde{g}(i) + \frac{q}{12}$. La valeur $q = 0,7$ a été trouvée satisfaisante. Par exemple, les valeurs de la loi correspondant à l'accord $\{\mathrm{do}_3, \mathrm{mi}_3, \mathrm{sol}_3, \mathrm{do}_4\}$ (représentées en vecteur) seront : $\frac{1-q}{4}(2,0,0,0,1,0,0,1,0,0,0,0) + \frac{q}{12}\mathbb{1}$.

La vraisemblance de chaque état est ensuite calculée grâce au modèle décrit dans [15]. La valeur d'un vecteur de chroma v extrait de l'audio est considérée comme l'histogramme d'un tirage aléatoire d'après la loi g, correspondant à un accord c. La probabilité d'obtenir cet histogramme est alors

$$p\big(v|c\big) = Z \prod_{i=1}^{12} g(i)^{\alpha v(i)}, \quad (2)$$

où Z dépend uniquement de l'observation et α est un paramètre d'échelle. La valeur de ce paramètre n'influant pas sur l'optimum cherché, sa valeur est fixée à 1.

3.2 Flux spectral

Pour prendre en compte les transitoires présents aux attaques de notes, nous utilisons le descripteur de *flux spectral*, dont l'efficacité pour la tâche de détection de tempo est illustrée dans [16]. Nous employons ce descripteur pour un détecteur de transitoires "probabilisé".

Tout d'abord, les valeurs du flux spectral sont normalisées afin que la valeur maximale soit 1. Un seuil local est alors calculé par un filtre d'ordre aux 67$^{\text{ème}}$ centile, sur une longueur de 200 ms. Notre fonction de détection d'attaque est alors obtenue en retranchant ce seuil au flux spectral. Enfin, la vraisemblance d'une attaque est calculée grâce à un simple modèle logistique. Soir A la variable aléatoire de Bernoulli représentant l'évènement "attaque". Pour une valeur f de notre fonction de détection, on a

$$p\big(A=1|f\big) = \frac{e^{bf}}{1+e^{bf}} \quad (3)$$

où b est un paramètre positif à déterminer, qui contrôle la "confiance" accordée au détecteur de transitoire : plus il est grand, plus la décision sera proche d'un détecteur déterministe (valeurs 0 ou 1).

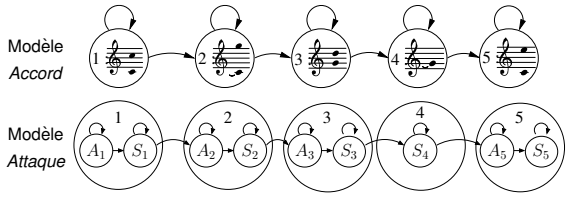


Figure 2 – *Structure des systèmes* Accord *et* Attaque *pour la même partition (*A *et* S *représentent respectivement* attaque *et* soutien*).*

4 Performance des systèmes d'alignement

4.1 Système *Accord* et système *Attaque*

Dans le cadre statistique présenté précédemment, deux structures de modèles sont utilisées. Dans la plus simple, qui constitue le système *Accord*, un accord est représenté par un unique état, quels que soient son contenu ou sa durée théorique. Seule les observations de chroma sont alors considérées. Le flux spectral n'est donc pas utilisé, et les vraisemblances des accords sont calculées d'après (2).
Le système *Attaque* est une modification du système précédent qui prend en compte les transitoires. Dans ce modèle, un "niveau de hiérarchie" inférieur est ajouté afin de pouvoir modéliser deux phases différentes à l'intérieur d'un accord : la phase d'*attaque* et la phase de *soutien*. Un accord comprenant au moins une attaque de note est donc représenté par deux états successifs, qui partagent le même modèle de chroma. Les descripteurs d'attaque et de chroma sont supposés indépendants. Un état S est alors un couple accord/phase (C, A), et sa vraisemblance est exprimée par

$$p\big(v, f|C, A\big) \propto p\big(v|C\big)p\big(A|f\big)$$

d'après les équations (2) et (3). Six valeurs différentes sont testées pour le paramètre b de l'équation (3) : 0 ; 0,1 ; 1 ; 10 ; 50 et 100.
Dans les deux cas, seules deux transitions sont autorisées à partir d'un état : vers lui-même ou vers l'état suivant. La figure 2 illustre les différences de structure entre les deux systèmes. Dans le système *Attaque*, le "bouclage" sur l'état d'attaque est utile pour modéliser une phase d'attaque (valeur élevée du flux spectral) durant plusieurs trames.

4.2 Système de référence : *Multi-scale Dynamic Time Warping*

Les modèles précédents sont comparés à un système utilisant l'algorithme multi-scale Dynamic Time Warping (MsDTW) [9, 17], qui ne correspond pas au cadre statistique présenté. L'algorithme MsDTW cherche l'alignement de coût minimal entre deux séquences, d'abord à un niveau grossier (avec une faible résolution temporelle des descripteurs), puis à un niveau plus fin. À chaque niveau de précision, l'alignement est calculé en explorant uniquement un voisinage de δ trames autour du chemin d'alignement du niveau supérieur.
Cet algorithme est utilisé pour synchroniser la séquence de vecteurs de chroma extraite de l'audio avec une séquence construite à partir de la partition (l'information d'attaque n'est pas prise en compte). Pour cela, on effectue une "pseudo-synthèse" de la partition, en associant à chaque accord un vecteur de chroma type. Les valeurs de vecteur-type sont les valeurs de la loi de probabilité vue en 3.1. Cette pseudo-synthèse est ensuite dilatée pour que sa durée soit égale à celle de l'audio.
Trois niveaux de précision sont utilisés pour l'algorithme MsDTW : le plus fin utilise les vecteurs de chroma initiaux, avec une résolution temporelle de 50 Hz. Les niveaux supérieurs utilisent ces descripteurs moyennés sur respectivement 10 trames (200 ms) et 50 trames (1 s), avec des résolutions respectives de 10 Hz et 2 Hz. La mesure de distance locale utilisée est la distance cosinus. La largeur du voisinage considéré pour la réduction hiérarchique de l'espace de recherche est fixée à $\delta = 1000$, dans le but de supprimer le moins de chemins possible, tout en conservant une complexité raisonnable.

4.3 Évaluation

Nous évaluons les performances des systèmes d'alignement grâce à une base de données de 94 chansons de 2 à 6 minutes, tirées de la base RWC-pop [18]. Ces morceaux sont polyphoniques, multi-instrumentaux et la plupart contient des percussions. Afin de réduire la taille des données, ces morceaux (initialement échantillonés à 44,1 kHz) sont sous-échantillonés à 16 kHz. La vérité-terrain est fournie par des fichiers MIDI synchronisés avec l'audio. Les partitions utilisées sont ces mêmes fichiers MIDI, dans lesquels plusieurs changements de tempo arbitraires ont été ajoutés.
Les scores sont mesurés par le taux de reconnaissance, défini comme la proportion des attaques de notes détectées correctement, dans un intervalle de 300 ms autour de l'instant d'attaque réel. Ce seuil de 300 ms est choisi égal à celui de l'évaluation MIREX'06 [19].
Les scores ainsi que la complexité moyenne sont compilés dans le tableau 1. La complexité est mesurée par le nombre de cellules (couples trame audio - état ou trame audio - trame de pseudo-synthèse) évaluées ramené au nombre de cellules nécessaires à l'algorithme DTW (le carré du nombre de trames audio).
Tout d'abord, on peut noter que la MsDTW obtient de meilleurs résultats que le système *Accord*. La raison en est que la MsDTW modélise implicitement les durées des notes dans la phase de "pseudo-synthèse", alors que le modèle statistique n'en tient pas compte. Cela augmente la précision, mais aussi la complexité (68,4% contre 16,2%).
Cependant, on voit que l'utilisation du flux spectral permet au système *Attaque* de surpasser la MsDTW, en conservant une complexité significativement moindre. En effet, un taux de reconnaissance de 87,2% est obtenu avec la valeur $b = 50$, contre 78,8% pour la MsDTW, alors que la complexité reste à 26,3%.

Système	Score	Complexité
MsDTW	**78,8**%	68,4%
Accord	64,5%	16,2%
Attaque ($b = 0$)	69,7%	26,3%
Attaque ($b = 0, 1$)	70,5%	
Attaque ($b = 1$)	73,1%	
Attaque ($b = 10$)	82,9%	
Attaque ($b = 50$)	**87,2**%	
Attaque ($b = 100$)	84,7%	

Tableau 1 – *Taux de reconnaissance et complexité moyenne (proportion de la complexité DTW) en fonction du système d'alignement.*

L'augmentation de la complexité peut être bénéfique à la précision de l'alignement. En effet, même avec la valeur $b = 0$ (le flux spectral n'est pas pris en compte), le taux de reconnaissance passe de 64,5% (système *Accord*) à 69,7% (système *Attaque*). Cela tient au fait que la plupart des accords sont alors représentés par deux états. Ainsi, la durée minimale passée dans chaque accord est de deux trames au lieu d'une. Cela évite au système de passer très rapidement d'un état à un autre état éloigné, et conduit à un chemin d'alignement plus régulier.

5 Approche hiérarchique pour un décodage approché

5.1 Principe

Afin d'accélérer encore la phase de décodage du modèle statistique, nous présentons ici une approche hiérarchique de réduction de complexité inspirée de l'algorithme MsDTW. Comme dans cette méthode, l'idée est d'effectuer un alignement d'abord à un niveau grossier, puis d'affiner ce résultat en considérant uniquement le voisinage du chemin d'alignement obtenu.

Pour ces alignement grossiers, nous tirons parti de structures musicales plus longues que les accords, à savoir les *temps* (ou *beats*, que nous emploierons pour éviter des confusions) et les *mesures*. Pour chacun de ces niveaux, on peut construire un automate, dont les états correspondent respectivement aux beats et aux mesures de la partition. Ces automates forment des modèles à états cachés dont le décodage fournira un alignement. Puisque les unités temporelles considérées sont plus grandes, on utilise des descripteurs calculés sur des fenêtres plus longues et ayant une résolution temporelle plus faible. La figure 3 illustre la construction des automates et le calcul des observations aux trois niveaux de hiérarchie utilisés.

L'algorithme se déroule comme suit : on calcule le chemin optimal $\hat{\mathbf{S}} = \hat{S}_1, \ldots, \hat{S}_N$ dans l'automate de plus haut niveau. Une passe "retour" est ajoutée à l'algorithme de Viterbi, pour calculer

$$\tilde{P}(n, s) = \max_{\mathbf{S} \in \mathcal{S}, \hat{S}_n = s} \{P(\mathbf{y}|\mathbf{S})\}$$

pour tout état s et tout instant n. $\mathcal{S}$ est l'ensemble des chemins acceptables et $\mathbf{y}$ est la séquence d'observations. Cette valeur est la vraisemblance du meilleur chemin passant par l'état s à l'instant n.

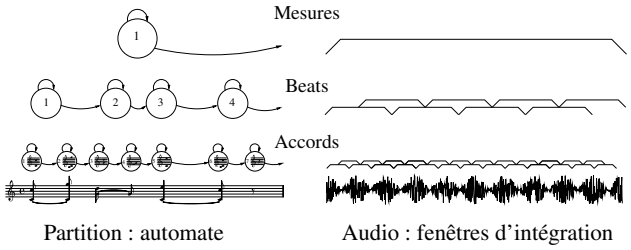


Figure 3 – *Automates finis (modélisant la partition) et fenêtres d'intégration (sur lesquelles sont calculées les observations) aux trois niveaux de hiérarchie considérés.*

La structure de l'automate est gauche-droite, on peut donc définir un ordre total sur ses états : $s \leq s'$ ssi il existe un chemin de s vers s'. On calcule alors, pour chaque instant n, les "états admissibles les plus lointains" S_n^- et S_n^+, définis par :

$$S_n^- = \min \{s \,/\, \tilde{P}(n, s) \geq \frac{P(\mathbf{y}|\hat{\mathbf{S}})}{\eta}\}$$

$$S_n^+ = \max \{s \,/\, \tilde{P}(n, s) \geq \frac{P(\mathbf{y}|\hat{\mathbf{S}})}{\eta}\}$$

où η est un paramètre contrôlant la vraisemblance minimale considérée au niveau plus bas. On définit alors les *rayons de tolérance* δ_- et δ_+ comme le maximum (pour n dans $\{1, \ldots, N\}$) du nombre d'états séparant respectivement S_n^- de $\hat{S}_n$ et $\hat{S}_n$ de S_n^+.

Ces rayons de tolérance définissent un ensemble d'états possibles autour du chemin optimal, qui est utilisé pour réduire l'espace de recherche au niveau inférieur. Un alignement au niveau inférieur est alors calculé, en explorant uniquement le domaine défini précédemment. La figure 4 illustre ce processus de réduction de l'espace de recherche.

La séquence de descripteurs audio correspondant aux niveaux supérieurs est constituée de versions intégrées (moyennées) des vecteurs de chroma initiaux, avec une résolution temporelle réduite. Le flux spectral n'est pas pris en compte à ces niveaux. Les durées d'intégration sont choisies en fonction des tempos les plus rapides acceptables. Pour le niveau beat, cette durée est de 200 ms, correspondant à un tempo très rapide de 300 bpm. Pour le niveau mesure, la fenêtre d'intégration est de 1 s, soit une mesure à 4 temps à un tempo de 240 bpm. Une recouvrement de moitié est utilisé, ce qui donne des résolutions temporelles respectives de 10 Hz et 2 Hz. Le même modèle d'observation vu en 2 est utilisé. Pour un état (beat ou mesure) la loi g est la superposition des lois associées aux accords que contient cet état, pondérées par leurs durées.

La grande différence entre cette approche et la méthode à l'origine de la MsDTW est que les rayons de tolérance δ ne sont pas des paramètres, mais sont calculés à partir des données de façon adaptative, contrôlé par le paramètre η. Il

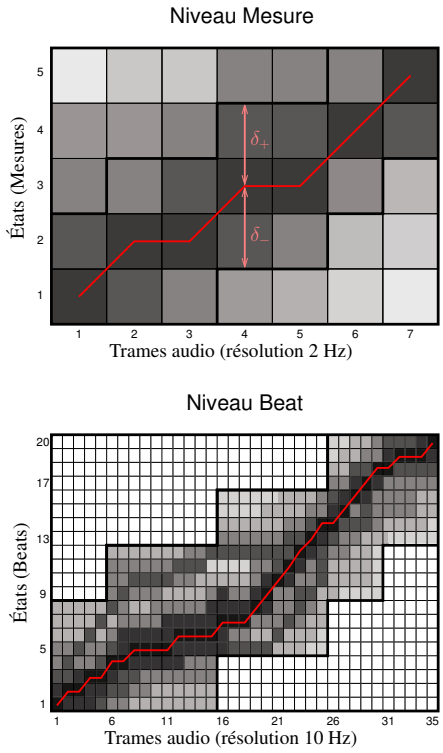


Figure 4 – *Principe de la méthode d'élagage hiérarchique. Le niveau de gris d'une cellule temps-état correspond à la vraisemblance maximale des chemins passant par cette cellule. Au niveau* beats*, seul le domaine grisé est exploré.*

est souvent plus avantageux de régler la tolérance en terme de score (le paramètre η) plutôt qu'en terme de déviation du chemin d'alignement (paramètres δ). En effet, il est possible qu'un mauvais chemin obtienne un score légèrement supérieur à celui du "vrai" chemin d'alignement à un niveau grossier, par exemple s'il suit une différente répétition d'une phrase musicale. Si ce chemin est trop éloigné du "vrai" alignement, ce dernier pourra être supprimé si on considère un rayon de tolérance fixe. En revanche, on peut supposer que le "vrai" chemin a un score élevé, et qu'il n'est pas supprimé avec notre méthode.

5.2 Expériences

Nous testons cette approche hiérarchique sur la base de données déjà présentée. Le système de bas-niveau utilisé est le système *Attaque* avec $b = 50$. Plusieurs valeurs du paramètre η sont testées et les principaux résultats sont présentés dans le tableau 2. La complexité est présentée en fraction du nombre de cellules explorées en rapport au nombre de cellules de l'algorithme DTW. La complexité au niveau mesure est égale à 0,16% pour la MsDTW, et à 0,04% pour tous les autres systèmes. Le temps d'exécution total et le nombre d'"erreurs d'élagage" sont présentés. Une erreur d'élagage se produit si une partie du "vrai" chemin d'alignement se trouve supprimée dans le processus de réduction de l'espace de recherche. Notre implémentation de l'algorithme est en MATLAB, et a été exécuté sur un

Système	Complexité		Temps	Erreurs
	Beats	Accords	d'exécution	(nb)
MsDTW δ=150	2,24%	14,02%	1180 s	0
Attaque b=50	–	26,26%	482 s	0
$\delta = 60$	0,81%	7,93%	362 s	0
$\eta = 1000$	0,42%	4,53%	300 s	0
$\eta = 200$	0,35%	4,07%	276 s	0
$\eta = 100$	0,33%	3,82%	265 s	0
$\eta = 50$	0,30%	3,59%	256 s	0
$\eta = 20$	0,26%	3,22%	240 s	0
$\eta = 10$	0,23%	2,97%	229 s	2
$\eta = 5$	0,19%	2,59%	215 s	2

Tableau 2 – *Performance de notre implémentation du système d'alignement, avec différents paramètres pour l'algorithme hiérarchique de réduction de complexité. Les erreurs comptent le nombre de morceaux pour lesquels le "vrai" chemin est supprimé au cours de l'élagage.*

Intel Core2, 2,66 GHz avec 3,6 Go de RAM, sous linux.
Trois systèmes de références sont considérés. Le premier utilise l'algorithme MsDTW avec le paramètre $\delta = 150$, qui est la valeur minimale n'entrainant aucune erreur d'élagage sur notre base de données. Le second système utilise le modèle *Attaque* sans élagage. Le dernier effectue un élagage hiérarchique, mais avec des rayons de tolérance constant $\delta_- = \delta_+ = 60$. Cette valeur est la plus basse pour laquelle aucune erreur n'est comptée.
En terme de précision d'alignement, tous les systèmes qui ne font pas d'erreurs d'élagage obtiennent le même score que le système de référence (87.16%). Ainsi, cette méthode approchée de décodage n'affecte pas les performances d'alignement.
Les résultats montrent l'avantage de cette approche, car la complexité et le temps d'exécution de tous les systèmes l'utilisant sont inférieurs à celle de la méthode de référence (sans élagage). Comme prévu, la complexité diminue avec la valeur de η. Aucune erreur d'élagage n'est à déplorer jusqu'à la valeur $\eta = 20$, dont le temps d'exécution correspondant est la moitié de celui du système de référence (240 s contre 484 s).
Le gain de cette méthode par rapport à l'utilisation d'un rayon de tolérance δ fixe est visible. En effet, le système utilisant un rayon fixe optimal $\delta = 60$ s'exécute en 362 s et présente une complexité en espace supérieure à notre stratégie d'élagage "adaptative".
Sur la figure 5 sont représentés les complexités en espace de trois systèmes d'alignement : sans élagage, avec un rayon fixe $\delta = 60$ et avec $\eta = 20$. Les deux stratégies d'élagage entrainent une réduction significative de la complexité sur tous les morceaux. De plus, il est visible que la taille de l'espace de recherche obtenu avec notre stratégie peut fortement varier suivant les morceaux, alors qu'elle est à peu près constante avec un rayon δ fixe (les variations sont dues aux différents nombres de notes par beat). Ces variations ne sont pas corrélées avec le nombre d'états initial du modèle, indiquant que notre approche adapte le processus d'élagage aux données. Ainsi, alors que dans cer-

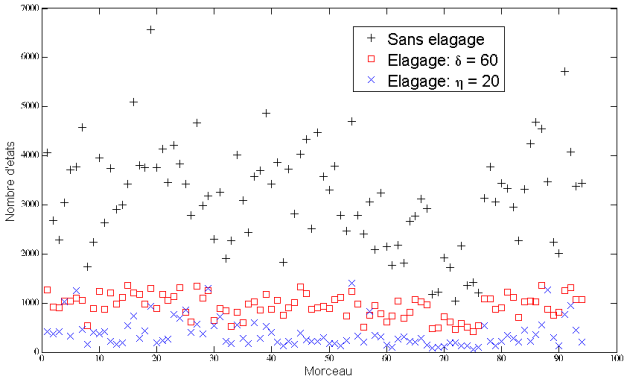


Figure 5 – *Nombre d'états explorés par trame, au cours l'alignement au plus bas niveau en fonction du morceau traité.*

tains cas la complexité de notre méthode est supérieure à celle d'un rayon fixe, elle est significativement plus faible pour la majorité des morceaux.

6 Conclusion

Dans cet article, nous montrons qu'une approche hiérarchique pour le décodage approché d'un modèle à états cachés peut fournir un alignement d'une très bonne précision pour une faible complexité. Nos expériences indiquent que les taux de reconnaissance sont supérieurs à ceux obtenus avec un système par DTW, quand une description des attaques de notes est utilisée en plus des vecteurs de chroma, et cela en conservant une complexité inférieure pour la phase de décodage.

La méthode hiérarchique d'élagage de l'arbre de recherche réduit encore la complexité, sans affecter la précision du système. L'avantage de notre stratégie par rapport à celle utilisée dans [9, 17] est que les rayons de tolérance peuvent s'adapter aux données, ce qui conduit à une meilleure efficacité globale.

La suite de ces travaux sera consacrée à l'utilisation de modèles plus élaborés pour le niveau le plus bas, dont l'utilisation est rendue abordable grâce à la réduction de complexité engendrée par l'élagage. Nous tenterons aussi de réduire encore le nombre d'états du modèle statistique en tirant parti des répétitions dans la structure musicale.

Références

[1] Lorin Grubb et Richard Dannenberg. A stochastic method of tracking a vocal performer. Dans *Proc. of ICMC*, 1997.

[2] Christopher Raphael. A probabilistic expert system for automatic musical accompaniment. *Journal of Computational and Graphical Statistics*, 10, 2001.

[3] Diemo Schwarz, Nicola Orio, et Norbert Schnell. Robust polyphonic midi score following with hidden markov models. Dans *Proc. of ICMC*, 2004.

[4] Arshia Cont. A coupled duration-focused architecture for realtime music to score alignment. *IEEE Trans. on Pattern Analysis and Machine Intelligence on*, 32(6) :974–987, June 2010.

[5] Ning Hu, Roger B. Dannenberg, et George Tzanetakis. Polyphonic audio matching and alignment for music retrieval. Dans *Proc. of IEEE WASPAA*, 2003.

[6] Christian Fremerey, Michael Clausen, Sebastian Ewert, et Meinard Müller. Sheet music-to-audio identification. Dans *Proc. of ISMIR*, 2009.

[7] Sebastian Ewert, Meinard Müller, et Peter Grosche. High resolution audio synchronization using chroma onset features. Dans *Proc. of IEEE ICASSP*, 2009.

[8] Hagen Kaprykowsky et Xavier Rodet. Globally optimal short-time dynamic time warping : Application to score to audio alignment. Dans *Proc. of IEEE ICASSP*, 2006.

[9] Meinard Müller, Henning Mates, et Frank Kurth. An efficient multiscale approach to audio synchronization. Dans *Proc. of ISMIR*, 2006.

[10] Pedro Cano, Alex Loscos, et Bonada. Score-performance matching using hmms. Dans *Proc. of ICMC*, 1999.

[11] Arshia Cont, Diemo Schwarz, et Norbert Schnell. Training ircam's score follower. Dans *Proc. of IEEE ICASSP*, 2005.

[12] Nicola Orio et Diemo Schwarz. Alignment of monophonic and polyphonic music to a score. Dans *Proc. of ICMC*, 2001.

[13] Cyril Joder, Slim Essid, et Gaël Richard. A comparative study of tonal acoustic features for a symbolic level music-to-score alignment. Dans *Proc. of IEEE ICASSP*, 2010.

[14] Yongwei Zhu et M.S. Kankanhalli. Precise pitch profile feature extraction from musical audio for key detection. *IEEE Trans. on Multimedia*, 8(3) :575–584, June 2006.

[15] Christopher Raphael. Aligning music audio with symbolic scores using a hybrid graphical model. *Machine Learning Journal*, 2006.

[16] Miguel Alonso, Gaël Richard, et Bertrand David. Extracting note onsets from musical recordings. Dans *Proc. of ICME*, 2005.

[17] Stan Salvador et Philip Chan. Fastdtw : Toward accurate dynamic time warping in linear time and space. Dans *KDD Workshop on Mining Temporal and Sequential Data*, pages 70–80, 2004.

[18] M. Goto. Rwc music database : Popular, classical, and jazz music databases, 2002.

[19] Music information retrieval evaluation exchange 2006, score following task : `http://www.music-ir.org/mirex/2006/index.php/Score_Following_Proposal`.

Compression sans perte et presque sans perte d'images médicales à l'aide d'un prédicteur hiérarchique orienté et adaptatif

J. Taquet[1] C. Labit[1]

[1] INRIA, Centre Inria Rennes Bretagne Atlantique

IRISA, Campus de Beaulieu,
35042 Rennes cedex – FRANCE

{Jonathan.Taquet, Claude.Labit}@inria.fr

Résumé

Cet article présente une nouvelle approche pour la compression sans perte et presque sans perte d'images médicales. Elle combine l'adaptativité des schémas DPCM avec un prédicteur hiérarchique orienté (HOP) afin de proposer une compression efficace tout en permettant une transmission progressive en résolution. Nous obtenons des taux de compression sans perte qui sont de l'ordre de 4% supérieurs à ceux du standard JPEG2000 (également progressif en résolution) et proches de ceux de CALIC (non progressif) sur une base d'images conséquente. L'algorithme HOP est également bien adapté pour la compression presque sans perte, proposant un compromis débit/distorsions intéressant en comparaison au standard JPEG-LS, et un PSNR similaire voir meilleur que JPEG2000 sur des images médicales natives (bruitées).

Mots clefs

Compression, sans perte, presque sans perte, prédiction hiérarchique, images médicales.

1 Introduction

L'utilisation de l'imagerie médicale [1] s'est largement accrue ces dernières années et surtout en ce qui concerne l'imagerie par résonance magnétique (IRM) et la tomographie calculée (TC). Ces deux techniques d'acquisition permettent de produire des images volumiques pouvant être vues comme des séquences d'images bidimensionnelles (coupes). Ces images, assez volumineuses, nécessitent des techniques de compression efficaces pour leur archivage à long terme (pouvant s'étendre à plus de 30 ans) mais également intéressantes pour leur consultation au travers de réseaux. L'une des particularité souvent requise afin de satisfaire des contraintes juridiques et éthiques est la compression sans perte, même si parfois une compression presque sans perte pourrait être suffisante.

Cet article se concentrera sur les techniques 2D ne nécessitant pas une optimisation, gourmandes en temps de calcul, de paramètres optimaux pour chaque image. Les algorithmes de compression sans perte qui ont su se montrer les plus efficaces suivent un schéma DPCM. Ils effectuent une décorrélation prédictive, ligne par ligne et colonne par colonne, de chaque pixel à l'aide d'approches adaptatives exploitant l'information causale. Suivant ce schéma, LOCO-I [2] utilisé par le standard JPEG-LS et CALIC [3] sont deux algorithmes de référence. Bien qu'efficaces, ces codeurs ne permettent pas une transmission progressive, fonctionnalité importante pour la navigation et la consultation d'images médicales à distance. Cette progressivité en qualité ou en résolution est permise par des approches effectuant une prédiction hiérarchique par interpolation (HIP : *hierarchical interpolative prediction*) telles que HINT [4, 5] ou encore la transformée en ondelettes entières (IWT : *integer wavelet transform*) intégrée au standard JPEG2000 [6] pour son mode sans perte.

La IWT est souvent effectuée à l'aide de schémas de lifting entier [7] se déroulant généralement en trois étapes : 1) la séparation de l'ensemble des pixels en deux sous-bandes L et H (par sous échantillonnage), 2) la prédiction des pixels de H à l'aide de ceux de L (H contient alors les coefficients hautes fréquences) et 3) la mise à jour des pixels de L à l'aide des résidus de H (L contient alors les coefficients basses fréquences). La HIP peut se résumer par un schéma de lifting simplifié aux étapes 1) et 2). Afin d'obtenir une progressivité en résolution ou en qualité, les données résiduelles peuvent être compressées à l'aide de codeurs comme EBCOT [8] ou SPIHT [9]. Les approches utilisant la IWT ont l'avantage d'avoir un meilleur rapport débit/PSNR que la HIP et de proposer une représentation multi-résolution anti-crénelée, mais ont des taux de compression sans perte légèrement inférieurs. La HIP permet également une compression presque sans-perte plus aisée.

Nous proposons de combiner les approches prédictives DPCM (non hiérarchiques, orientées, adaptatives) et HIP (hiérarchiques, non orientées, non adaptatives) usuelles afin d'en définir une nouvelle, intitulée approche prédictive hiérarchique orientée adaptative (HOP : *hierarchical oriented prediction*). Cette contribution permet une représentation progressive en résolution seulement,

mais exploite l'information des pixels déjà connus dans une même sous-bande afin d'améliorer la décorrélation en comparaison à la IWT et la HIP. L'approche sans perte de notre contribution sera présentée dans la section 2, son extension presque sans perte dans la section 3 et le codage entropique des résidus dans la section 4. Enfin les résultats expérimentaux seront commentés dans la section 5.

2 Prédicteur hiérarchique orienté

Comme pour les approches HIP, la HOP s'effectue échelle par échelle. Pour chaque niveau de prédiction, elle se déroule en deux étapes E_H et E_V (cf. figure 1). La première permet de prédire séquentiellement (ligne par ligne, colonne par colonne) les pixels horizontalement impairs (H) avec l'aide des pixels déjà connus : les pixels pairs (L), comme pour la HIP, mais également ceux précédemment prédits dans H. Cette décomposition permet d'obtenir une image sous échantillonnée horizontalement (L) ainsi que les résidus de prédiction (H). La seconde étape est la transposée mathématique de E_H, appliquée sur L afin d'obtenir une image basse résolution LL et les résidus des pixels verticalement impairs LH. Le schéma dyadique visant à décomposer également H en deux bandes HL et HH n'a pas été retenu, les résidus de H étant suffisamment décorrélés.

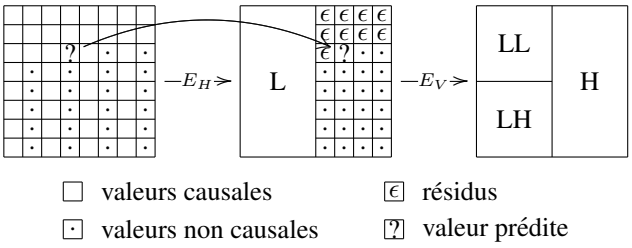


Figure 1 – *Approche prédictive hiérarchique proposée*

Par la suite, seule le prédicteur de l'étape E_H sera présenté, sa transposition pour E_V étant immédiate.

2.1 Définition du prédicteur

Le prédicteur HOP a principalement été conçu pour des images bruitées contenant des objets peu texturés et aux contours assez fortement marqués. Il utilise le motif contextuel de la figure 2 afin de s'adapter au contenu local de l'image.

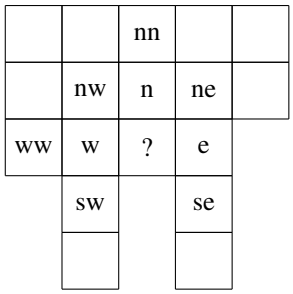


Figure 2 – *Motif contextuel de prédiction*

Ce nouveau type de motif de prédiction étend le modèle DPCM classique (lignes précédentes et pixels précédents de la même ligne que le pixel à prédire) en le fusionnant avec un motif HIP (colonnes paires).
En s'inspirant du prédicteur de CALIC qui a fait ses preuves, et en utilisant le motif contextuel étendu, une estimation de l'orientation des contours (ou de la texture) est effectuée à l'aide d'une estimation du gradient selon 4 orientations :

$$d_{\mathrm{k}} = \frac{1}{\operatorname{Card} D_k} \sum_{(x_i,x_j)\in D_k} \frac{|I(x_i) - I(x_j)|}{\|x_i - x_j\|}, \qquad (1)$$

avec I l'image, $k \in \left\{\mathrm{h}, \mathrm{v}, \frac{\pi}{4}, \frac{3\pi}{4}\right\}$ et D_k l'ensemble des pixels reliés dans la figure 3.

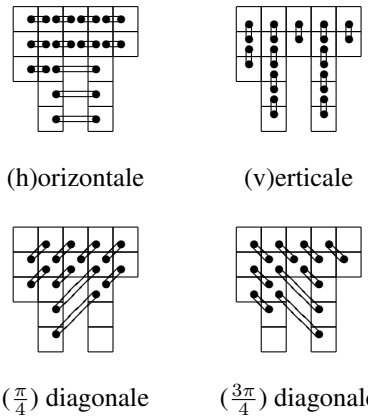


Figure 3 – *Pixels utilisés pour l'estimation du gradient*

L'orientation avec le gradient minimal :

$$o = \begin{cases} \underset{x\in\{\frac{\pi}{4};\frac{3\pi}{4}\}}{\operatorname{argmin}} & d_x, \quad \text{si } \left|d_{\frac{\pi}{4}} - d_{\frac{3\pi}{4}}\right| > |d_{\mathrm{v}} - d_{\mathrm{h}}|, \\ \underset{x\in\{\mathrm{v};\mathrm{h}\}}{\operatorname{argmin}} & d_x, \quad \text{sinon}, \end{cases} \qquad (2)$$

permet d'estimer l'activité d'un contour par :

$$G_o = \begin{cases} \left|d_{\frac{\pi}{4}} - d_{\frac{3\pi}{4}}\right|, & \text{si } o \in \{\frac{\pi}{4}; \frac{3\pi}{4}\}, \\ |d_{\mathrm{v}} - d_{\mathrm{h}}|, & \text{si } o \in \{\mathrm{v}; \mathrm{h}\}. \end{cases} \qquad (3)$$

Lorsque G_o est inférieur à un seuil de bruit $\mathrm{S}_{\text{Bruit}}$ (estimé à l'aide de la formule de Donoho [10] sur les plus hautes fréquences de la transformée orthonormale de Haar), la prédiction est non orientée et se comporte comme un filtre passe bas : $\hat{x} = \frac{\mathrm{n}+\mathrm{w}+\mathrm{e}+\frac{1}{\sqrt{2}}(\mathrm{nw}+\mathrm{ne}+\mathrm{sw}+\mathrm{se})}{3+\frac{4}{\sqrt{2}}}$. Pour cet article, lorsque $o \in \{\mathrm{h}, \frac{\pi}{4}, \frac{3\pi}{4}\}$, une prédiction par interpolation linéaire selon cette orientation est effectuée et lorsque $o = \mathrm{v}$, une prédiction linéaire ($\hat{x} = \mathrm{n}$) est retenue.

2.2 Adaptativité du prédicteur

Selon certaines configurations du contexte de prédiction, des erreurs systématiques (également appelées biais) ont

tendance à survenir. Prenant en compte le voisinage causal v (pouvant inclure la valeur de prédiction $\hat{x}$), une technique usuelle consiste à corriger la prédiction pour un contexte $C(v)$, à l'aide de la moyenne $\mu^{\epsilon}(C(v))$ des erreurs précédemment obtenues après prédiction dans un contexte identique :

$$\hat{x} \leftarrow \hat{x} + \mu^{\epsilon}(C(v)). \tag{4}$$

Cette approche peut également être étendue en utilisant une moyenne pondérée des erreurs survenant dans différents contextes [11] :

$$\hat{x} \leftarrow \hat{x} + \sum_k \alpha_k \mu_k^{\epsilon}(C_k(v)). \tag{5}$$

Comme la plupart des techniques de décorrélation orientées, la HOP est peu efficace dans les régions texturées. L'idée utilisée dans CALIC afin de capturer la texture a donc été retenue et étendue au schéma de prédiction hiérarchique. Elle consiste à binariser une information texturelle $\{t_0, \ldots, t_4\} = \{\mathrm{n}, \mathrm{w}, \mathrm{e}, 2\mathrm{n}-\mathrm{nn}, 2\mathrm{w}-\mathrm{ww}\}$ avec l'aide de la valeur de prédiction : $b_k = (t_k \geq \hat{x})$, pour former un nombre sur 5 bits $B = b_4 b_3 \ldots b_0$. L'implémentation de la HOP pour cet article utilise 3 contextes de correction. Chacun d'entre eux combine B avec le numéro du prédicteur sélectionné (0.bruit, 1.horizontal, 2.vertical, 3.diagonal $\frac{\pi}{4}$ ou 4.diagonal $\frac{3\pi}{4}$) et avec une quantification de G_o différente pour chaque contexte.

Lorsque la correction est appliquée séquentiellement :

$$\hat{x}_k \leftarrow \hat{x}_{k-1} + \alpha_k \mu_k^{\epsilon}(C_k(v, \hat{x}_{k-1})), \tag{6}$$

avec $\hat{x}_0 = \hat{x}$, le motif texturel binaire B est raffiné après chaque étape de réduction du biais. Cette nouvelle extension de (5) permet d'améliorer la sélection des contextes multiples lorsqu'ils dépendent de la valeur de prédiction, et d'obtenir ainsi une correction légèrement meilleure.

3 Compression presque sans perte

Le terme presque sans perte est couramment employé pour qualifier une image $\tilde{I}(N \times M)$ obtenue à l'aide d'un algorithme de compression avec pertes contraint par une erreur absolue maximale (PAE : *peak absolute error*) :

$$\mathrm{PAE} = \max_k \left\| I(k) - \tilde{I}(k) \right\| = \left\| I - \tilde{I} \right\|_{\infty}. \tag{7}$$

En compression prédictive, l'approche majoritairement retenue consiste a toujours effectuer la prédiction à partir des valeurs causales reconstruites presque sans perte. L'erreur de prédiction est alors généralement quantifiée à l'aide d'un quantificateur scalaire uniforme Q_δ :

$$\epsilon = Q_\delta(x - \hat{x}) = \mathrm{sign}(x - \hat{x}) \left\lfloor \frac{|x - \hat{x}| + \delta}{2\delta + 1} \right\rfloor, \tag{8}$$

δ étant égal au PAE souhaité.

LOCO-I a adopté une approche légèrement différente, favorisant un codage RLE tant que celui-ci satisfait le PAE et utilisant (8) sinon. Lorsque le PAE est important cette approche génère des artefacts visuellement dérangeants, mais permet une compression efficace sans utiliser un codeur arithmétique.
Pour les approches par transformée, une simple compression avec pertes de l'image, couplée à la compression de l'image d'erreur quantifiée : $Q_\delta(I - \tilde{I})$, s'avère être l'une des techniques les plus efficaces [12, 13].

La HOP presque sans perte suit l'approche prédictive habituelle : comme pour les HIP, une quantification de l'image à la plus basse résolution est appliquée avant de prédire les bandes de plus hautes résolution, et comme pour les DPCM, dans chaque sous bande les résidus sont quantifiés et permettent de générer les valeurs presque sans perte à utiliser pour la prédiction des pixels suivants.

4 Codage du résidu

Les résidus de la HOP sont compressés résolution par résolution, en commençant par la plus basse, ligne par ligne et colonne par colonne. Durant la prédiction, et comme souvent utilisé en compression prédictive afin de réduire la taille de l'alphabet à utiliser d'un facteur 2, les valeurs résiduelles sont réorganisées dans l'intervalle $[0, Q_\delta(I_{max} - \hat{x}) - Q_\delta(I_{min} - \hat{x})]$ afin d'obtenir un alphabet avec une distribution de probabilité quasi-décroissante. Un contexte d'estimation du modèle entropique est sélectionné à l'aide d'une quantification logarithmique de l'énergie résiduelle des pixels causaux voisins de la même sous-bande et des voisins hiérarchiques de la sous-bande de résolution inférieure. Un codeur adaptatif basé sur un codage arithmétique est alors utilisé pour compresser l'information.

Comme la taille de l'alphabet des symboles à compresser est encore potentiellement importante, une compression en deux phases a été retenue pour chacun des contextes. La première phase effectue le codage à l'aide d'un modèle adaptatif possédant peu de paramètres. Chaque symbole s est décomposé en $m = s \bmod 2^k$ et $d = s/2^k$, avec k le nombre de bits nécessaires pour représenter $Q_\delta(\mathrm{T_{Noise}})$. Cette décomposition similaire à la représentation des codes golomb-rice est compressée par $k+1$ codeurs arithmétiques binaires adaptatifs : un pour chaque bit de m, et un pour le codage de la représentation unaire de d (séquence de d bits à 0 suivie d'un bit d'arrêt à 1).
Durant la première phase, les statistiques d'un alphabet de taille réduite et adaptative, incluant un symbole d'échappement, sont apprises pour chaque contexte. Lorsque les symboles les plus fréquents, pour un contexte donné, sont suffisamment appris (en comparant le nombre d'occurrences des symboles de l'alphabet a celui du symbole d'échappement), la permutation pour la seconde phase est effectuée. Désormais, les symboles se présentant au

Bases	TC			IRM		
	V	N	res. X,Y	V	N	res. X,Y
CIPR (8 bits)	4	557	256,256	4	251	256,256
MeDEISA	5	1050	512,512	8	777	256,256
				2	58	512,512
VHP-Female	1	1734	512,512	18	923	256,256
VHP-Male	2	2395	512,512	18	973	256,256
VHP-Harvard	1	463	512,512	3	228	256,256
Harvard-3D	-	-	-	1	229	512,512

Tableau 1 – *Composition des bases d'images : V pour nombre de volumes, N pour nombre total de coupes.*

codeur du contexte seront compressés à l'aide d'un codeur arithmétique adaptatif n-aire guidé par les statistiques de l'alphabet précédemment appris. Lorsqu'un symbole non présent dans celui-ci est rencontré, le symbole d'échappement est transmis, et la valeur du symbole original est codée à l'aide du codeur de la première phase.

Cette compression en deux phases permet une adaptation rapide pour un codage sous-optimal des résidus des petites résolutions (pour les plus petites, la seconde phase n'a pas le temps de survenir), et un codage plus optimal lorsque la seconde phase survient.

5 Résultats expérimentaux

Les expérimentations ont été menées sur un ensemble conséquent d'images comportant plus de 6000 coupes tomographiques et 3000 coupes IRM. Il inclut des bases d'images 12 bits : les TC et IRM natives de NLM-VHP (National Library of Medicine's Visible Human Project)[1], contenant les acquisitions de deux corps complets (VHP-Male/Female) et d'une tête (VHP-Harvard) ; MeDEISA (Medical Database for the Evaluation of Image and Signal Processing Algorithms)[2] ; ainsi que la base 8 bits du Mallinckrodt Institute of Radiology, Image Proc. Lab., disponible au CIPR[3], qui est souvent utilisée comme référence pour les codeurs volumiques (bien que ces coupes 8 bits soient beaucoup moins bruitées que des images natives 12 bits). Quelques détails sur la composition de ces bases sont donnés dans le tableau 1.

5.1 Compression sans perte

Dans le tableau 2, les résultats de la compression sans perte par HOP sont comparés à ceux des standards JPEG-LS (JLS, prédictif, non progressif) et JPEG2000 (J2K, ondelettes, progressif en qualité et/ou résolution) ainsi qu'à ceux obtenus à l'aide des logiciels de référence de SPIHT (ondelettes, progressif en qualité) et CALIC (prédictif, non progressif). Toutes bases confondues, CALIC montre toujours son efficacité. HOP propose des taux de compression intéressants, et en particulier sur les IRM pour lesquelles il est équivalent voir plus performant que CALIC

[1]http ://www.nlm.nih.gov/research/visible/visible_human.html
[2]http ://www.medeisa.net
[3]http ://www.cipr.rpi.edu/resource/sequences/sequence01.html

		Débits (bpp)				
		non progressifs		progressifs		
	Bases	CALIC	JLS	SPIHT	J2K	HOP
TC	CIPR (8bits)	**1.92**	1.96	2.19	2.17	**2.06**
	MeDEISA	**4.72**	4.87	**4.74**	4.84	4.99
	VHP-Female	**4.65**	4.73	4.87	4.91	**4.73**
	VHP-Male	**4.75**	4.82	4.97	5.01	**4.83**
	VHP-Harvard	**5.03**	5.05	5.28	5.35	**5.18**
IRM	CIPR (8bits)	**2.69**	2.79	2.82	2.96	**2.72**
	MeDEISA	**3.15**	3.29	3.24	3.33	**3.11**
	VHP-Female	**4.41**	4.62	4.50	4.61	**4.42**
	VHP-Male	**4.90**	5.08	4.91	5.01	**4.80**
	VHP-Harvard	**4.68**	4.85	4.96	5.05	**4.67**
	Harvard-3D	**3.70**	4.13	**3.58**	3.82	3.78

Tableau 2 – *Débits moyens obtenus lors de la compression sans perte. Les meilleurs résultats sont soulignés et ceux dans chacune des catégories (progressifs / non progressifs) sont en gras.*

avec l'avantage non négligeable de proposer une progressivité en résolution. Sur les TC, il apporte des résultats généralement meilleurs que les deux codeurs ondelettes avec une exception sur les TC de la base MeDEISA. Ces dernières on subit un filtrage, elle contiennent très peu de hautes fréquences et ne sont donc pas natives. Cette constatation révèle la supériorité de la décorrélation des ondelettes sur des images relativement douces et peu bruités en comparaison à la HOP. Ainsi lorsque des images ont subi un filtrage, impliquant des contours moins marqués et un bruit plus faible, des codeurs de type ondelettes peuvent se révéler très efficace et parfois surpasser les approches prédictives. Des résultats atypiques ont également été obtenus sur une IRM de base VHP-Harvard. Celle-ci, reconstruite de manière volumique (identifiée par Harvard-3D), contient un bruit texturé (effet de Gibbs) important qui peut favoriser les ondelettes et semble être particulièrement bien capté par SPIHT.
Comme la compression sans perte prend plus de sens sur des images natives, les résultats sont comparés sur l'ensemble de la base VHP. Sur les images tomographiques, HOP a un débit 1.8% supérieur à CALIC, et 0.4% supérieur à JPEG-LS, mais améliore la compression de 2.7% par rapport à SPIHT et de 3.6% par rapport JPEG2000. Sur les IRM, HOP améliore la compression de 0.7% relativement à CALIC, 5.1% relativement à JPEG-LS, 1.8% par rapport à SPIHT et 4.3% par rapport à JPEG2000. Ainsi l'approche HOP peut permettre d'obtenir des taux de compression proches des DPCM tout en fournissant un codage multi-résolution. Ceci apporte une légère amélioration de l'ordre de 4% en comparaison aux standard JPEG2000. On peut également noter que HOP se comporte mieux fasse aux IRM qui sont plus contrastées que les tomographies, et qui possèdent un bruit d'arrière plan plus uniforme. L'effet de Gibbs apparaissant près des contours peut également contribuer à sélectionner une orientation o plus adéquate avec un estimateur G_o plus stable qu'en tomographie ou les artefacts sont des enchevêtrements très orientés.

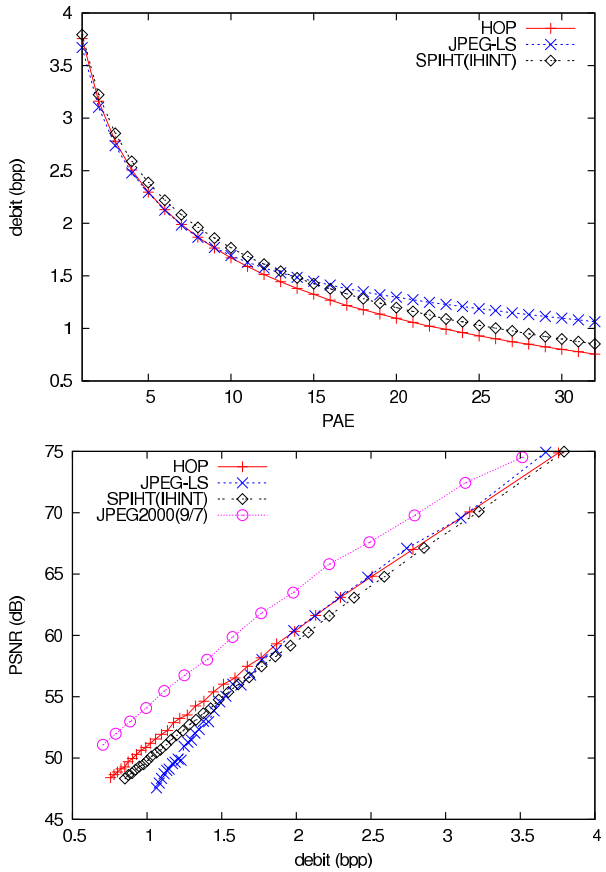


Figure 4 – *Résultats de la compression presque sans perte de l'image Fig. 6-(a). Débits sans perte : JPEG-LS*=4.90*, HOP*=5.02*, SPIHT(IHINT)*=5.05.

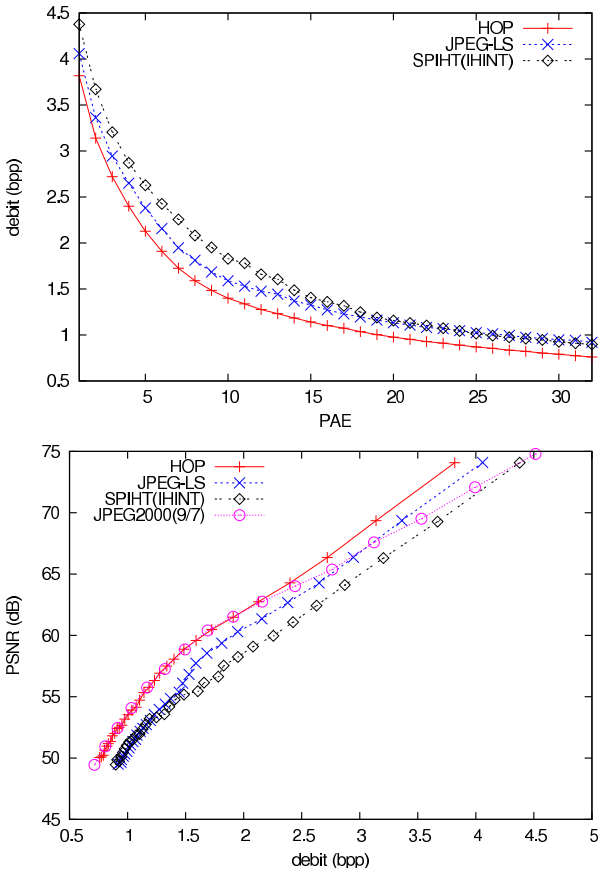


Figure 5 – *Résultats de la compression presque sans perte de l'image Fig. 6-(b). Débits sans perte : JPEG-LS*=5.59*, HOP*=5.35*, SPIHT(IHINT)*=5.94.

5.2 Compression presque sans perte

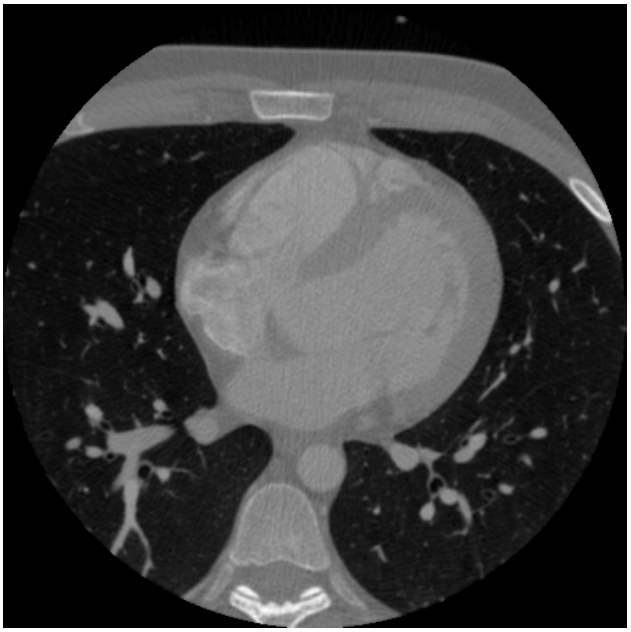
Les résultats presque sans perte de HOP ont été comparés avec ceux du standard JPEG-LS ainsi que l'extension presque sans perte du prédicteur hiérarchique dyadique IHINT [5] compressé avec une adaptation sans perte de l'algorithme SPIHT implémenté dans la bibliothèque QccPackSPIHT[4] (SPIHT(IHINT)). Les performances en PSNR des codeurs presque sans perte, contraints par le PAE, ont également été comparées avec celles du codeur JPEG2000 en mode compression irréversible (noyau 9/7), contraint par le débit. Les figures 4 et 5 illustrent les types de résultats pouvant être obtenus. La figure 4 montre les résultats sur l'une des coupes de tomographie cardiaque de la base MeDEISA pour laquelle les taux de compression sans perte de HOP sont parmi les plus mauvais en comparaison à JPEG-LS. HOP devance JPEG-LS à partir d'un PAE égal à 4. La figure 5 présente les résultats sur une section transverse d'une IRM du cerveau de la base native VHP-Male. Le premier exemple illustre également la supériorité en PSNR a haut débit des ondelettes sur des images contenant peu de hautes fréquences. Ceci peut alors favoriser les approches avec-pertes plus résidu pour la compression presque sans perte, tandis que le second exemple montre que sur les image bruitées (natives) le codage prédictif semble plus adapté lorsqu'un PAE raisonnable est choisi.

D'une façon générale, les résultats peuvent quelques peu varier mais SPIHT(IHINT) tend à dépasser JPEG-LS (débit/PSNR) pour des PAE importants (> 16) et peu permettre une progressivité en résolution ou en qualité. HOP s'est toujours montré plus efficace (débit/PSNR) que SPIHT(IHINT) et même dans les situations où JPEG-LS est plus performant en sans perte, HOP le devance rapidement en presque sans perte.

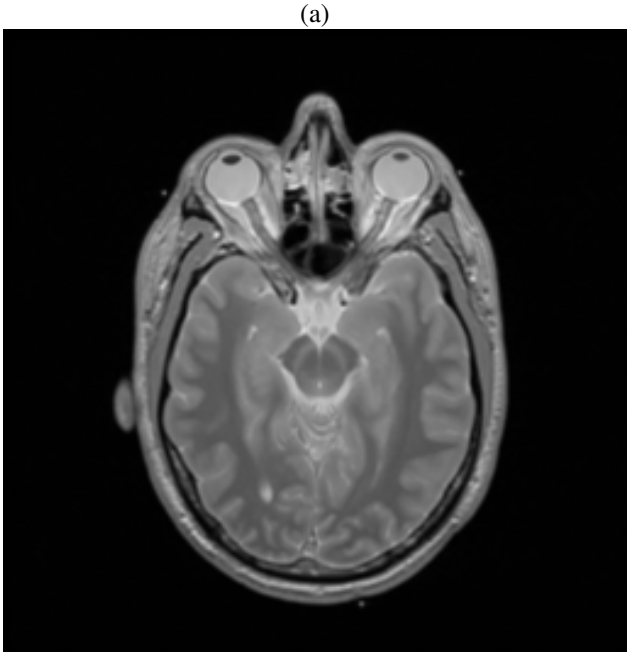
6 Conclusion et perspectives

Nous avons présenté un algorithme de compression sans perte et presque sans perte offrant une progressivité en résolution. Bien qu'il puisse être amélioré, les premiers résultats montrent qu'une approche combinée hiérarchique/DPCM peut permettre d'améliorer la compression sans perte d'images médicales de 4% par rapport à JPEG-2000, tout en conservant une scalabilité en résolution. En presque sans perte, les taux de compression sont généralement meilleurs que le standard JPEG-LS et

[4] http ://qccpack.sourceforge.net/



(a)



(b)

Figure 6 – *Images utilisées pour les résultats présentés dans les figures 4 et 5 : (a) coupe tomographique cardiaque de MeDEISA (CT_data_1(91)) et (b) coupe IRM cérébrale de VHP-Male (pd-1(21)).*

sur des images natives (bruitées) les distorsion introduites sont également plus faibles en PSNR que celles obtenus avec le standard JPEG2000 à haut débit.

Les résultats étant encourageants, des travaux sont envisageables pour l'amélioration des divers éléments de la chaîne de traitements, pour l'extension volumique de l'approche, ainsi que pour l'optimisation des prédicteurs qui permettrait d'obtenir de meilleurs taux de compression pour l'archivage, en particulier sur les images peu bruitées.

Remerciements

Ce travail a été réalisé grâce a un co-financement de la Région Bretagne pour une subvention de recherche doctorale INRIA (n° 4591).

Références

[1] A. Naït-Ali et C. Cavaro-Ménard, éditeurs. *Compression des images et des signaux médicaux*. LAVOISIER, 2007.

[2] M.J. Weinberger, G. Seroussi, et G. Sapiro. The LOCO-I lossless image compression algorithm : principles and standardization into JPEG-LS. *IEEE Trans. Image Proc.*, 9(8) :1309–1324, August 2000.

[3] X. Wu et N. Memom. Context-based, adaptative, lossless image coding. *IEEE Trans. Communications*, 45(4) :437–444, April 1997.

[4] P. Roos, M.A. Viergever, M.C.A. van Dijke, et J.H. Peters. Reversible intraframe compression of medical images. *IEEE Trans. Medical Imaging*, 7(4) :328–336, December 1988.

[5] A. Abrardo, L. Alparone, et F. Bartolini. Encoding-interleaved hierarchical interpolation for lossless image compression. *Signal Proc.*, 56(3) :321–328, February 1997.

[6] D. Taubman et M. Marcellin. *JPEG2000 : Image Compression Fundamentals, Standards and Practice*. Springer, 2001.

[7] W. Sweldens. Wavelets and the lifting scheme : A 5 minute tour. *Zeitschrift für Angewandte Mathematik und Mechanik*, 76 (Suppl. 2) :41–44, 1996.

[8] D. Taubman. High performance scalable image compression with EBCOT. *IEEE Trans. Image Proc.*, 9(7) :1158–1170, July 2000.

[9] A. Said et W.A. Pearlman. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. Circuits and Systems for Video Technology*, 6(3) :243–250, June 1996.

[10] D.L. Donoho. De-noising by soft-thresholding. *IEEE Trans. Information Theory*, 41(3) :613 –627, May 1995.

[11] G. Ulacha et R. Stasinski. A new simple context lossless image coding algorithm based on adaptive context arithmetic coder. Dans *Proc. IWSSIP International Conference on Systems, Signals and Image Proc.*, pages 45–48, June 2008.

[12] A. Krivoulets. A method for progressive near-lossless image compression. Dans *Proc. IEEE Int. Conf. Image Proc.*, volume 2, pages 185–188, Sept. 2003.

[13] S. Yea et W.A. Pearlman. A wavelet-based two-stage near-lossless coder. *IEEE Trans. Image Proc.*, 15(11) :3488–3500, Nov. 2006.

Vidéo 3D : quel débit pour la profondeur ?

Emilie Bosc[1] Vincent Jantet[2] Luce Morin[1] Muriel Pressigout[1] Christine Guillemot[2]

[1] IETR - INSA Rennes - 20 avenue des Buttes de Coësmes - 35043 Rennes, France
[2] INRIA Rennes, Bretagne Atlantique - Campus de Beaulieu - 35042 Rennes, France
[1]{Emilie.Bosc, Luce.Morin, Muriel.Pressigout}@insa-rennes.fr
[2]{Vincent.Jantet, Christine.Guillemot}@irisa.fr

Résumé

Cet article s'intéresse à la répartition du débit entre la texture et la profondeur lors de la compression de séquences multi-vues plus profondeur (MVD). Les effets de la quantification sur les deux types de données sont étudiés. La distorsion est mesurée sur les images synthétisées à partir des séquences MVD encodées et décodées par la méthode MVC. Bien que la profondeur soit codée sur une seule composante (contre trois pour la texture), allouer 25% du débit à la profondeur n'est pas le choix optimal. Les résultats montrent que plus de la moitié du débit doit être réservée aux données de profondeur.

Mots clefs

télévision 3D, vidéo multi-vues, synthèse de vue, compression, MVC, allocation de débit, cartes de profondeur

1 Introduction

Les vidéos 3D sont considérées comme l'évolution de la télévision conventionnelle actuelle. Le changement radical attendu est comparé à celui qu'occasionna l'introduction de la couleur à la télévision. Aujourd'hui, l'innovation majeure vient de l'apport de l'impression de profondeur générée par l'exploitation du phénomène de stéréopsie (perception de la profondeur de champ relative de deux stimuli présentés dans le champ visuel).

Ainsi, les représentations telles que les vidéos multi-vues (*MultiView Video* en anglais, MVV) permettent la création de vidéos 3D. Il s'agit de plusieurs séquences vidéo conventionnelles prises avec plusieurs caméras synchronisées et à des positions différentes dans la scène. Lorsque l'on associe ces vidéos à des vidéos dites de profondeur on parle de données *Multiview Video-plus Depth*, MVD. La Figure 1 illustre ce type de données constitué de séquences en couleur et de séquences de profondeur.

FIGURE 1 – *Les données de type MVD comportent des séquences de texture (à gauche) et de profondeur (à droite)*

La connaissance de la géométrie de la scène (issue des vidéos de profondeur) facilite la génération d'images virtuelles selon des points de vue différents de ceux réellement acquis par les caméras. Ces représentations de données permettent des applications telles que la télévision tridimensionnelle (3DTV) et le libre choix du point de vue (*Free viewpoint TeleVision*, FTV). 3DTV donne une impression de profondeur ou de relief à la scène alors que FTV offre à l'utilisateur la possibilité de choisir interactivement un point de vue arbitraire.
Les données traitées sont de taille considérable et une méthode efficace de compression est indispensable. De nombreuses solutions ont été proposées pour réaliser l'encodage des données de type MVD. Une problématique récurrente reste l'évaluation de l'allocation de débit qui optimise la qualité visuelle des images reconstruites à partir des données de texture et de profondeur.
Nous présentons dans ce document les études existantes ainsi que nos propres expériences réalisées pour répondre à ce problème. La section 2 revient sur les principes de synthèse de vue basée sur les cartes de profondeur. La section 3 rappelle les stratégies de compression de données MVD couramment utilisées. Enfin la section 4 montre les résultats des expériences réalisées pour déterminer le budget binaire optimal à attribuer aux vidéos de profondeur pour assurer des synthèses de vue de bonne qualité.

2 Synthèse de vue

La représentation MVD des vidéos 3D consiste en plusieurs séquences vidéo standard dites "vidéos de texture", et les vidéos de profondeur associées. Les cartes de pro-

fondeur (ou images de profondeur) qui composent les vidéos de profondeur sont des représentations bidimensionnelles de la scène 3D saisie. Ce sont des images en niveaux de gris, et de mêmes dimensions que les images de texture. Chaque pixel de la carte de profondeur indique la valeur de la profondeur du pixel de texture correspondant. Les valeurs de profondeur sont comprises dans un champ $[z_{near} - z_{far}]$, quantifié sur 8 bits. z_{near} est le point le plus proche de la caméra, et correspond à la valeur 255. z_{far} est le point le plus éloigné de la caméra, et correspond à la valeur 0.
Grâce à la géométrie projective [1], une carte de profondeur peut être convertie en une représentation tridimensionnelle. L'algorithme que nous utilisons pour cette opération de synthèse de vue à partir de cartes de profondeur est l'algorithme VSRS (View Synthesis Reference Software [2], version 3.5) fourni par MPEG. La Figure 2 illustre la relation entre un point 3D réel de la scène X et un point x_1 dans l'image de la caméra C_1 et son correspondant x_2 dans l'image de la caméra C_2. Cette transformation géométrique est possible grâce à la connaissance de la profondeur des pixels et des paramètres intrinsèques et extrinsèques des caméras. Ces dernières informations sont estimées lors du calibrage des caméras.

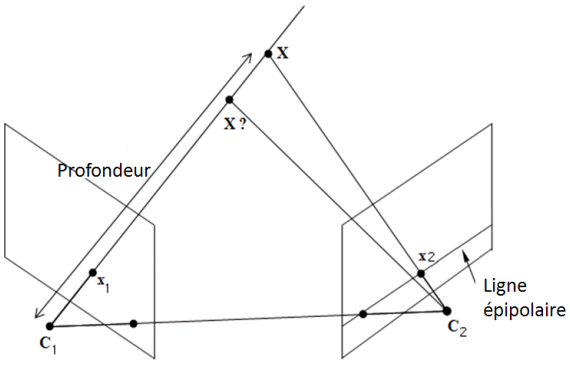


FIGURE 2 – *Relation entre les points de l'image et le monde réel [3]*

Selon le même principe, on peut projeter des points du monde réel sur le plan image d'une caméra virtuelle ayant un point de vue quelconque. Dans ce cas, l'algorithme VSRS nécessite les paramètres C_v de la caméra virtuelle. A partir des textures des deux vues adjacentes et des cartes de profondeur associées, l'algorithme estime la vue rendue selon le point de vue intermédiaire désiré.
La Figure 3 illustre le procédé utilisé par VSRS. Les deux cartes de profondeur sont projetées dans le point de vue intermédiaire. Il en résulte deux nouvelles cartes de profondeur correspondant au point de vue désiré. Ces cartes comportent des zones sans valeurs : ce sont les zones pour lesquelles on ne dispose pas d'informations dans la vue de référence (gauche, ou respectivement droite). Les zones manquantes de petites tailles peuvent être comblées par l'application d'un filtre médian. Les images de texture de référence sont ensuite elles-mêmes projetées dans le point de vue désiré, en fonction des profondeurs précédemment calculées. Il en résulte deux images de texture pour le même point de vue virtuel. Les vastes zones manquantes sont comblées en y attribuant les informations disponibles dans la deuxième image de texture générée, et vice versa. Enfin, ces deux images sont fusionnées, et on applique une méthode d'inpainting pour combler les dernières zones manquantes.

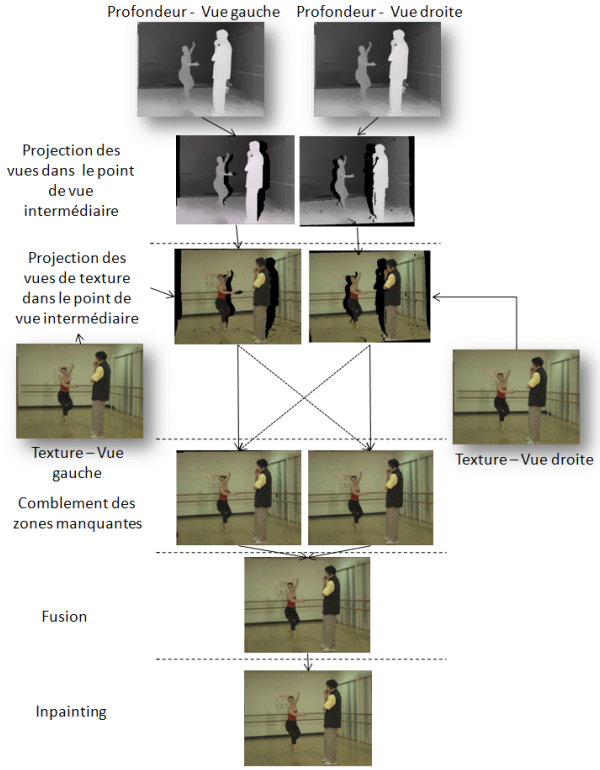


FIGURE 3 – *Synthèse d'une image intermédiaire avec VSRS*

Lors de la construction, expliquée précédemment, les valeurs de texture T_v de la vue virtuelle rendue sont attribuées aux pixels selon leur profondeur. Il y a bien trois cas possibles :
– Les profondeurs des pixels des vues adjacentes sont nulles, ce qui correspond à une zone occultée.
– Seule une des valeurs de profondeur est non-nulle, qui correspond à des régions occultées selon la nouvelle perspective.
– Les valeurs de profondeur des deux pixels sont non-nulles.
Ceci se traduit par l'expression suivante :

$$T_v = \begin{cases} 0, & \text{si } (u,v) \text{ n'est pas visible} \\ T_1(u,v), & \text{si } d_1(u,v) \neq 0 \\ & \text{et } d_2(u,v) = 0 \\ T_2(u,v), & \text{si } d_1(u,v) = 0 \\ & \text{et } d_2(u,v) \neq 0 \\ (1-\alpha)T_1(u,v) + \alpha T_2(u,v), & \\ & \text{si } d_1(u,v) \neq 0 \\ & \text{et } d_2(u,v) \neq 0 \end{cases}$$

où (u, v) fait référence aux coordonnées d'un pixel de l'image synthétisée, $d_1(u, v)$ la profondeur de ce pixel par rapport à la caméra C_1, $d_2(u, v)$ la profondeur de ce pixel par rapport à la caméra C_2, et α un facteur de pondération dépendant de la distance ($\alpha < 1$). A ce stade certaines causes d'artefacts peuvent être identifiées. Elles sont d'ordre géométrique : les régions occultées entraînent des zones dont la texture est nulle sur l'image rendue. Elles peuvent être dues aux arrondis des coordonnées des pixels calculées lors de la projection du monde réel vers le plan de l'image. L'inpainting fait partie des techniques de post-traitement inclues dans l'algorithme utilisé, qui permettent de pallier ce type d'artefacts.

3 Compression de vidéos multi-vues plus profondeur

Cette section aborde la compression de données de type MVV dans un premier temps, puis celle de données de type MVD. Enfin, les problématiques d'allocation de débit sont présentées.

3.1 Compression multi-vues (H.264/MVC)

L'idée la plus simple pour l'encodage des vidéos multi-vues consiste à encoder indépendamment chaque vue en utilisant un codeur de l'état de l'art. On appelle cette méthode *simulcast coding*. Ce procédé conduit à négliger l'exploitation des redondances existantes entre chaque vue, or ces redondances permettent de réaliser un gain significatif. Ainsi, en vue de surpasser les performances du codage simulcast, des modes de prédiction inter-vues sont ajoutés aux modes de prédiction spatiale et temporelle existant dans les codeurs simples de l'état de l'art. La compression H.264/AVC est considérée comme la meilleure méthode de compression pour les vidéos simples. Le codage multi-vues (*Multiview Video Coding* [4] en anglais) est basé sur cette méthode et a donné naissance au standard H.264/MVC [5]. La compression H.264/AVC [6, 7] exploitant déjà les redondances spatiales et temporelles au sein d'une même vue, l'algorithme MVC y ajoute l'utilisation des corrélations entre vues adjacentes.

La structure du codeur MVC consiste en une matrice de codeurs simples qui utilisent les redondances spatiales et temporelles. Les structures de bases des méthodes de codage sont conservées. Il s'agit des images codées *intra* (I) qui font appel aux pixels voisins à l'intérieur de la même image, des images *prédictives* (P) et des images *bidirectionnelles* (B) qui font référence aux pixels d'images voisines précédemment encodées ou à encoder.

L'originalité de la compression H.264 vient du fait qu'elle utilise des prédictions hiérarchiques, dont l'efficacité est prouvée [8]. Il existe des niveaux de prédictions, certaines images B pouvant devenir des références pour d'autres images B.

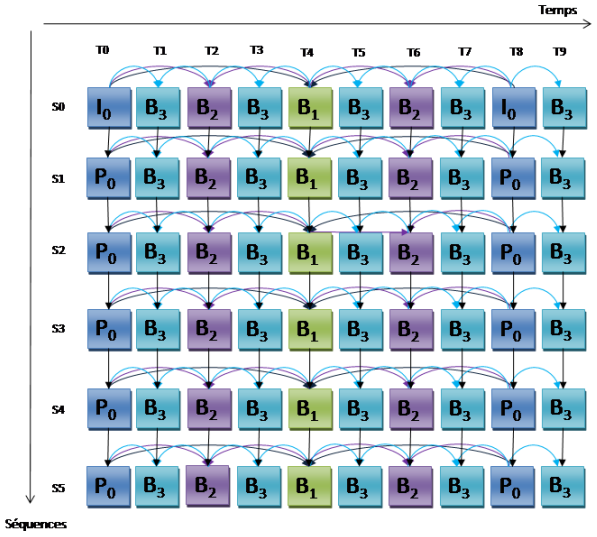


FIGURE 4 – *Schéma de la structure de la méthode H.264/MVC*

La prédiction inter-vues utilise la compensation en mouvement. La Figure 4 montre que les images encodées à un niveau de prédiction (et appartenant à une certaine vue) peuvent contribuer à la prédiction des images des vues adjacentes.

3.2 Compression de données MVD

Les séquences de profondeur peuvent être codées de façon particulière, en tenant compte de leurs caractéristiques [9], ou en réutilisant le codeur MVC, comme le proposent de nombreux travaux [10, 11, 12]. Lors d'une réutilisation du codeur MVC, rien n'impose d'utiliser les mêmes jeux de paramètres pour les cartes de profondeurs que pour les textures. Pour une utilisation judicieuse des canaux de transmission, restreints en bande passante, il convient de définir la distribution du budget binaire entre les séquences de texture et celles de profondeur.

En effet, compte tenu des propriétés des vidéos de profondeur, bien que ces dernières puissent être traitées comme des signaux monochromes et comprimées par les méthodes de l'état de l'art, il convient de leur appliquer un taux de compression adapté. Les codeurs de l'état de l'art sont souvent optimisés pour respecter des propriétés statistiques ou bien pour respecter la perception humaine des couleurs. Cependant, comme expliqué précédemment, les cartes de profondeur sont foncièrement différentes des images de texture, ce ne sont pas des images réelles. Par conséquent les artéfacts liés à la compression des cartes de profondeur n'impliquent pas les mêmes gênes que celles occasionnées par la compression d'images réelles. Pour cette raison, il convient de bien évaluer la quantification limite à appliquer aux vidéos de profondeur pour une qualité de synthèse optimale.

La synthèse de vue, décrite dans la section précédente, implique déjà des artefacts de projection à cause des zones

d'occultation et des imprécisions possibles de l'estimation de la carte de profondeur : certains pixels de la vue virtuelle peuvent ne pas trouver de point correspondant dans les vues de référence adjacentes. Ils nécessiteront une opération d'inpainting (interpolation par rapport aux valeurs des pixels voisins). La compression ajoute à cela de nouveaux artefacts de projection dans la vue rendue. Des expériences, menées dans [10], consistent à faire varier la compression des séquences de profondeur et à en observer l'impact sur les vues de synthèse rendues (existantes ou virtuelles). Les comparaisons sont basées sur des mesures de PSNR.

Les résultats confirment l'influence sérieuse de la qualité de l'information de la profondeur sur la qualité de la vue intermédiaire rendue. Les taux de compression élevés de la profondeur conduisent à des distorsions importantes au niveau des discontinuités, sur les bords des objets. En revanche, les auteurs constatent que la variation de la qualité de la texture influence moins la qualité de la vue rendue. Pourtant, des mesures objectives et subjectives menées dans [13] indiquent que les observateurs tolèrent plus facilement les artefacts de profondeur que ceux de texture, ce qui suggère que l'on peut favoriser une compression plus importante pour les informations de profondeur.

3.3 Allocation texture/profondeur

Dans la plupart des méthodes d'encodage, le débit affecté aux séquences de profondeur est restreint, et fixé à un certain pourcentage du débit total (généralement entre 10% et 15%). Dans [14], les auteurs estiment que seulement 5% à 10% du débit total suffisent pour l'encodage des séquences de profondeur avec la méthode H.264/AVC.

Dans les travaux présentés dans [15], le budget est fixe mais les images de profondeur sont sous-échantillonnées. Cette approche est basée sur le fait que sous-échantillonner une image vers une résolution moindre, puis l'encoder et l'agrandir à la résolution initiale améliore les performances en termes de PSNR. Il s'agit de réaliser un compromis entre la distorsion introduite par le sous-échantillonnage et la distorsion introduite par la quantification. A budget binaire fixé, le sous-échantillonnage permet une quantification moins brutale. Les résultats montrent un gain de 33% du débit total.

Face à l'approche consistant à attribuer un budget fixe, il existe également des méthodes d'optimisation de l'allocation du débit. Elles se basent sur des critères de qualité de l'information de la profondeur, ou bien de qualité de l'image de synthèse. C'est le cas dans [16], où le rapport optimal entre le débit et la distorsion est calculé en sélectionnant le groupe de paramètres de quantification qui favorise la meilleure qualité de vue rendue. Il y est proposé un algorithme d'optimisation de la répartition du débit par la recherche hiérarchique des paramètres de quantification idéaux. Pour un débit total de 75kbits, un gain de 1dB est enregistré sur la qualité de l'image rendue (en PSNR), comparé à une méthode d'allocation fixe (10% pour la profondeur).

Des études ont cherché à évaluer quantitativement la façon dont la compression de la texture et de la profondeur affecte la qualité des images rendues, en utilisant la méthode de compression H.264 comme dans [17]. Les auteurs proposent un modèle mathématique de la distorsion maximum possible. Cette dernière dépend de l'erreur induite par la compression de la profondeur et de l'erreur induite par la compression de la texture. Les résultats de ces expériences montrent que le modèle présenté est une indication approximative de la qualité de la vue virtuelle rendue et peut constituer un outil pour optimiser le rapport entre le débit et la distorsion due à la compression des deux types de séquences. La méthode d'optimisation d'allocation proposée obtient une amélioration de 0.3 à 1dB sur la qualité de l'image de synthèse comparée à une méthode d'allocation constante de débit (20% du débit total alloué à la profondeur).

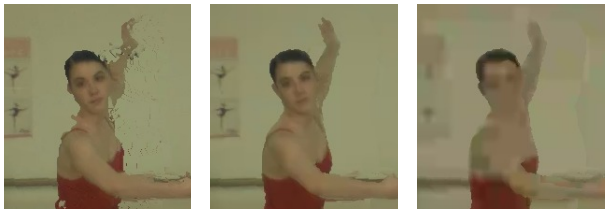
D'autres travaux ont mis en évidence les effets de la compression de la profondeur sur le rendu de vues [18] : comparé à une méthode de compression basée *platelet*, H.264/MVC obtient de meilleurs résultats en termes de PSNR. En revanche, la méthode basée platelet, adaptée aux caractéristiques de la carte de profondeur, préserve mieux les contours et donc la géométrie de la scène, d'où une meilleure qualité visuelle de l'image rendue.

La distribution optimale du débit entre les séquences de texture et de profondeur dépend fortement de l'utilisation qui sera faite des vidéos. Dans la section suivante, nous présentons une série d'expérimentations permettant d'évaluer le ratio optimal entre les débits de la texture et de la profondeur, pour une utilisation de la séquence MVD par VSRS.

4 Simulation et résultats expérimentaux

Nous avons souhaité évaluer la qualité d'images compressées par l'algorithme MVC en faisant varier la répartition du débit entre la texture et la profondeur [19]. Nous avons testé différents paramètres de quantification pour les textures et les cartes de profondeurs, afin d'observer la qualité du rendu obtenu par l'algorithme VSRS (version 3.5). Comme expliqué précédemment, VSRS est un algorithme développé par MPEG qui permet de synthétiser une vue intermédiaire à partir de deux vues adjacentes (textures) et des cartes de profondeur associées.

Allouer des débits différents à la texture et à la profondeur a un impact important sur la qualité finale d'une vue synthétisée. Pour un débit total fixé à 13 Mbit/s, la Figure 5 montre quelques images synthétisées à partir de vidéo multi-vues ("*Ballet*" de Microsoft Research), mais avec des proportions Texture/Profondeur différentes. Elles ont été construites avec VSRS, à partir de textures et de profondeur préalablement encodées par l'algorithme MVC.



(a) PSNR = 30.0dB ; Prof. = 3% débit ; (b) PSNR = 33.8dB ; Prof. = 60% débit ; (c) PSNR = 30.8dB ; Prof. = 95% débit ;

FIGURE 5 – *Images synthétisées à partir de séquences MVD à 13 Mbit/s, et pour différents rapports de débit entre texture et profondeur.*

– Avec 3% du débit alloué à la profondeur, la Figure 5(a) montre une importante dégradation des contours. Une trop forte compression de la carte de profondeur induit une perte de précision sur la position des discontinuités, provoquant des erreurs lors de la projection. Même si une grande partie de l'arrière plan est correct, le PSNR chute à cause des erreurs importantes commises sur les pixels de contours.
– Avec 95% du débit alloué à la profondeur, la Figure 5(c) montre des contours nets, mais des textures floues. La quantification trop forte des textures a entraîné une perte d'information importante, et une baisse de la qualité visuelle. Le PSNR chute également, à cause des erreurs commises sur tous les pixels de l'image.
– Avec 60% du débit alloué à la profondeur, la Figure 5(b) montre une qualité visuelle bien supérieure aux exemples précédents. Les cartes de profondeur sont suffisamment précises pour permettre les calculs de projections, et les textures sont suffisamment détaillées pour éviter des artéfacts de compression.

La qualité visuelle de ces images laisse penser qu'il existe un budget binaire optimal à attribuer à la profondeur.

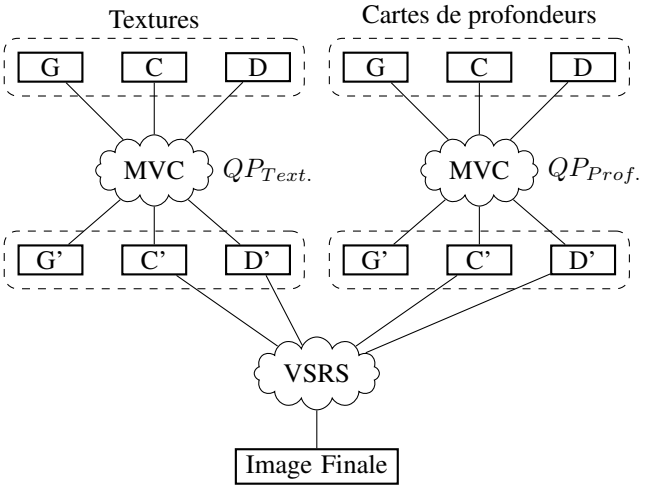


FIGURE 6 – *Représentation schématique du processus expérimental.*

La Figure 6 présente le protocole expérimental qui a permis d'évaluer le rapport de débit idéal entre la texture et la profondeur. Ce protocole se scinde en deux étapes distinctes :

– Dans un premier temps, les premières images des vues 2, 4 et 6 de la vidéo "*Ballet*" de Microsoft Research ont été sélectionnées. Sur le schéma 6, elles correspondent aux notations D pour droite, C pour centre, et G pour gauche. MVC a été utilisé pour compresser et décompresser ces textures d'une part, et les cartes de profondeur associées d'autre part. Notons $QP_{Text.}$, le pas de quantification des séquences de texture, et $QP_{Prof.}$ le pas de quantification des séquences de profondeur. Dans les deux cas, la vue 4 est la vue de référence, utilisée pour prédire les vues 2 et 6. Le débit de la séquence, calculé comme la somme des débits de la texture et de la profondeur, est directement lié aux pas de quantifications choisis.
– Dans un second temps, les vues décompressées 2 et 4 sont utilisées par le logiciel VSRS pour la synthèse d'une vue intermédiaire 3. Dans le cadre d'une utilisation en synthèse de vue, le PSNR entre cette vue générée et la vue 3 originale est le critère d'évaluation de la qualité de la vidéo multi-vues (nous disposons d'une vérité terrain).

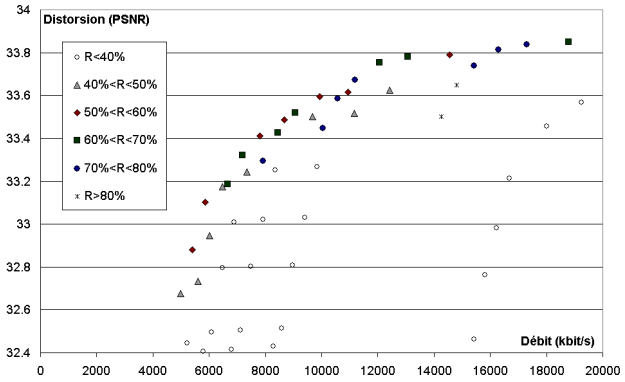


FIGURE 7 – *Courbe débit-distorsion pour différents rapports du débit profondeur sur le débit total.*

La Figure 7 rassemble les résultats de ce protocole expérimental, appliqué à un large choix de paramètres de quantification différents. Chaque point du graphique est représenté par un symbole, qui traduit le rapport R entre le débit alloué à la profondeur et le débit total. La position du point dans le graphique représente le PSNR (en dB) de la vue synthétisée en fonction du débit total. L'ensemble des points est délimité par une enveloppe convexe qui définit le rapport idéal entre le débit texture et profondeur.

La texture est une image à trois composantes, alors que la profondeur est représentée par une image à une seule composante. En conséquence, une hypothèse simple consisterait à penser que la profondeur nécessite moins de bande passante. Pourtant, les échantillons dont la profondeur représente moins de 40% du débit total sont très mal posi-

tionnés dans ce graphique. Au contraire, ce sont les échantillons de rapport R entre 50% et 70% qui approchent le mieux l'enveloppe. C'est ce rapport R voisin de 60% qui permet d'obtenir les meilleurs compromis entre le débit et la distorsion de l'image reconstituée. Ce rapport optimal n'est pas dépendant du débit total ciblé, ou de la distorsion maximale autorisée. Il apparait nécessaire d'allouer à la profondeur plus de la moitié, voir les deux tiers, du débit, pour assurer une bonne qualité de synthèse de vue.

5 Conclusion

A partir de vidéos MVD, nous avons utilisé le codeur MVC pour compresser les textures d'un coté, et les cartes de profondeur de l'autre. En faisant varier les pas de quantifications, on observe l'évolution du PSNR lors de la synthèse de vue par VSRS. La compression des textures fait apparaitre des zones floues, et la compression des cartes de profondeur entraîne des distorsions géométriques. A débit total fixé, il existe un rapport idéal entre le débit alloué à la texture et celui alloué à la profondeur, maximisant la qualité visuelle de l'image générée. Pour assurer la meilleure qualité d'image lors de la synthèse de vue avec VSRS, nous avons montré qu'il est nécessaire d'attribuer aux cartes de profondeur environ 60% du débit total.

Remerciements

Ces travaux ont été réalisés dans le cadre des projets ANR-PERSEE, ANR-CAIMAN et DGE-Région FUTURIM@GE. Nous souhaitons également remercier Microsoft Research pour la mise à disposition de la séquence multi-vues "*Ballet*".

Références

[1] R. Hartley et A. Zisserman. *Multiple view geometry in computer vision*. Cambridge Univ Pr, 2003.

[2] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, et Y. Mori. Reference softwares for depth estimation and view synthesis. Avril 2008.

[3] C. Lee et Y. S Ho. View synthesis tools for 3D Video.ISO/IEC JTC1/SC29/WG11 MPEG2008/M15851, Octobre 2008.

[4] Y. Chen, Y. K Wang, K. Ugur, M. M Hannuksela, J. Lainema, et M. Gabbouj. The emerging MVC standard for 3D video services. *EURASIP Journal on Advances in Signal Processing*, 2009(1), 2009.

[5] ISO/IEC JTC1/SC29/WG11,. Text of ISO/IEC 14496-10 :200X/FDAM 1 multiview video coding, 2008.

[6] ITU-T Recommendation H.264. Advanced video coding for generic Audio-Visual services, 2009.

[7] T. Wiegand, G. J Sullivan, G. Bjontegaard, et A. Luthra. Overview of the h. 264/AVC video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7) :560–576, 2003.

[8] H. Schwarz, D. Marpe, et T. Wiegand. Analysis of hierarchical b pictures and MCTF. Dans *Proc. ICME*, page 1929–1932, 2006.

[9] Y. Morvan, D. Farin, et P. H.N De. Depth-Image compression based on an RD optimized quadtree decomposition for the transmission of multiview images. Dans *IEEE Int. July*, page 105–108, 2007.

[10] P. Merkle, A. Smolic, K. Muller, et T. Wiegand. Multi-view video plus depth representation and coding. Dans *Proceedings of ICIP*, page 201–204, 2007.

[11] P. Merkle, K. Muller, A. Smolic, et T. Wiegand. Efficient compression of multi-view video exploiting inter-view dependencies based on h. 264/MPEG4-AVC. Dans *Proc. ICME*, page 9–12, 2006.

[12] I. Daribo, M. Kaaniche, W. Miled, M. Cagnazzo, et B. Pesquet-Popescu. Dense disparity estimation in multiview video coding. Dans *Proc. of the IEEE Workshop on Multimedia Signal Processing (MMSP), Rio de Janeiro, Brazil*, 2009.

[13] A. Tikanmaki, A. Gotchev, A. Smolic, et K. M\üller. Quality assessment of 3D video in rate allocation experiments. Dans *IEEE Int. Symposium on Consumer Electronics (14-16 April, Algarve, Portugal)*, 2008.

[14] E. Martinian, A. Behrens, J. Xin, et A. Vetro. View synthesis for multiview video compression. Dans *Picture Coding Symposium*, volume 37, page 38–39, 2006.

[15] E. Ekmekcioglu, S. T Worrall, et A. M Kondoz. Bitrate adaptive down-sampling for the coding of multiview video with depth information. Dans *3DTV Conference : The True Vision-Capture, Transmission and Display of 3D Video, 2008*, page 137–140, 2008.

[16] Y. Morvan, D. Farin, et P.H.N. de With. Joint depth/texture bit-allocation for multi-view video compression. *Picture Coding Symposium*, 10(1.66) :4349, 2008.

[17] Y. Liu, Q. Huang, S. Ma, D. Zhao, et W. Gao. Joint video/depth rate allocation for 3D video coding based on view synthesis distortion model. *Signal Processing : Image Communication*, 24(8) :666–681, Septembre 2009.

[18] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Muller, et T. Wiegand. The effects of multiview depth video compression on multiview rendering. *Singal Processing : Image Communication*, 24(1-2) :73–88, 2009.

[19] K. Muller, A. Smolic, K. Dix, P. Merkle, et T. Wiegand. Coding and intermediate view synthesis of multiview video plus depth. pages 741–744, Novembre 2009.

Génération, Compression et Rendu de LDI

Vincent Jantet [1] Luce Morin [2] Christine Guillemot [1]

[1] INRIA Rennes, Bretagne Atlantique, Campus de Beaulieu, 35042 Rennes – France

[2] IETR - INSA Rennes, 20 avenue des Buttes de Coësmes, 35043 Rennes – France

Résumé

Ce papier présente l'utilisation de Layered Depth Images (LDI) dans le cadre de la compression de vidéo MVD. Leur construction incrémentale (I-LDI) permet l'élimination d'une grande partie des corrélations inter-vues, sans nécessiter de caméras rectifiées. Leur compression par le codeur MVC permet en plus l'exploitation des corrélations temporelles. Une solution aux effets de fantômes (Ghosting) est proposée, basée sur une détection de contours dans les cartes de profondeur, suivie d'une classification locale premier plan/arrière plan. L'élimination des découvrements (Disocclusions) et craquelures de textures est permise par une projection ordonnée des pixels, suivie d'une projection inverse. Les tests sur des vidéos MVD montrent une réduction de 80% du nombre de pixels sur les couches supplémentaires des I-LDI. La comparaison des courbes débit-distorsion de MVC appliqué sur les vues, et de MVC appliqué sur les couches I-LDI, montre un apport de qualité pour des débits faibles.

Mots clefs

Codage vidéo, Réalité virtuelle, Vidéo Multi-Vues, Layered Depth Video, Interpolation de vues.

1 Introduction

La banalisation des dispositifs stéréoscopiques impose l'ajout d'une nouvelle dimension aux données multimédias classiques. L'acquisition d'une scène par plusieurs caméras synchronisées (MVV pour *Multi-View Video*) est un des procédés permettant l'estimation de la géométrie de la scène. La connaissance de cette géométrie est indispensable aux fonctionnalités 3D comme l'affichage en relief (3DTV) ou la sélection de points de vue (FVV pour *Free Viewpoint Video*).

Les informations géométriques peuvent être de différentes formes (maillage, fonction plénoptique, polygones, cartes de profondeur, ...). Chacune de ces formes constitue un format de données indépendant, et est associée à un algorithme de synthèse de vues approprié. Ces algorithmes de rendu peuvent être classés selon deux catégories, en fonction du type d'informations géométriques qu'ils utilisent.

– Les rendus basés géométrie (GBR) utilisent un modèle 3D évolué et très détaillé. Ils sont idéaux pour le rendu de scènes synthétiques, mais inadaptés à des données réelles, où le modèle géométrique est difficile à obtenir :
– Les rendus basés images (IBR) nécessitent une géométrie moins détaillée (souvent des cartes de profondeur), mais ont besoin de plus d'information de texture. Les données d'entrée sont moins compactes, mais les images synthétisées sont de meilleure qualité et plus réalistes.

Les LDI (pour *Layered Depth Images*) [1, 2] constituent l'une de ces approches IBR. La géométrie est modélisée par une image en plusieurs couches, où chaque couche est accompagnée de la carte de profondeur correspondante. Chaque point dans l'image en couche est donc associé à un ensemble de pixels, constitués chacun d'une couleur et d'une profondeur. Ces pixels représentent les objets (visibles ou masqués dans la scène) traversés par un rayon reliant le centre optique de la caméra au point image considéré. Cette représentation selon un seul point de vue, réduit efficacement les redondances inter-vues, tout en permettant une synthèse de vue réaliste pour des points de vue éloignés du point de vue de référence.

Deux méthodes de construction de LDI [3] sont détaillées dans la section 2, dont une approche incrémentale permettant d'éliminer les corrélations entre les couches. La compression de ces LDI [4, 5] par le codeur MVC [6] est analysée dans la section 3. Une méthode de synthèse de vues permettant de combler les découvrements et d'éliminer les effets de fantôme est décrite dans la section 4. Les résultats présentés dans la section 5 permettent la validation des trois étapes mises à la chaîne : Construction, Compression puis Rendu.

2 Construction des LDI

Les méthodes de construction de LDI s'appuient en entrée sur des vidéos multi-vues [3], où chaque vue est accompagnée de sa carte de profondeur (MVD pour *Multi-View + Depth*). Le logiciel de référence pour les construire (MVD2LDV du groupe MPEG [7]) présente un certain nombre de limitations :

– Les caméras d'entrée doivent être rectifiées, c'est-à-dire que les lignes épipolaires sont toutes alignées horizontalement. Cette hypothèse simplifie les équations de projection (présentées dans la section 2.1), mais impose aux caméras d'avoir des poses précises et des paramètres intrinsèques identiques. Cette propriété, difficile à avoir

dès l'acquisition, est généralement obtenue par un traitement sur les vues acquises.
– Le nombre de vues en entrée est limité à trois, et le nombre de couches générées est limité à deux.

Deux approches sont proposées pour remédier à ces limitations. La première, un peu naïve, est présentée dans la section 2.2. Simple à mettre en œuvre, elle présente le défaut de générer un grand nombre de couches, partiellement corrélées les unes aux autres. La seconde approche, détaillée dans la section 2.3, est une construction incrémentale. Les couches générées sont moins nombreuses, et moins remplies, ce qui facilite leur compression ultérieure.

2.1 Equations de projection

Dans une image, chaque pixel p est la projection perspective d'un point 3D M selon les équations (1). Ces équations nécessitent la connaissance des paramètres intrinsèques K, de la matrice de rotation R et de la position t de la caméra dans le repère global.

Les équations inverses (2) permettent de retrouver la position 3D $M = (X, Y, Z)$ d'un point, à partir du pixel p et de sa profondeur Z dans l'image.

$$\omega p = KR^{-1}(M - t) \quad (1)$$

$$M = \omega RK^{-1}p + t \quad (2)$$

avec ω le coefficient de normalisation.

Ces équations sont utilisées par les techniques de synthèse de vues basées sur des cartes de profondeur [2, 3]. Elles génèrent des effets de craquelures sur les textures et des zones de découvrements qui seront détaillés dans la section 4. Ces paramètres peuvent être estimés à partir des vidéos MVD, qui dans certains cas peuvent être rectifiés. Nous supposerons donc dans la suite que les paramètres caméras (K, R et t) sont connus.

2.2 Construction naïve des LDI

Une LDI est un concentré d'information selon un même point de vue. La méthode de construction naïve consiste donc à projeter toute l'information présente dans la MVD selon un unique point de vue de référence, puis à fusionner cette information et à l'organiser par couches.

Le point de vue de référence peut être choisi quelconque, et n'a pas à être identique à un des points de vue de la MVD. Les vues sont considérées comme des ensembles de pixels, chacun ayant une couleur et une profondeur. Tous les pixels de chaque vue sont projetés selon le point de vue de référence. Lorsque deux pixels p_1 et p_2, des vues respectives V_1 et V_2, représentent le même point 3D M, alors ils doivent être projetés sur le même pixel p' selon le point de vue de référence, et à la même profondeur Z. Pour éviter cette redondance, un test sur la profondeur est utilisé et les pixels trop proches (seuil Δ_Z) sont fusionnés. Cette construction simpliste est schématisée dans la figure 1.

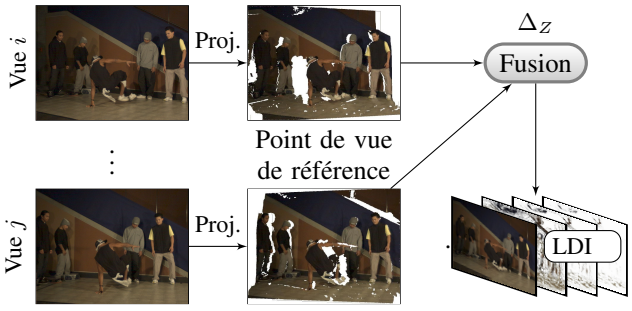


Figure 1 – *Schéma de construction naïf.*

Les trois premières couches d'une LDI sont présentées sur la figure 2. Chaque couche contient l'ensemble des pixels les plus proches de la caméra, qui ne soient pas déjà dans une des couches précédentes. Ainsi, la première couche contient tous les pixels visibles, alors que les autres couches sont partiellement vides car ne contiennent que les pixels masqués par un objet de la scène.

(a) 1ere couche (b) 2eme couche (c) 3eme couche

Figure 2 – *Premières couches d'une LDI naïve (Ref. = Vue 4) générée en utilisant les 8 caméras d'acquisition. ($\Delta_Z = 0.1$)*

Le seuil Δ_Z est critique pour l'élimination des redondances, mais difficile à déterminer à l'avance. Une autre méthode de construction est donc proposée, utilisant un critère totalement différent.

2.3 Construction incrémentale des LDI

Pour éliminer les corrélations entre les vues, une autre construction est proposée, désignée par I-LDI, pour Incremental-LDI [8]. Basée sur l'extraction d'information résiduelle [9], elle permet de réduire le nombre de couches et leur taux de remplissage.

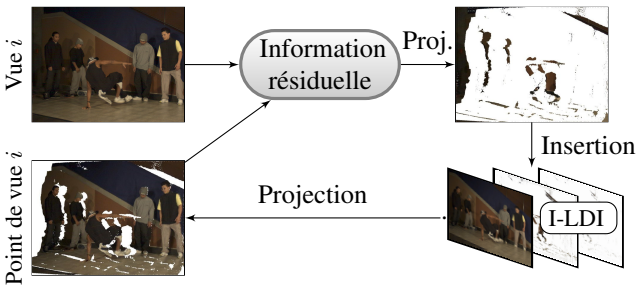


Figure 3 – *Extraction de l'information résiduelle de la vue i lors d'une itération de la construction I-LDI.*

L'algorithme peut se décomposer de la façon suivante :

Initialisation : Le point de vue de référence est fixé arbitrairement. Il n'a pas besoin d'être un des points de vue des caméras d'acquisition. La I-LDI est initialisée vide, elle ne contient aucun pixel.

Itérations : La I-LDI est projetée selon le point de vue d'une caméra d'acquisition, ce qui génère des zones de découvrement. Cette image virtuelle est comparée à l'image acquise de ce point de vue. L'ensemble des pixels permettant de combler les zones de découvrement est appelé information résiduelle. Seule cette information résiduelle est re-projetée selon le point de vue de référence pour être ajoutée dans la I-LDI. Cette étape est réitérée pour chacune des caméras d'acquisition, ajoutant de plus en plus de pixels dans la I-LDI.

Finalisation : L'ensemble des pixels est réparti par couches, chaque couche contenant les pixels les plus proches, non déjà contenus dans une couche précédente. Par construction, la première couche est l'image 2D+Z observée selon le point de vue de référence.

Les trois premières couches d'une I-LDI sont présentées dans la figure 4. La seconde couche ne contient que les textures des zones réellement occultées dans la scène.

(a) 1ere couche (b) 2eme couche (c) 3eme couche

Figure 4 – *Premières couches d'une I-LDI (Ref. = Vue 4), générée en utilisant les 8 caméras d'acquisition dans un ordre B-hiérarchique : (4 ; 2 ; 6 ; 1 ; 3 ; 5 ; 7 ; 0).*

La figure 5 présente le taux de remplissage des huit premières couches d'une LDI et d'une I-LDI. Dans les deux cas, la première couche est complète. Dans une LDI, le nombre de pixels cumulé de toutes les couches supplémentaires représente plus de 50% du nombre de pixels dans la première couche. Ce pourcentage est réduit à moins de 10% par la construction incrémentale. Dans ce cas, les couches à partir de la 3eme sont pratiquement vides.

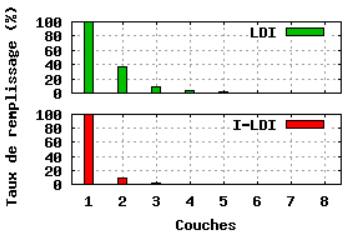


Figure 5 – *Taux de remplissage moyen des couches d'une LDI et d'une I-LDI*

3 Compression des LDI

Les vidéos MVD représentent un important volume de données, mais très fortement corrélées. Les corrélations inter-vues viennent s'ajouter aux corrélations temporelles, déjà présentes dans une vidéo classique. Pour compresser ces vidéos, une des propositions mise en œuvre par MPEG est le codeur MVC, présenté dans la partie 3.1. Ce codeur présente certaines limitations dans le cadre multi-vues, c'est pourquoi la partie 3.2 présente une adaptation de ce codeur aux couches d'une LDI.

3.1 Codeur MVC

Dans un codeur vidéo 2D comme H.264/AVC, quelques images clés servent de référence, et sont encodées indépendamment des autres. Les images restantes sont prédites à partir de ces images clés, par l'utilisation d'un champ de mouvement, codé et transmis avec les images clés. L'erreur de prédiction est également encodée, pour permettre d'être corrigée par le décodeur et d'améliorer la qualité. Ce principe à été étendu aux séquences multi-vues par le codeur MVC (pour *Multi-View Coding*, une extension de H.264/AVC). L'une des vues est choisie comme vue de référence, et est encodée comme une vidéo classique. Chaque image des vues supplémentaires est alors prédite à partir des images au même instant des vues déjà encodées et des images de la même vue mais à des instants différents.

Comparée à une compression Simulcast, où chaque vue est encodée indépendamment par H.264/AVC, la compression MVC présente les résultats suivants [6] :

– À débit équivalent, MVC apporte un gain en qualité entre 1 et 1.5 dB.
– À qualité équivalente, MVC apporte un gain de seulement 15% à 30% du débit.
– Plus de 90% des corrélations dans une vidéo multi-vues sont des corrélations temporelles.
– Le débit du flux MVC est linéairement lié au nombre de vues supplémentaires, rendant son utilisation difficile lorsque ce nombre de vues augmente.

3.2 Codage des LDI par MVC

Les déformations géométriques entre les vues sont prédites par MVC à l'aide d'un champ de mouvement par blocs, ce qui n'est pas adapté. L'avantage de la construction LDI est de corriger ces déformations géométriques, tout en éliminant une grande partie des corrélations inter-vues. Le codeur MVC est utilisé sur la séquence de LDI, en considérant chaque couche comme une vidéo différente. Pour cela, les couches partiellement vides sont complétées par le contenu de la couche précédente, de sorte que le codeur MVC n'ait pas à re-coder ces zones. MVC ne supportant que des vidéos à trois composantes, le codeur MVC est utilisé sur l'ensemble des textures, puis séparément sur l'ensemble des cartes de profondeur.

La figure 6 présente le graphique débit/distorsion obtenu par la compression MVC des LDI, et par la compression MVC des données MVD. En utilisant les quatre vues (4 ;

2 ; 6 ; 0) de la vidéo "Ballet" (MSR), les LDI sont générées d'une part, puis limitées à seulement deux couches. Les textures et les profondeurs sont encodées indépendamment par MVC, en utilisant les mêmes pas de quantification. Après décodage, les couches de la LDI sont utilisées pour générer la vue intermédiaire 5. Ces quatre mêmes vues sont encodées d'autre part directement par MVC (texture et profondeur au même pas de quantification). Les vues 4 et 6 décodées sont utilisées par VSRS, le logiciel de synthèse de vue de MPEG [7], pour synthétiser la vue 5.

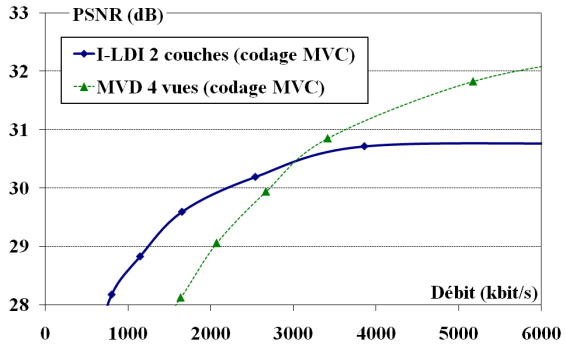


Figure 6 – *Comparaison débit-distorsion sur la vidéo "Ballet" (MSR), entre le format LDI codé par MVC et le format MVD codé par MVC.*

Le format LDI proposé présente un gain en qualité significatif comparé au standard MVC classique, lorsque les débits cibles sont faibles (de l'ordre de 3Mbits/s). Pour des débits plus élevés, le format LDI atteint rapidement une saturation, due aux erreurs de projection et d'alignement des contours.

4 Synthèse de vues

Les LDI contiennent suffisamment d'informations pour permettre la synthèse de vues virtuelles. Cette synthèse de vue se heurte à plusieurs difficultés. La section 4.1 présente les solutions classiques contre le problème de découvrement. La section 4.2 propose une solution plus complète, résolvant également les craquelures de textures. La section 4.3 explique l'apparition de fantômes, et propose une solution pour les éviter.

4.1 Inpainting

La projection de chaque pixel d'une image selon un nouveau point de vue génère des craquelures (erreurs d'échantillonnage) et des découvrements (texture inconnue), visibles sur la figure 7(a). Les craquelures peuvent être comblées par application d'un filtre médian sur la carte de profondeur, puis re-synthèse de la texture par projection inverse [10]. Les découvrements sont plus difficiles à compléter, puisqu'il manque de l'information. Plusieurs solutions d'inpainting ont été proposées, utilisant les textures avoisinantes.

– La bibliothèque OpenCV propose une méthode ne tenant pas compte de la profondeur. La figure 7(b) montre que la couleur d'avant plan semble s'étaler dans l'arrière plan.
– Une méthode dérivée utilise la profondeur pour remplacer temporairement l'avant plan par de l'arrière plan. L'inpainting est ensuite utilisé avec uniquement des textures d'arrière plan. Enfin, le remplacement inverse est effectué [10].

Ces deux méthodes ne permettent ni de détecter, ni de combler, les craquelures sur des textures d'avant plan. La solution proposée dans la section 4.2 s'appuie sur une projection ordonnée des pixels, dont le résultat est visible sur la figure 7(c). Un effet d'étirement est observable dans toute la zone remplie.

(a) Découvrements (b) Inpainting OpenCV (c) Proj. Ordonnée

Figure 7 – *Découvrements 7(a) et comparaison des résultats d'inpainting d'openCV 7(b) et de la projection ordonnée 7(c).*

4.2 Projection Ordonnée

McMillan [11] propose de projeter les pixels dans un ordre déterminé, pour appliquer l'algorithme du peintre et se passer de Z-Buffer. Cet ordre de parcours a pour second avantage de détecter les découvrements pendant la projection. La différence d'abscisse, après projection, entre un pixel et son voisin, permet de détecter les recouvrements et les découvrements. L'ordre est défini en fonction de l'épipôle, de sorte que les derniers pixels posés soient des pixels d'avant plan.

La figure 8(b) présente le principe lorsque les caméras sont rectifiées, c'est-à-dire que les pixels peuvent être traités par ligne. Soit un pixel $P_i = (x_i, y_i)$ de l'image d'origine se projetant en $P'_i = (x'_i, y'_i)$ dans l'image finale. L'égalité $y_i = y'_i$ implique que chaque ligne, que nous supposerons parcourue de gauche à droite, est indépendante des autres. Dans ce cas :

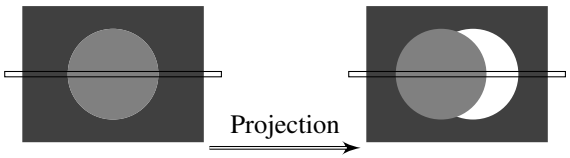
– Si $x'_i < x'_{i-1}$, alors il y a recouvrement, et le pixel courant est un pixel d'avant plan.
– Si $x'_i > x'_{i-1}$, alors il y a découvrement, et le pixel courant est un pixel d'arrière plan, pouvant servir à compléter.

Ce procédé est implémenté dans LDVRS, le logiciel de synthèse de vues rectifiées à partir des LDI de MPEG [7].

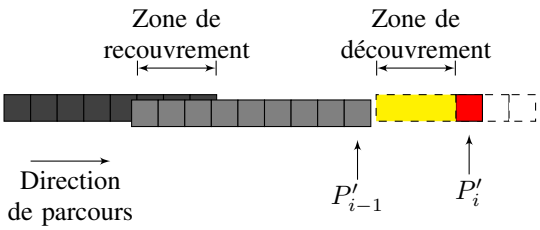
La figure 8(c) présente le principe dans le cadre de projections quelconques. L'égalité de y_i et de y'_i n'est plus assurée, et les lignes doivent être parcourues simultanément (en pratique, l'image est parcourue par colonne). Pour chaque ligne j de l'image finale, on définit la limite X^j_{max} des pixels déjà remplis par l'équation (3).

$$X^j_{max} = \max_{\{P'_i\,:\,y'=j\}} (x') \tag{3}$$

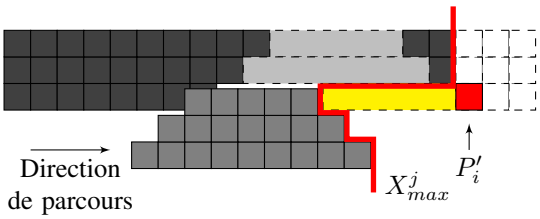
Dans ce cas :
- Si $x'_i < X^j_{max}$, alors il y a recouvrement.
- Si $x'_i > X^j_{max}$, alors il y a découvrement, et le pixel courant est un pixel d'arrière plan, pouvant servir à compléter la zone découverte.



(a)



(b) Détails d'une ligne de pixels pour une projection rectifiée.



(c) Détails de quelques lignes de pixels pour une projection non rectifiée.

Figure 8 – *Principe de la projection ordonnée.*

Cette projection n'est pas plus coûteuse en temps de calcul qu'une projection classique, mais du fait de son ordonnancement, son implémentation sur GPGPU est difficile.

4.3 Fantômes

Dans les images acquises pas les caméras, les contours des objets ne sont jamais nets. La couleur des pixels de contour reçoit la contribution de l'objet et celle de l'arrière plan. Par contre, la profondeur de ces pixels est tantôt celle de l'objet, tantôt celle de l'arrière plan. Après projection, ces pixels se retrouvent éloignés de la nouvelle frontière de l'objet, et génèrent un effet de fantôme.
La solution proposée est de ne pas tenir compte des pixels trop proches des frontières d'objets. Pour cela, un détecteur de Canny est appliqué sur la carte de profondeur, afin de localiser les contours. Une fenêtre de taille W est définie autour des contours, dans laquelle tous les pixels sont classés en fonction de leur profondeur. Les pixels dont la profondeur est supérieure à la profondeur moyenne dans la fenêtre, sont éliminés.

La figure 9 présente les résultats de synthèse de vues à partir d'une LDI à deux couches. Le fantôme 9(a) est détecté et effacé par cette méthode 9(b), puis remplacé par de l'inpainting 9(c).

(a) Fantôme. (b) Suppression. (c) Remplissage.

Figure 9 – *Suppression des effets de fantôme. ($W = 7$)*

5 Résultats

Des mesures de la qualité du rendu ont été effectuées, à partir des vidéos MVD "Ballet" et "Breakdancers" [12]. Les I-LDI à deux couches ont été générées selon le point de vue 4, à partir de la première image de chacune des 8 vues, utilisées dans un ordre B-hiérarchique (4 ; 0 ; 7 ; 2 ; 6 ; 1 ; 5 ; 3).
La figure 10 présente les PSNR, mesurés entre les vues générées et les vues acquises, pour la vidéo "Breakdancers". Plus la caméra virtuelle s'éloigne du point de vue de référence, plus les erreurs de projection sont accentuées, réduisant ainsi la qualité de l'image générée.

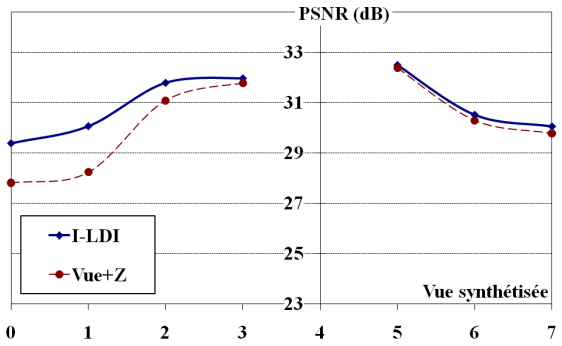


Figure 10 – *Qualité de rendu des vues virtuelles selon le point de vue. (Séquence : "Breakdancers" ; Ref. : Vue 4)*

Pour mesurer l'apport de la seconde couche de la I-LDI, les mêmes mesures ont été faites à partir uniquement de l'image 2D+Z de la vue de référence (c'est-à-dire la première couche de la I-LDI). Les PSNR obtenus sont toujours plus faibles qu'en utilisant les deux couches, mais dans des proportions très différentes selon la position de la caméra virtuelle. La géométrie de la scène fait que, pour une caméra virtuelle numérotée entre 0 et 3, les zones découvertes ont des textures uniformes, facilement estimées par les solutions d'inpainting. A l'inverse, pour une caméra virtuelle numérotée entre 5 et 7, des zones fortement texturées apparaissent, rendant pertinent le contenu de la seconde couche. Le logiciel LDVRS de MPEG [7] à été utilisé pour comparer les résultats précédents avec ce qui se fait déjà en

synthèse de vue à partir de LDI. Ce logiciel utilise des LDV (semblables aux LDI à deux couches) pour générer des vues virtuelles rectifiées par rapport au point de vue de référence. VSRS [7] a été utilisé sur la vidéo MVD non compressée, pour générer un ensemble de vues rectifiées pouvant servir de références aux calculs de distorsions.
La figure 11 présente les résultats obtenus sur la vidéo "Ballet". Les mêmes I-LDI sont utilisées pour un rendu par projection ordonnée, et pour un rendu par LDVRen, un des algorithmes de LDVRS. Quel que soit le point de vue généré, les améliorations proposées dans cet article apportent un gain en PSNR comparé au rendu de LDVRen. En effet, LDVRen est conçu pour synthétiser des vues très proches de la vue de référence. Lorsque la caméra s'éloigne, il commet des erreurs qui peuvent le rendre moins performant que des algorithmes basés sur une seule vue 2D+Z.

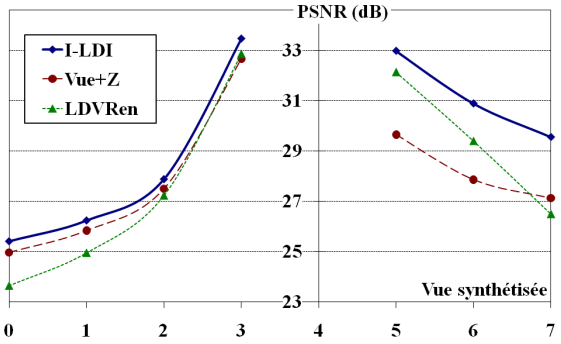


Figure 11 – *Qualité de rendu pour des vues virtuelles rectifiées. (Séquence : "Ballet" ; Ref. : Vue 4)*

6 Conclusions

Cet article propose une approche basée sur les LDI pour la compression et le rendu de vidéos MVD non rectifiées. La construction incrémentale permet d'éliminer les corrélations inter-vues tout en réduisant le nombre de couches produites. Ainsi, la seconde couche ne contient que 10% de pixels différents, comparée à la première couche. La compression MVC appliquée sur ces couches permet, en plus, l'exploitation des corrélations temporelles. Les I-LDI produites sont presque indépendantes du nombre de vues utilisées, ce qui rend le format très compétitif comparé à la compression MVC directement appliquée sur les vues. Quelques méthodes sont proposées pour améliorer la qualité des vues virtuelles générées. La totalité des zones découvertes sont comblées par une solution d'inpainting basée sur un ordonnancement des pixels lors de leur projection. Les effets de fantômes sont éliminés par une détection de contours dans la carte de profondeur. La comparaison avec LDVRen pour des vues rectifiées permet d'observer une amélioration de la qualité du rendu.
Les travaux futurs utiliseront un algorithme d'inpainting directionnel pour éliminer l'effet d'étirement produit par la projection ordonnée. Une étude sera également menée sur l'utilisation conjointe de plusieurs I-LDI.

Références

[1] J. Shade, S. Gortler, L. W. He, et R. Szeliski. Layered depth images. Dans *SIGGRAPH '98 : Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 231–242, New York, NY, USA, 1998. ACM.

[2] S.-U. Yoon, E.-K. Lee, S.-Y. Kim, et Y.-S. Ho. A framework for representation and processing of multi-view video using the concept of layered depth image. *Journal of VLSI Signal Processing Systems for Signal Image and Video Technology*, 46 :87–102, 2007.

[3] X. Cheng, L. Sun, et S. Yang. Generation of layered depth images from multi-view video. *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, 5 :V –225–V –228, 16 2007-Oct. 19 2007.

[4] S.-U. Yoon, E.-K. Lee, S.-Y. Kim, Y.-S. Ho, K. Yun, S. Cho, et N. Hur. Coding of layered depth images representing multiple viewpoint video. 2006.

[5] J. Duan et J. Li. Compression of the layered depth image. *Image Processing, IEEE Transactions on*, 12(3) :365–372, Mars 2003.

[6] P. Merkle, A. Smolic, K. Müller, et T. Wiegand. Efficient prediction structures for multiview video coding. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(11) :1461–1473, Novembre 2007.

[7] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, et Y. Mori. *Reference Softwares for Depth Estimation and View Synthesis*. ISO/IEC JTC1/SC29/WG11MPEG2008/M15377, April 2008.

[8] V. Jantet, L. Morin, et C. Guillemot. Incremental-ldi for multi-view coding. Dans *3DTV-Con2009*, Potsdam, Germany, Avril 2009.

[9] K. Müller, A. Smolic, K. Dix, P. Kauff, et T. Wiegand. Reliability-based generation and view synthesis in layered depth video. *Multimedia Signal Processing, 2008 IEEE 10th Workshop on*, pages 34–39, Oct. 2008.

[10] K. J. Oh, S. Yea, et Y. S. Ho. Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-d video. Dans *PCS'09 : Proceedings of the 27th conference on Picture Coding Symposium*, pages 233–236, Piscataway, NJ, USA, 2009. IEEE Press.

[11] L. McMillan. A list-priority rendering algorithm for redisplaying projected surfaces. Rapport technique, Chapel Hill, NC, USA, 1995.

[12] C.-L. Zitnick, S.-B. Kang, M. Uyttendaele, S. Winder, et R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. Graph.*, 23(3) :600–608, 2004.

Solution d'allocation de puissance efficace pour de la transmission vidéo scalable en environnement Outdoor réaliste

W. Hamidouche[1] C. Perrine[1] Y. Pousset[1] C. Olivier[1]

[1] Laboratoire XLIM, CNRS UMR 2172, Département SIC

Université de Poitiers Téléport 2 Bd Marie et Pierre Curie
BP 30179, 86962 Futuroscope Cedex (France)

{hamidouche, perrine, pousset, olivier}@sic.sp2mi.univ-poitiers.fr

Résumé

Dans ces travaux nous considérons la transmission d'information vidéo au format H.264/SVC dans un contexte de canaux MIMO réalistes tenant compte de techniques de précodage. Ainsi, d'une part, le standard H.264/SVC fournit une qualité de scalabilité spatio-temporelle de l'information vidéo. D'autre part, les précodeurs subdivisent le canal MIMO en sous canaux SISO indépendants associés à des TEB faibles. Ainsi, nous proposons un nouveau schéma exploitant à la fois la hiérarchie intrinsèque du standard H.264/SVC et les qualités de quatre précodeurs (Max-SNR, WF, QoS et E-d_{min}). Le schéma proposé est testé sur un canal MIMO réaliste. De plus, nous évaluons ses performances vis-à-vis des erreurs d'estimation du canal selon la norme IEEE802.11n.

Mots clefs

Allocation optimale de la bande passante, H.264/SVC, MIMO, techniques de précodage, précodeur QoS, précodeur E-d_{min}, OFDM, tracé de rayons 3D.

1 Introduction

Lors d'une communication sans fil, l'interaction des ondes électromagnétiques avec l'environnement ainsi que la mobilité des utilisateurs engendrent des affaiblissements importants et imprévisibles du signal reçu. Ainsi, la bande passante limitée du canal radio et sa nature instable représentent les éléments principaux à prendre en compte pour maintenir une bonne qualité de service des applications multimédias. Dans cet article nous nous intéressons plus particulièrement à la transmission d'un flux vidéo scalable sur un canal MIMO (Multiple Input Multiple Output). Au niveau de la couche application, nous utilisons le standard H.264/SVC [1] pour compresser la vidéo originale en un flux vidéo scalable et hiérarchisé.

Plusieurs auteurs se sont intéressés à la transmission de vidéos scalables sur un canal MIMO [2]-[5]. Parmi ces solutions, nous nous focalisons sur celles qui considèrent la connaissance de l'information du canal CSI (Channel State Information) au niveau de l'émetteur (Tx-CSI) [3]-[5], notamment appelées systèmes à boucle fermée. Un schéma d'allocation de puissance optimale à été proposé dans [3] minimisant la distorsion du flux vidéo scalable MPEG-4. Cette solution considère la connaissance parfaite de Tx-CSI ainsi que l'information sur le débit-distorsion de la version scalable du codeur vidéo MPEG-4. Un autre schéma de sélection adaptative des sous canaux SISO (Single Input Single Output) pour la transmission de flux vidéo H.264/SVC a été proposé dans [4]. Cette solution, basée sur l'information partielle du canal résumée par le Rapport Signal à Bruit (RSB), affecte chaque flux vidéo, suivant son importance, au sous canal SISO correspondant. Ce schéma offre une stratégie de protection inégale contre les erreurs de transmission aux flux vidéo sans aucune redondance ou une stratégie d'allocation de puissance. Les mêmes auteurs ont proposé par la suite dans [5] un nouveau schéma qui combine les travaux dans [4], le précodeur Water Filing (WF) et une technique de ré-allocation de puissance. Cette combinaison permet d'une part une transmission sans erreur du flux vidéo de qualité de base, et d'autre part, une transmission des flux d'amélioration de qualité avec une modulation de haute efficacité spectrale.

Les travaux présentés précédemment utilisent des techniques d'allocation de puissance standards, telle que le précodeur WF qui est largement surpassé par les nouveaux précodeurs en terme de Taux d'Erreurs Binaire (TEB) et de flexibilité dans le processus d'allocation de puissance. De plus, les stratégies d'allocation de puissances utilisées ne sont pas finement adaptées à l'importance des flux vidéo, tel que le précodeur WF qui plutôt maximise la capacité du canal. Enfin, ces schémas ne sont testés que sur des canaux statistiques, alors que les spécificités environnementales et la corrélation des antennes MIMO ont un grand impact sur la qualité des liens MIMO [9].

Dans ce papier, nous proposons un nouveau schéma de transmission temps réel de vidéos codées H.264/SVC sur un canal MIMO en utilisant quatre précodeurs, à savoir Max-SNR [6], WF [7], QoS [7] et E-d_{min} [8]. Ces précodeurs décomposent le canal MIMO en sous canaux SISO indépendants et parallèles de différentes impor-

tances, réduisant ainsi la complexité du décodage de Maximum par Vraisemblance (MV) et maintiennent des valeurs de TEB relativement faibles. Les spécificités communes à ces précodeurs leur permettent d'être très adaptées à une transmission temps réel d'un flux vidéo hiérarchisé.
Nous exploitons la flexibilité du précodeur QoS afin d'adapter finement l'allocation de puissance suivant l'importance des flux vidéo et l'état instantané du canal de transmission. Les performances du schéma proposé sont évaluées sur un canal MIMO réaliste. Nous utilisons un modèle de propagation déterministe [10] qui prend en compte toutes les spécificités environnementales (géométrique, électrique) d'un environnement réel fournissant en sortie la Réponse Impulsionnelle Complexe (RIC) correspondante. De plus, nous évaluons la robustesse de ce schéma vis-à-vis des erreurs d'estimation du canal (EC) suivant la norme IEEE802.11n [11].

2 Description du système

2.1 Standard H.264/SVC pour une transmission sans fil

La norme H.264/SVC est la version scalable du standard de compression de vidéos H.264/SVC. La principale attente de cette nouvelle norme est de supporter les scalabilités spatiale, temporelle et en résolution tout en maintenant l'efficacité de compression de la norme H.264/AVC. Ainsi, H.264/AVC conserve la même structure de codage que le standard H.264/AVC, à savoir un codage en deux couches : Video Coding Layer (VCL) et Network Abstraction Layer (NAL). La scalabilité temporelle est efficacement établie en utilisant le concept hiérarchique d'images bidirectionnelles (image-B). La scalabilité spatiale est effectuée en utilisant le mécanisme de prédiction entre les différentes couches. Quant à la scalabilité en résolution, elle est assurée par le concept général d'un codage spatial appelé dans le standard H.264/SVC Meduim Grain quality Scalability (MGS). Pour plus détails sur le standard H.264/SVC le lecteur pourra se référer à [1].
Le codec H.246/SVC encode la vidéo originale en plusieurs couches de résolutions temporelles. Le codage de scalabilité en résolution est appliqué sur chacune des couches temporelles afin de fournir les couches de scalabilité en qualité (raffinement ou amélioration). Dans ce papier, le codeur source fournit N_T couches de résolution temporelle et N_Q couches de résolution en qualité. Suivant une stratégie de transmission, ces couches constitueront les N flux vidéo à transmettre, définis dans ce papier par $(\rho_0, \rho_1, ..., \rho_{(N-1)})$.
Au niveau de la couche physique, les N flux vidéo constituant les paquets NAL seront transmis dans des trames IEEE802.11n de taille constante. Tout paquet vidéo NAL reçu dans une trame erronée sera perdu. Ces pertes de paquets se traduisent soit par des dégradations temporelles ou en résolutions de la vidéo reçue. Afin de faire face aux pertes d'images causées par la perte des couches de basse résolution, quatre méthodes de robustesses proposées dans [2] ont été implémentées. Ces dernières permettent de remplacer les images perdues par des images correctement reçues appartenant au même GOP (Group Of Pictures). Parmi ces méthodes, on utilise celle qui offre les meilleures performances en terme de qualité des vidéos reçues. La méthode sélectionnée remplace l'image perdue par l'image suivante correctement reçue. Enfin, la distorsion des vidéos reçues est mesurée par le Peak Signal to Noise Ratio (PSNR) moyen de la composante luminance et des deux composantes de chrominance.

2.2 Précodeurs dans un contexte de transmission vidéo

Dans ce papier nous étudions la transmission de vidéos H.264/SVC sur un canal MIMO impliquant trois précodeurs diagonaux (Max-SNR, WF, QoS) et un précodeur non diagonal E-d_{min}. Ces précodeurs nécessitent la connaissance parfaite du Tx-CSI et permettent d'optimiser un critère propre à chaque précodeur. Par exemple, les précodeurs Max-SNR, WF et E-d_{min} optimisent respectivement le RSB, la capacité du canal et la distance euclidienne. Quant au précodeur QoS il permet de maintenir tout rapport de RSB entre les sous canaux SISO. On définit un système MIMO avec n_T antennes à l'émission et n_R antennes à la réception, noté système MIMO $(n_T \times n_R)$. L'équation correspondante à ce système est donnée par

$$y = GHFs + Gn \tag{1}$$

où s et y représentent respectivement les vecteurs émis et reçus de taille $(b \times 1)$ avec $b = trace(H) < min(n_R \times n_T)$, H la matrice du canal de taille $(n_R \times n_T)$, F la matrice de précodage de taille $(n_T \times b)$, G la matrice de décodage de taille $(b \times n_R)$ et n le vecteur de bruit additif de taille $(n_R \times 1)$.
L'étape commune à ces précodeurs est appelée transformation virtuelle. Cette étape permet de diagonaliser le canal et blanchir le bruit. Elle est généralement réalisée par la méthode de décomposition en valeurs singulières. Après la transformation virtuelle du canal, le système MIMO s'écrit sous la forme suivante

$$y = G_d H_v F_d s + G_d n_v \tag{2}$$

où F_d et G_d sont respectivement les matrices de précodage et de décodage, $H_v = G_v H F_v$ est la matrice des valeurs singulières, et $n_v = G_d v_n$ est le vecteur de bruit de covariance $R_{nv} = I_b$ (I_b est la matrice identité de taille b).
La puissance du système MIMO doit être limitée à la puissance totale d'émission E_T

$$\|F_d\|_F^2 = E_T \tag{3}$$

avec $||.||_F$ est la norme de Frobeniuos. Les matrices G_v et F_v sont unitaires et sélectionnées afin de blanchir le bruit, diagonaliser le canal et réduire sa dimension à b. Ainsi, le canal MIMO est décomposé en b sous canaux SISO

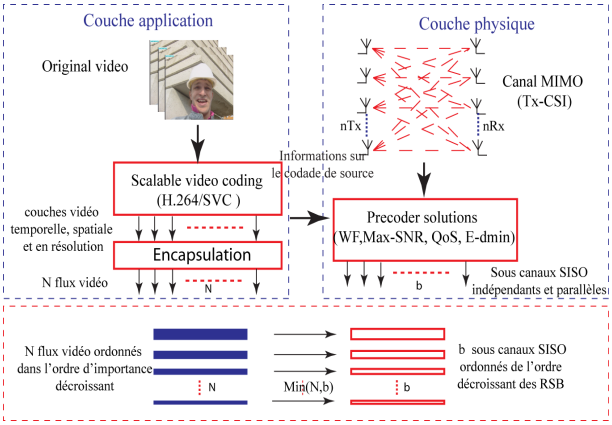


Figure 1 – *Stratégie de transmission temps réel du flux vidéo H.264/SVC entre la couche application et la couche physique*

indépendants et parallèles représentés par la matrice diagonal H_v

$$H_v = diag(\sigma_0, \sigma_1, ..., \sigma_{b-1}) \quad (4)$$

H_v est appelé la matrice du canal virtuel, et la valeur σ_i^2 représente le RSB du sous canal i. Le gain total (σ) d'un système MIMO $(n_T \times n_R)$ est calculé par

$$\sigma = \sqrt{\sum_{i=0}^{b-1} \sigma_i^2} \quad (5)$$

Les précodeurs diagonaux sont définis par leurs matrices diagonales F_d et G_d. Quant au précodeur $E-d_{min}$ ses matrices de précodage et décodage ne sont plus diagonales. Ce précodeur offre les meilleures performances en terme de TEB comparé aux précodeurs diagonaux [8]. Le précodeur QoS permet une grande flexibilité dans l'allocation de puissance aux sous canaux SISO, et offre le meilleur compromis entre complexité, performance de TEB et flexibilité [7]-[8]. Ainsi, il nous parait le plus approprié pour une transmission temps réel d'un flux vidéo scalable.

2.3 Stratégie de transmission proposée

Le schéma proposé dans ce papier est composé de deux couches : couche application et couche physique. Les flux vidéo (ρ_i) alimentent la couche physique qui traite chaque flux vidéo d'une manière spécifique selon son importance. Le concept général de notre couche physique est similaire à celui de la couche physique de la norme IEEE802.11n. Les précodeurs présentés dans la section précédente sont implémentés, testés puis intégrés dans la couche physique IEEE802.11n. On utilise les codes correcteurs d'erreurs Low Density Parity Check (LDPC) pour protéger les flux vidéo avec une stratégie EEP (Equal Error Protection). Ainsi, les flux vidéos sont traités par tous les blocs de la chaîne de transmission IEEE802.11n offrant en sortie des trames de taille constante prêtes à la transmission. Au niveau du récepteur, un algorithme basé MV est utilisé pour décoder les symboles reçus, qui sont par la suite, démodulés, dé-entrelacés et corrigés par le code correcteur d'erreurs LDPC. Après le processus de correction, tous les paquets vidéo NAL transportés sur des trames contenant des erreurs résiduelles seront perdus. Le reste des paquets vidéo, reçus sans erreur, alimentent la couche application pour la reconstruction de la vidéo reçue.

Dans ce papier nous adoptons une stratégie de transmission efficace entre le codeur H.264/SVC et les précodeurs. D'une part, au niveau de la couche application, le codeur H.264/SVC fournit N flux vidéo $(\rho_0, \rho_1, ..., \rho_{(N-1)})$ ordonnés dans un ordre d'importance décroissant. D'autre part, au niveau de la couche physique, les précodeurs décomposent le canal MIMO en sous canaux SISO ordonnés par ordre de RSB décroissant $(f_0^2\sigma_0^2, f_1^2\sigma_1^2, ..., f_{b-1}^2\sigma_{b-1}^2)$ avec f_i le coefficient de précodage affecté au sous canal SISO i. Ainsi, la stratégie de transmission proposée consiste à affecter les flux vidéos aux sous canaux SISO en considérant une association fine entre l'importance du flux vidéo et la puissance qui lui est allouée par le précodeur. En effet, le flux vidéo (ρ_i) est transmis sur le sous canal SISO de RSB égal à $f_i^2\sigma_i^2$. La figure 1 illustre la stratégie de transmission adoptée entre la couche application et la couche physique.

3 Résultats de simulation

3.1 Contexte de simulation

Confi.	Zone 1	Zone 2	Zone 3
état du canal	Mauvais	Moyen	Bon
RSB relatives	(0.7, 0.3, 0, 0)	(0.4, 0.3, 0.3, 0)	(0.3, 0.3, 0.25, 0.15)

Tableau 1 – *Valeurs relatives de RSB utilisées par le précodeur QoS à chaque état de canal*

Dans cette section nous présentons les résultats de simulations afin de montrer les performances du schéma proposé dans un environnement réaliste Outdoor. Nous utilisons un modèle de propagation déterministe 3-D [10], développé au laboratoire XLIM-SIC, pour calculer les RIC d'un trajet de 180 m sur le campus de l'Université de Poitiers. L'émetteur reste statique et le récepteur se déplace sur la trajectoire avec une vitesse de 10 m/s. Les paramètres utilisés au niveau de chaque couche sont présentés ci-dessous :

1. La couche application : on utilise la séquence vidéo de test Foreman au format QCIF 176×144 pixels/image. Les 288 premières images de la vidéo sont codées avec le codec H.264/SVC à 30 images/seconde. Le codeur de vidéo est paramétré pour fournir 4 couches de résolutions temporelles $(N_T = 4, GOP = 8)$ et 4 couches de résolutions en qualité $(N_Q = 4)$. Ici, les 4 couches en résolution qualité forment quatre flux vidéo de même taille à transmettre $(N = 4)$ $(\rho_0, \rho_1, \rho_2, \rho_3)$ Ainsi, le flux ρ_i est exploitable seulement si tous les flux ρ_j d'identifiant $j < i$ sont correctement reçus.
2. La couche physique : les paramètres de transmis-

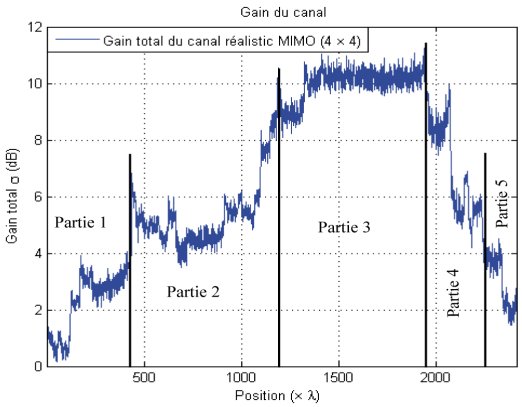


Figure 2 – *Variation du gain en fonction de la position sur le trajet*

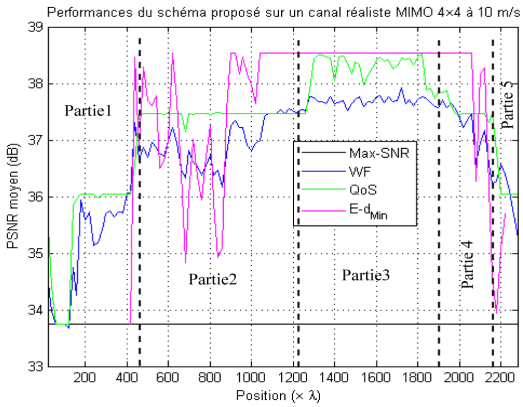


Figure 3 – *Performance du schéma proposé sur un canal MIMO* (4×4) *réaliste*

sion (puissance d'émission, fréquence porteuse, système OFDM) sont fixés suivant la norme IEEE802.11n. Nous considérons un canal MIMO (4×4). La distance entre les antennes est fixée à 0.4λ, ou λ représente la fréquence porteuse. Cela correspond généralement à la corrélation minimale entres les antennes [9]. Nous considérons la modulation 4-QAM pour tous les flux vidéo, qui sont protégés par le code correcteur LDPC de rendement $1/2$. Cette configuration de couche physique nous permet de transmettre à un débit utile de $12 \times b \;\; Mb/s$.
Suivant l'état du canal de transmission, nous considérons trois configurations du précodeur QoS. Ces configurations tiennent en compte, d'une part, de l'état instantané du canal en sélectionnant uniquement un sous ensemble des canaux SISO, et d'autre part, de l'importance des flux vidéo en adoptant une stratégie UEP. Les valeurs relatives de RSB utilisées à chaque état de canal sont données dans le Tableau 1, suivant trois zones sur lesquelles nous revenons dans le section suivante. Nous voulons notifier que cette configuration est choisie expérimentalement parmi d'autres configurations moins performantes, et qu'elle ne représente en aucun cas la solution optimale.

3.2 Résultats et discussions

Le gain total σ du canal réaliste MIMO (4×4) en fonction de la position du récepteur est illustré sur la figure 2. Suivant les valeurs du gain total, on peut classer la trajectoire du récepteur en trois zones d'études. La première zone représente un canal de mauvaise qualité et couvre les parties 1 et 5 de la trajectoire. La zone 2 représente un canal de qualité moyenne et couvre les parties 2 et 4 de la trajectoire. Quant à la partie 3 de la trajectoire, elle représente un canal de bonne qualité et correspond à la zone 3. La figure 3 illustre les performances du schéma proposé le long de la trajectoire du récepteur. Ces courbes montrent la variation du PSNR des vidéos reçues en fonction du déplacement du récepteur sur la trajectoire pour les différents précodeurs utilisés. Nous constatons les hautes performances du précodeur QoS en utilisant la configuration appropriée à chaque zone du canal (Tableau 1). Le précodeur Max-SNR quant à lui assure une réception sans erreur du flux vidéo de qualité de base et cela quelque soient les conditions du canal. Nous pouvons aussi remarquer l'adaptation des précodeurs WF et QoS aux variations du canal de transmission. En effet, la qualité des vidéos reçues en utilisant ces deux précodeurs est améliorée suivant l'augmentation de la puissance du signal reçue. Dans de mauvaises conditions du canal (zone 1), le précodeur E-d_{min} n'arrive même pas à assurer une bonne réception du flux vidéo de qualité de base, alors qu'en zone 3 ce précodeur offre les meilleures performances en recevant correctement les quatre flux vidéos. Nous constatons aussi que les précodeurs adoptants une stratégie UEP, tels que QoS et WF assurent en moyenne les meilleures performances. De plus, une considération plus fine du contenu de l'information transmise améliore davantage les performances du précodeur QoS par rapport à celles du précodeur WF. Le gain en terme de qualité visuelle apportée par le précodeur QoS par rapport au précodeur WF est illustré sur les figures 4(a) et 4(b). Cette figure montre clairement que la qualité visuelle de la vidéo reçue avec le précodeur QoS est nettement meilleure à celle reçue avec le précodeur WF.
Nou pouvons aussi constater sur la figure 3 que les performances du précodeur E-d_{min} ne sont pas calculées dans la zone1. Pour cette raison, ses performances moyennes sur la trajectoire ne peuvent être objectivement comparées avec les trois autres précodeurs (diagonaux). Ainsi, sur les Figures 5(a) et 5(b) nous illustrons séparément à la fois les performances moyennes du schéma proposé sur la trajectoire ainsi que sa robustesse vis-à-vis des erreurs d'EC pour respectivement les précodeurs diagonaux et le précodeur non-diagonal E-d_{min}. La figure 5(a) montre que le précodeur QoS assure en moyenne les meilleures performances comparées aux autres precodeurs diagonaux. De plus, on remarque que tous les précodeurs diagonaux restent très robustes aux erreurs d'EC et maintiennent les mêmes performances avec et sans erreurs d'estimation. Cependant, sur la 5(b) nous remarquons que le précodeur E-

d_{min} devient très sensible aux erreurs d'EC. En effet, les erreurs d'EC à une vitesse de 10 m/s entrainent une perte entre 1 et 2 dB en termes de PSNR sur les vidéos reçues.

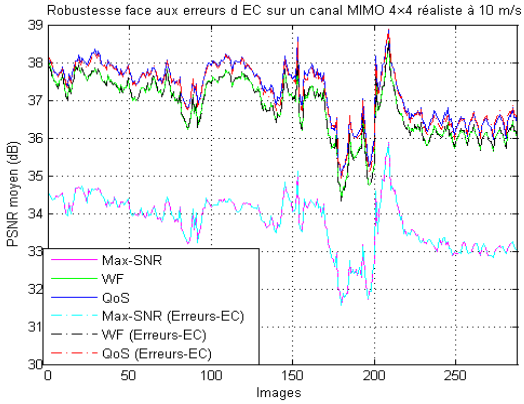
(a) PSNR = 35.67 dB

(b) PSNR = 34.11 dB

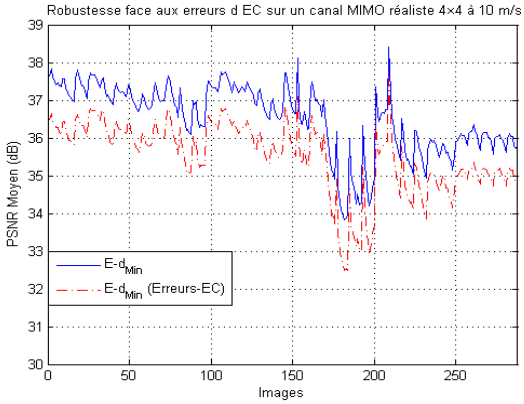
Figure 4 – *Illustration de la qualité visuelle des vidéos reçues par les précodeurs QoS (haut) et WF (bas) sur un canal MIMO réaliste (zone 1)*

4 Conclusion

Dans ce papier nous avons développé une solution prometteuse pour la transmission temps réel d'un flux vidéo H.264/SVC sur un canal MIMO. La hiérarchie fournie par le codeur vidéo H.264/SVC est efficacement exploitée par quatre précodeurs. Nous avons montré que les précodeurs les plus performants en terme de TEB (E-d_{min}) ne sont pas nécessairement les plus appropriés pour une transmission de vidéo. Cependant, une meilleure prise en compte du contenu des flux vidéo (QoS) permet de garantir la meilleure qualité de service. De plus, contrairement au précodeur E-d_{min}, le precodeur QoS reste très robuste aux erreurs d'EC même avec des vitesses de déplacement relativement élevées de l'ordre de 10 m/s.
Ces travaux montrent clairement qu'une adaptation adéquate de l'allocation de puissance en fonction de l'importance du flux vidéo et l'état instantané du canal améliore la qualité de service. Ainsi, dans la suite de nos travaux nous développerons un précodeur QoS adaptatif qui minimise la distorsion de la vidéo reçue. En se basant sur un critère pertinent, cette solution considérera conjointement les couches application et physique et permettra d'atteindre la solution d'allocation de puissance optimale quelque soient les conditions du canal.



(a) Précodeurs diagonaux



(b) Précodeur non-diagonal

Figure 5 – *Robustesse du schéma proposé vis-à-vis les erreurs d'EC sur un canal réaliste MIMO* (4×4) *à 10 m/s*

Remerciement

Ce travail entre dans le cadre du projet ANR CAIMAN.

Références

[1] T. Wiegand, G. Sullivan, J. Reichel, H. Schwarz, M. Wien, *Joint Draft of SVC Amendment*, Joint Video Team (JVT), Doc. JVT-W201, San Jose, CA, 2007.

[2] M. K. Jubran, M. Bansal, L. P. Kondi and R. Grover, *Accurate Distortion Estimation and Optimal Bandwidth Allocation for Scalable H.264 video transmission Over MIMO systems*, IEEE Trans. Image Processing, Vol. 10, no. 1 pp. 106-116, Jan. 2009.

[3] Z. Ji, Q. Zhang, W. Zhu, Z. Guo, J. Lu, *Power efficient MPEG-4 FGS video transmission over MIMO-OFDM*

systems, Proc. IEEE ICC, Vol. 5, pp. 3398-3402, May 2003.

[4] D. Song, C. W. Chen, *Scalable H.264/AVC Video Transmission over MIMO Wireless Systems with adaptive channel selection based on partial channel information*, IEEE Trans. Circuits Syst. Video Technol., Vol. 17, no. 9, pp. 1218-1226, Sep. 2007.

[5] D. Song, C. W. Chen, *Maximum-throughput delivery of SVC-based video over MIMO systems with times-varying channel capacity*, Elsevier Journal of Visual Commun. And Image Represent Vol. 19, no. 8, pp. 520-528, Dec. 2008.

[6] P. Stoica and G. Ganesan, *Maximum-SNR spatial-temporal formatting designs for MIMO channels*, IEEE Trans. Signal Processing, vol. 50, no. 12, pp. 3036-3042, Dec. 2002.

[7] H. Sampath, P. Stoica, and A. Paulraj, *Generalized Linear Precoder and Decoder Design for MIMO Channels Using the Weighted MMSE Criterion*, IEEE Trans. Commun. Vol. 49, no. 12, pp. 2198-2206, Dec. 2001.

[8] B. Vrigneau, J. Letessier, P. Rostaing, L. Collin and G. Burel, *Extension of the MIMO Precoder based on the Minimum Euclidean Distance : a cross-form matrix*, IEEE Signal Processing Vol. 2, no. 2, pp. 135-146, May 2008.

[9] C. Pereira, Y. Pousset, R. Vauzelle, P. Combeau, *Sensitivity of the MIMO Channel Characterization to the Modeling of the Environment*, IEEE Transactions on Antennas and Propagation, Vol. 57, no. 4, pp. 1218-1227, Mar. 2009.

[10] F. Mora, L. Aveneau, *Optimized scanning of a visibility graph data structure for efficient ray-tracing*, ECWT, Paris, Oct. 2005.

[11] IEEE Standard for Information Technology-Part 11 : *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment : Enhancements for Higher Throughput (802.11n)*, 2009.

Affiches

Tatouage d'image performant utilisant la Quantification Codée Treillis dans un domaine indépendant

C. DELPHA[1] I. BENKARA[2] S. BRACI[1] R. BOYER[1] M. KHAMADJA[2]

[1]Laboratoire des Signaux et Systèmes (L2S),

CNRS / Supelec / Université Paris Sud 11,
Plateau du Moulon, 3, rue Joliot Curie, 91192 Gif sur Yvette – FRANCE

[2]Laboratoire de Traitement du Signal (LTS),

Département d'électronique et de l'ingénieur, 3, Rue Joliot Curie,
Université Mentouri, Route Ain El Bey, 25000 Constantine – ALGERIE

{delpha, sbraci, rboyer}@lss.supelec.fr, {ilhem_bm, m_khamadja}@yahoo.fr

Résumé

Les besoins de protection pour la diffusion sécurisée des images ont permis de mettre en avant les solutions de tatouage numérique dans le domaine du data hiding. Dans ce domaine l'idée principale est d'insérer une information dans un document numérique en s'assurant des propriétés de robustesse, de capacité et d'invisibilité (perceptuelle et statistique). De récentes études ont montré que pour des raisons de sécurité l'invisibilité statistique était un critère très important rarement évalué. Nous proposons ici d'étudier un schéma de tatouage à information adjacente basé sur une quantification codée en treillis (TCQ) pour lequel nous ferons une insertion dans un domaine indépendant pour renforcer à la fois les propriétés de robustesse et d'indiscernabilité statistique du schéma. Nous montrons, en comparaison avec les systèmes référents, que la robustesse est nettement améliorée avec la solution proposée sans réduire de façon drastique la capacité (quantité d'information à transmettre sans erreur pour un niveau de bruit donné). Puis, grace à une étude des fonctions de densité de probabilité prouvons que le niveau de sécurité du schéma proposé est également amélioré grâce à l'utilisation de l'Analyse en Composantes Indépendantes (ICA).

Mots clefs

Digital watermarking, SCS, TCQ, ICA.

1 Introduction

Ces dernières années, de nombreuses méthodes de dissimulation d'information ont été étudiées pour des applications du tatouage numérique. Parmi les plus performantes d'entre elles, on peut citer notamment les méthodes basées sur l'approche de Costa [1] souvent combinées à des méthodes de transformation [2]. Elles offrent généralement, dans des configurations de schéma de tatouage avec une clef privée, de bonnes performances en termes de robustesse à des attaques. Des études ont toutefois déjà mises en évidence certaines faiblesses quant à la sécurité de ces schémas par exemple avec l'utilisation de l'Analyse en Composante Indépendantes (ICA) pour séparer signal hôte et message pour ainsi permettre l'estimation du message [3].
Initialement, l'ICA a été proposée pour résoudre des problèmes de séparation aveugle de sources dans de nombreux domaines du traitement du signal [4]. Les propriétés majeures de cette méthode sont la maximisation de l'information et la minimisation de la distorsion induite en décomposant les données en sources indépendantes.

Dans certaines études afférentes au tatouage numérique, l'utilisation de l'ICA et ses propriétés a été proposée aussi bien lors de la phase d'insertion que d'extraction de l'information à cacher. Des travaux tels que ceux de González-Serrano et al. [5] ou encore J.J. Murillo-Fuentes [6] ont décrit une méthode modifiant les composantes indépendantes d'un signal pour effectuer l'opération de marquage en s'appuyant essentiellement sur des méthodes basées sur l'étalement de spectre. On peut aussi citer les travaux de Bounkong et al. [7], mais aussi ceux de Benkara et al [8] dans lesquels sont proposés le marquage par quantification de coefficients ICA pour des images en utilisant le principe du schéma à information adjacente à l'encodeur. Dans tous ces travaux le critère essentiel qui a été évalué est la robustesse aux attaques.
Pour ce travail, nous nous intéresserons à la sécurité telle que définie par Cachin [9] du schéma à information adjacente. Certaines études ont montré l'effet bénéfique d'une quantification codée treillis (TCQ) sur l'indiscernabilité statistique [10]. Après avoir rappelé les propriétés du schéma de tatouage par TCQ, nous voulons nous appuyer sur l'ICA aussi bien à l'insertion qu'à l'extraction

de l'information cachée et ainsi montrer l'effet bénéfique de l'ICA sur la sécurité du schéma TCQ-ICA. Nous montrons également l'amélioration des performances de capacité et de robustesse par rapport au schéma initial. Une comparaison avec les schémas offrant les meilleures performances de capacité-robustesse tels que le Scalar Costa Scheme (SCS) et le bien connu Spread-Transform Scalar Costa Scheme (ST-SCS) est proposée pour montrer les bonnes propriétés des schémas.

2 Tatouage informé dans un domaine indépendant

Dans cette partie, nous présentons le schéma de tatouage basé sur la quantification utilisant l'analyse en composantes indépendantes (ICA) à l'insertion et à l'extraction du message (Figure 1). L'ICA est une technique statistique dont l'objectif est de décomposer un vecteur $X \in \Re^m$ en une combinaison linéaire de sources indépendantes, c'est-à-dire $X = A \times \hat{X}$, où $\hat{X}$ est un signal à composantes indépendantes et A une matrice à coefficients réels notée matrice de mélange. Cette technique est appliquée à des problèmes où les sources peuvent être supposées indépendantes et pour lesquelles il est possible de trouver une matrice B de séparation des composantes telle que le vecteur $\hat{X}$ obtenu par action de B sur X ait des composantes les plus indépendantes possibles.

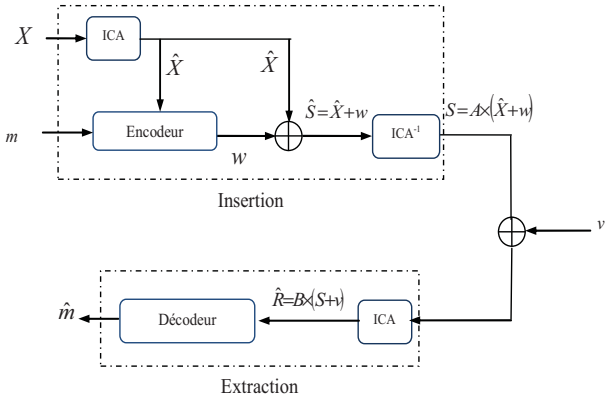


Figure 1 – *Schéma de tatouage informé exploitant les propriétés de l'ICA*

La procédure de tatouage proposée dans un domaine indépendant s'appuie sur les schémas basés sur la quantification et peut être décrite selon les principales étapes suivantes :

1. L'algorithme ICA est appliqué au signal hôte X pour obtenir les composantes indépendantes $\hat{X}$.
2. Les sources $\hat{X}$ obtenues sont alors quantifiées et tatouées, en utilisant le principe de Costa [1] pour obtenir les composantes indépendantes tatouées et marquées $\hat{S} = \hat{X} + w$.
3. Une multiplication des composantes indépendantes tatouées par la matrice de mélange A (transformation ICA inverse), nous permet de récupérer les données tatouées $S = A \times (\hat{X} + w)$ dans le domaine original du signal hôte.
4. Si on suppose que les données tatouées S sont transmises par l'intermédiaire d'un canal de communication pouvant introduire des distorsions qui seront modélisées par un bruit d'attaque V (on s'intéressera ici essentiellement a un bruit AWGN), le signal reçu $R = S + V$ sera multiplié par la matrice de séparation B pour produire les composantes indépendantes tatouées attaquées $\hat{R} = B \times (S + V)$.
5. Le décodage du message $\hat{m}$ se fait sans connaissance du signal hôte en utilisant le principe de Costa selon la méthode de quantification utilisée.

Pour notre étude, la matrice de séparation B sera considérée comme secrète pour rendre possible le décodage du message. En effet en s'appuyant sur une méthode de quantification scalaire uniforme (Schéma Scalaire de Costa) associé à l'ICA (schéma noté SCS-ICA), nous pouvons montrer que la parfaite connaissance de la matrice B est nécessaire pour le bon décodage du message $\hat{m}$ après ajout de bruit.
Comme le montre la figure 2, lorsque la matrice B est maintenue secrète, et donc partagée entre la phase d'insertion et la phase d'extraction, le message $\hat{m}$ peut être décodé avec un taux d'erreur binaire (BER) faible même lorsque le rapport marque à bruit (WNR) est important (faible bruit).

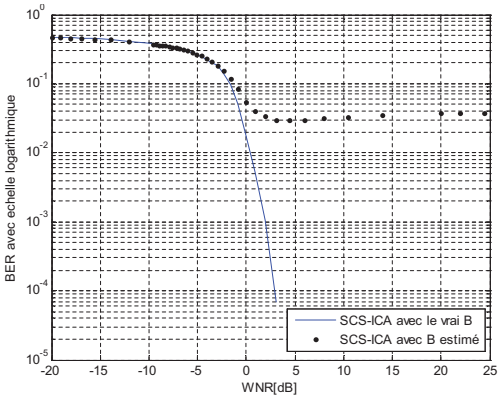


Figure 2 – *Importante de l'exactitude de la matrice de séparation B au décodage*

En cas d'absence de connaissance de la matrice B nous pouvons calculer une matrice $\tilde{B}$ permettant d'effectuer la séparation en composantes indépendante du signal reçu R. Si on considère $\tilde{B}$ comme une estimation de la matrice B, on voit que cette estimation n'est pas suffisante pour permettre le décodage du message avec un BER faible malgré un WNR élevé. Ceci laisse penser que la sécurité au sens de la fuite de l'information liée à l'estimation de la clé secrète sera renforcée sans une parfaite connaissance de B.

3 Tatouage utilisant la TCQ et l'ICA

3.1 Quantification par TCQ en Tatouage

La Quantification Codée en Treillis est une méthode de quantification qui utilise un dictionnaire (alphabet) structuré et qui a été appliquée au tatouage spécialement pour les méthodes basées sur la quantification. Inspirée de la TCM (Trellis Coded Modulation) [11], cette méthode était généralement utilisée pour réduire la complexité et les distorsions d'un système.
Cette technique de quantification est basée sur l'idée du partitionement proposée par Underboëck [12]. Elle consiste à partitioner un dictionnaire initial (dictionnaire structuré) en sous dictionnaires complémentaires de même taille associés aux transitions entre les différents états d'un code convolutif. Ainsi pour coder une séquence d'échantillons avec la TCQ, l'algorithm de Viterbi est utilisé [13] pour obtenir les transitions, associées au code convolutif, qui minimisent la distortion. Chaque transition dans le treillis est codé sur b bits. Ainsi, les $n-b$ bits résultants sont utilisés pour indexer le mot de code dans le sous-dictionnaire choisi associé à la transition.
La figure suivante représente un treillis avec 4 sous dictionnaires D_0, D_1, D_2, D_3.

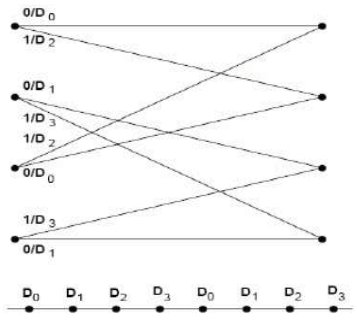


Figure 3 – *Treillis 4 états avec 4 sous dictionnaires*

Pour les applications en tatouage, une quantification avec la TCQ a été proposée pour éviter un partitionnement régulier généralement obtenu avec la quantification scalaire. Ce type de quantificateur est préféré pour améliorer le partitionnement du dictionnaire et ainsi améliorer l'invisibilité statistique des distorsions introduites sur la fonction de densité de probabilité du signal marqué [14].
Pour appliquer la TCQ au watermarking, les chemins du treillis sont fixés par les valeurs des bits du message. Les échantillons du signal hôte sont quantifiés en utilisant le sous dictionnaire correspondant aux transitions prises. Ainsi, le taux d'insertion est fixé à un bit par échantillons. Cette approche la plus courante dans les applications de tatouage est notée TCQ-PS (TCQ-Path selection) [15]. Elle peut être décrite selon la méthode qui suit.
Considérons un treillis défini par la fonction de transition :

$$E \times \{0,1\} \rightarrow e$$

$$t : (e_i, m[i]) \rightarrow e_{i+1}$$

tel que $E = \{0, 1, ...2^{r-1}\}$ corresponds à l'ensemble des états possibles du treillis.
Ainsi, la distortion relative à l'insertion de la marque dépends de l'état précédent du treillis et du symbole inséré :

$$E \times \{0,1\} \rightarrow [-\frac{\Delta}{2}, \frac{\Delta}{2}]$$

$$o : (e_i, m[i]) \rightarrow d[i])$$

Donc, le sous dictionnaire peut-être noté avec l'équation suivante :

$$U_m[i] = \{\breve{\kappa}\Delta + o(\hat{x}_i, m[i], \breve{\kappa} \in Z\}$$

avec $\breve{\kappa}$ la clé secrète utilisée à l'encodeur et au décodeur.
La variable auxiliaire $u^* \in U_m$ qui est plus proche de l'échantillon du signal $\hat{x}$ est calculée avec l'algorithme de Viterbi tel que :

$$u^* = \arg \min_{u \in U_m} \sum_{i=1}^{N} (\hat{x}[i] - u[i])^2$$

Pendant la phase de décodage, le signal reçu est à nouveau quantifié et l'algorithme de Viterbi est utilisé pour trouver les meilleurs chemins et transitions pour lesquelles nous sommes capables d'extraire le message inséré.

3.2 Application dans le domaine ICA

Pour l'étude qui nous intéresse, c'est cette procédure d'utilisation de la TCQ qui a été appliquée. Le mot de code u^* le plus proche de l'échantillon $\hat{x}$ après décomposition en composantes indépendantes est calculé en utilisant l'algorithme de Viterbi. On obtient alors le signal marqué S tel que $S = A \times (\hat{X} + \alpha(u^* - \hat{X})$. La marque est alors $w = \alpha(u^* - \hat{X})$ de telle sorte que le paramètre d'optimisation α est tel que celui défini dans le schéma idéal de Costa : $\alpha = \frac{\sigma_w^2}{\sigma_w^2 + \sigma_v^2}$, on notera σ_w^2 la variance de la marque et σ_v^2 la variance du bruit.
Pour appliquer cet algorithme aux images, une phase initiale de pré-traitement de l'image est nécessaire. Elle consiste à transformer cette image comme le montre la figure 4 pour que l'application de l'ICA soit possible.

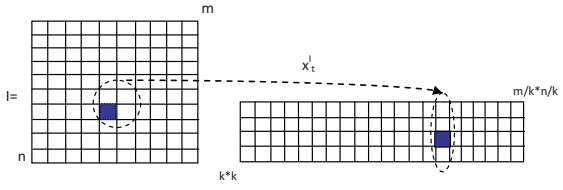


Figure 4 – *Etapes de pré-traitement de l'image à tatouer pour l'application de l'ICA - Exemple pour* $k = 2$

Considérons une image originale à tatouer I de taille $M \times N$, cette phase se décrit selon les étapes suivantes :

1. L'image I à tatouer est d'abord divisée en blocs $C_{p,q}$ de taille $k \times k$ pour lesquels la gamme de variation des indices est : $i, j = \{1, 2, ..., k\}$; $p = \{1, 2, ..., N/k\}$ et $q = \{1, 2, ..., M/k\}$. Donc nous pouvons noter les blocs $C_{p,q}(i, j) = I(k(p-1)+i, k(q-1)+j)$.
2. Les blocs $C_{p,q}$ sont alors transformés en vecteurs x_t^I pour lesquels $t = (p-1)M/k + q$, tels que $t = \{1, 2, ..., MN/k^2\}$. Nous obtenons alors un vecteur $x(k(i-1)+j)_{(M(p-1)/k+q)} = C_{p,q}(i, j)$. Cette transformation sera notée $\gamma(.)$ i.e $x_t^I = \gamma(I, k)$.
3. Les lignes sont alors projetées dans $l = k^2$ composantes indépendantes en utilisant l'algorithme FastICA [4], $\hat{x}_t^I = B \cdot x_t^I$.

C'est alors sur ces composantes indépendantes de l'Image que nous appliquons ensuite notre algorithme de tatouage basé sur la quantification avec la TCQ. Pour la suite nous noterons cet algorithme TCQ-ICA.

4 Résultats et Discussion

Cet algorithme de tatouage a été appliqué aux images et l'évaluation de ses performances en termes de robustesse, de capacité et d'invisibilité statistique a été effectuée. Nous avons utilisé pour notre étude une base d'image en niveau de gris de taille 512×512. Pour garantir l'invisibilité perceptuelle lors de l'insertion, nous avons choisi d'utiliser un rapport de puissance entre le signal original et la marque (DWR) suffisamment grand. Par exemple, dans le cas de l'image "Lena", nous avons utilisé un $DWR = 34.46dB$. Lors de notre étude, les images sont transformées selon la procédure indiquée précédemment et nous choisissons pour l'espace de projection $l = 4$ composantes indépendantes, ce qui correspond à une matrice de séparation A de taille 4×4.

Pour notre étude, nous avons évalué les performances de notre système TCQ-ICA et ainsi que celles d'autres systèmes référents comme éléments de comparaison à savoir : le SCS (Schéma de Costa exploitant une quantification scalaire), la TCQ (Schéma de costa exploitant une quantification codée treillis), le ST-SCS (le SCS utilisant une transformée par étalement), le ST-TCQ (la TCQ exploitant les propriétés d'une transformée par étalement), le SCS-ICA (le SCS dans un domaine indépendant).

4.1 Robustesse

Dans un premier temps nous avons évalué la robustesse de notre schéma face à une attaque par ajout de bruit AWGN. Aussi en faisant varier la puissance du bruit par rapport à celle de la marque, nous avons tracé la courbe du BER (pour un $DWR \approx 35dB$). Nous présentons les résultats obtenus sur les courbes de la figure 5.

Nous constatons que le schéma TCQ-ICA offre un niveau de robustesse meilleur que celui de la TCQ et du SCS. En fait, par rapport à ces 2 schémas de référence, pour un $BER = 10^{-2}$, la quantité de bruit tolérée peut-être augmentée de $5dB$ pour une même puissance d'insertion de la marque. Pour des taux d'erreur plus petits, cette augmentation du niveau de bruit toléré s'accroît dans le cas du SCS. mais reste quasiment la même dans le cas de la TCQ.

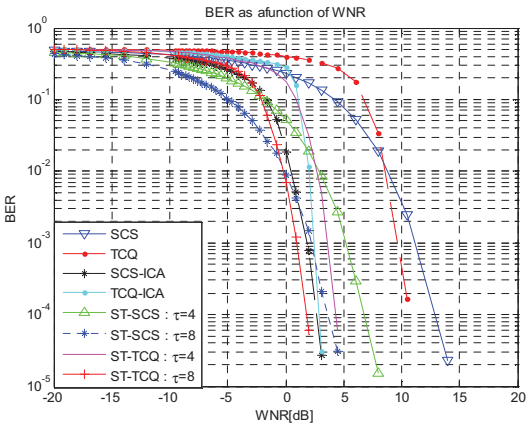


Figure 5 – *Etude comparative de la robustesse du schéma TCQ-ICA avec d'autres schémas de référence*

En comparaison avec les schémas mettant en oeuvre une transformée par étalement (ST-SCS et ST-TCQ), lorsque le facteur d'étalement $\tau = 4$, le taux d'erreur lors du décodage du message est un peu moins bon dans le cas du schéma proposé pour les $WNR < 2dB$. Par contre, au delà de cette valeur du WNR notre schéma tatouage basé sur la TCQ dans un domaine indépendant offre des taux d'erreur binaires nettement meilleurs. C'est seulement pour un facteur $\tau \geq 8$ que le schéma proposé présente une robustesse moins bonne pour cette gamme de WNR.

Comparé au schéma SCS-ICA, les performances obtenues pour notre schéma sont un peu moins bonnes mais restent tout de même tout à fait correctes pour une utilisation en tatouage robuste. Ces bons résultats s'expliquent par le fait que dans le domaine indépendant si on considère la matrice de mélange A composée des éléments $a_{i',j'}$, le WNR dans le domaine indépendant ($WNR_{(ICA)}$) peut s'écrire en fonction du WNR dans le domaine initial ($WNR_{(Init)}$) par la relation suivante :

$$WNR_{(ICA)_{j'}} = 10 \log_{10} \left(\sum_{i'=1}^{k^2} a_{i',j'}^2 . 10^{\frac{WNR_{(Init)_{j'}}}{10}} \right)$$

avec $i' = \{1, ..., k^2\}$ et $j' = \{1, ..., k^2\}$. Comme les éléments $a_{i',j'} > 1$, calculer un taux d'erreur pour un WNR donné revient à se placer dans le domaine indépendant à un $WNR_{(ICA)}$ beaucoup plus élevé, donc dans lequel le bruit est moins influent.

4.2 Capacité

Pour notre schéma nous nous sommes intéressé à la mesure de la capacité, c'est à dire de la quantité d'information que l'on peut décoder sans erreur pour un niveau de bruit donné. Les résultats obtenus sont présentés sur les courbes de la figure 6.

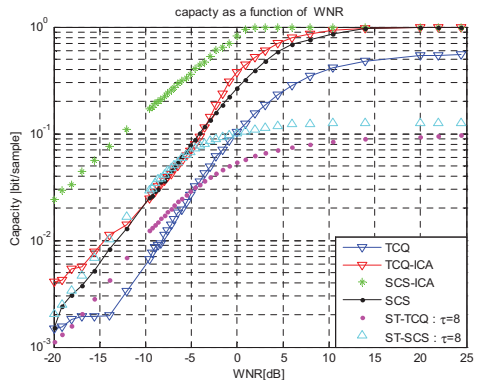


Figure 6 – *Etude comparative de la capacité du schéma TCQ-ICA avec d'autres schémas de référence*

Comparés aux schémas pour lesquels les performances de robustesse étaient proches, tels que ceux s'appuyant sur la transformée par étalement (ST) tels que le ST-SCS et le ST-TCQ, nous montrons sur cette figure que la capacité est semblable lorsque le WNR est faible ($WNR < -5dB$) mais elle est bien supérieure en dehors de cette gamme de valeur ($WNR > -5dB$). En effet il est bien connu que ces schémas basés sur le ST ont de bonnes performances de robustesse lorsque le facteur τ augmente mais une capacité qui diminue d'autant. Ce qui n'est pas le cas pour le schéma basé sur l'ICA. Par conséquent, le compromis de performances robustesse-capacité est meilleur pour notre schéma que pour ceux basés sur le ST.
En fait, par rapport aux autres schémas de référence choisis, la capacité que nous offre notre schéma TCQ-ICA est bien meilleure que celle de la TCQ et devient du coup semblable à celle du SCS. En revanche, elle reste moins bonne que celle du SCS-ICA. Comme pour la robustesse, ces bons résultats peuvent s'expliquer par la relation sur le $WNR_{(ICA)}$ donnée précédemment due au fait que la marque dans le domaine ICA a une puissance plus élevée que dans le domaine initial. En effet on peut écrire que :

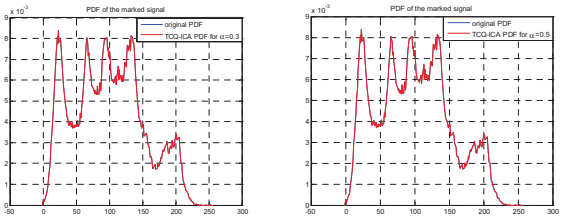
$$\sigma^2_{w'_{j'}} = \sum_{i'=1}^{k^2} a^2_{j'i'} \cdot \sigma^2_{w_{i'}}$$

avec $\sigma^2_{w'_{j'}}$ et $\sigma^2_{w_{i'}}$ respectivement les composantes des variances de la marque dans le domaine ICA $\sigma^2_{w'}$ et dans le domaine initial σ^2_w. Cette amplification de la puissance de la marque permet en effet de mieux décoder l'information sans erreur pour un niveau de bruit donné qui dans le domaine initial est plus faible.

4.3 Invisibilité

Pour ce critère de performance, nous avons cherché à évaluer aussi bien l'invisibilité aussi bien d'un point de vue statistique que perceptuel.

Indéctectabilité statistique. D'un point de vue indétectabilité statistique, nous cherchons en fait à évaluer la sécurité telle que définie par Cachin [9] en s'appuyant sur les fonctions de densité de probabilité (PDF) des signaux originaux et des signaux marqués. Nous présentons sur la figure 7 les PDF obtenues pour plusieurs valeurs de α dans le cas d'une image réelle.



(a) (b)

Figure 7 – *Comparaison entre la fonction de densité de probabilité avant et après marquage avec le schéma TCQ-ICA pour : (a)* $\alpha = 0.3$*, (b)* $\alpha = 0.5$

En observant ces courbes nous constatons que les distorsions dues au marquage sont faibles et peu visibles sur les statistiques des signaux. Aussi, nous avons effectué une étude plus approfondie de ces signaux en utilisant la Distance de Kullbach-Leibler (KLD) entre les PDF des signaux marqués et non marqués pour le schéma proposé. Nous comparons les résultats avec ceux obtenus pour les autres schémas de référence choisis (figure 8).

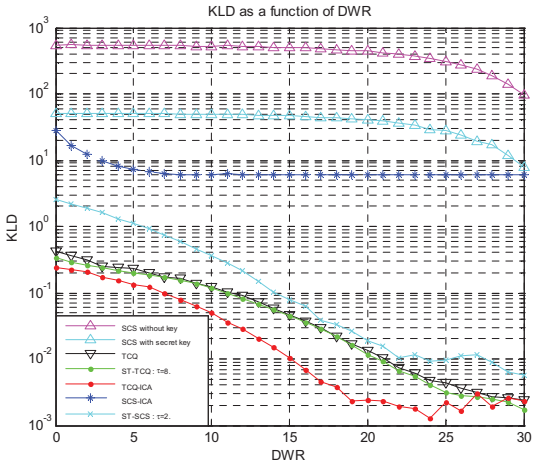


Figure 8 – *Etude comparative de l'indétectabilité du schéma TCQ-ICA avec d'autres schémas de référence*

Pour le schéma que nous proposons, nous obtenons une KLD faible qui diminue en fonction du rapport de puissance du signal avec la puissance de la marque (DWR). Cette réponse reste proche voire légèrement meilleure que celles obtenues pour la TCQ mais aussi pour le ST-SCS et ST-TCQ connus pour leurs bonnes performances en termes d'invisibilité statistiques [14], [10].
Sachant que le point de fonctionnement pour les images est

en général pour des valeurs du $DWR > 25dB$ (Imperceptibilité), de tels résultats signifient qu'avec notre schéma il sera impossible de distinguer statistiquement l'image marquée de l'image originale, ce qui peut être intéressant dans des approches de tatouage robuste et indécelable.

Imperceptibilité. En termes de perceptibilité, nous avons cherché à comparer les images originales et images marquées entre elles pour un même message inséré. Nous présentons sur la figure 9 un exemple de ce que nous avons obtenus avec l'image de Lena dans le cas des 2 schémas jusqu'ici jugés les plus performants : SCS-ICA, TCQ-ICA.

(a) (b) (c)

Figure 9 – *Exemple d'application aux images : (a) Originale (b) Marquée avec le SCS-ICA (c) Marquée avec le TCQ-ICA pour un* $DWR = 34,46dB$

Dans les 2 cas, le DWR lors se l'insertion a été fixé à $34,46dB$. Pour les 2 schémas il n'y a pas de différences visibles. La mesure du PSNR (Peak Signal to Noise Ratio), qui nous donne des valeurs d'environ $40dB$ dans les 2 cas, ne fait pas apparaître de différences perceptuelles significatives pour les 2 méthodes considérées. Aucun de ces 2 schémas n'introduit de dégradations perceptuelles significatives sur les images marquées.

5 Conclusion

Nous présentons ici une méthode de tatouage d'images basé sur la quantification par Treillis (TCQ) dans un domaine indépendant (ICA). Il s'agit d'une méthode de tatouage avec des propriétés de robustesse comparables à celles obtenus pour des schémas référents exploitant la transformée par étalement ST. Cependant, contrairement au ST, il y a pour cette méthode une très bonne capacité. De plus, cette méthode offre un niveau de sécurité intéressant en terme d'indétectabilité statistique.
Compte tenu des bonnes performances obtenues pour ce schéma de tatouage robuste et indécelable, nous envisageons d'étendre l'analyse de robustesse à d'autres attaques spécifiques aux images du types de celles décrites dans Stirmark (compression, mise à l'échelle, ...). Quant à l'indétectabilité, nous souhaitons aussi évaluer les performances en tenant comptes des statistiques à l'ordre 2.

6 Remerciements

Les auteurs remercient les gouvernements Français et Algériens pour leurs aides financières (ANR Projet MEDIEVALS et Programme PROFAS BAF).

Références

[1] M. H. M. Costa. Writing on dirty paper. *IEEE Trans. on Information Theory*, 29 :439–441, 1983.

[2] J. J. Eggers, R. Bauml, R. Tzchoppe, et B. Girod. Scalar costa scheme for information embedding. *IEEE Trans. on Signal Processing*, 51 :1003–1019, 2003.

[3] D. Yu, F. Sattar, et K.K. Ma. Watermark detection and extraction using independent component analysis method. *EURASIP Journal on Applied Signal Processing*, 1 :92–104, 2002.

[4] A. Hyvarinen. Survey on independent component analysis. *Neural Computing Surveys*, 2 :94–128, April 1999.

[5] F. J. González-Serrano, H. Y. Molina-Bulla, et J. J. Murillo-Fuentes. Independent component analysis applied to digital image watermarking. Dans *ICASSP 01*, pages 1997–2000, May 2001.

[6] J.J. Murillo-Fuentes. Independent component analysis in the blind watermarking of digital images. *Neurocomputing*, 70 :2881–2890, 2007.

[7] S. Bounkong, B. Toch, D. Saad, et D. Lowe. Ica for watermarking digital images. *Journal of Machine Learning Research*, 4(7-8) :1471–1498, 2003.

[8] I. Benkara Mostefa, C. Delpha, S. Braci, R. Boyer, et M. Khamadja. Improved performances of scalar costa scheme for images watermarking in an independent domain. Dans *6th Int'l Symposium on Image and Signal Processing and Analysis*, pages 477–482, Salzburg, Austria, Sept. 2009. IEEE.

[9] C. Cachin. An information-theoretic model for steganography. Dans *Lecture Notes in Computer Science*, volume 1525, pages 306–318. Springer, Jan. 1998.

[10] S. Braci, C. Delpha, R. Boyer, et G. Le Guelvouit. Informed stego-systems in active warden context : Statistical undetectability and capacity. Dans *IEEE Int. Workshop on MultiMedia Signal Processing*, Cairns, Australia, Oct. 2008.

[11] M.W. Marcellin et T.R. Fischer. Trellis-coded quantization of memoryless and gauss-markov sources. *IEEE Trans. on Com.*, 38 :83–93, Jan. 1990.

[12] G. Ungerboek. Trellis coded modulation with redundant signal sets. *IEEE Communications Magazine*, 25(2) :5–25, Feb. 1987.

[13] A.J. Viterbi. Error bounds for convolutional codes and asymptotically optimum decoding algorithm. *IEEE Trans. on Information Theory*, 13(2) :260–269, April 1967.

[14] G. Le Guelvouit, A. Ould Bouya, J. Bourgeois, C. Delpha, et R. Boyer. Analyse stéganographique du schéma scalaire de costa. Dans *GRETSI*, 2007.

[15] E. Esen et A.A. Alatan. Data hiding using trellis coded quantization. Dans *International Conference on Image Processing*, Singapore, Oct. 2004. IEEE.

Optimal fusion scheme selection framework based on genetic algorithms, for multimodal face recognition

W. Ben Soltana[1] M. Ardabilian[1] L. Chen[1] C. Ben Amar[2]

[1] LIRIS (Laboratoire d'InfoRmatique en Image et Systèmes d'information)

Université de Lyon, CNRS
Ecole Centrale de Lyon, LIRIS, UMR5205, F-69134, France

{wael.ben-soltana, Mohsen.Ardabilian, Liming.Chen}@ ec-lyon.fr

[2] REGIM (REsearch Group on Intelligent Machines)

University of Sfax
National Engineering School of Sfax, 3038, Sfax, Tunisia

chokri.benamar@ieee.org

Abstract

In this paper, we consider the problem of feature selection and classifier fusion and discuss how they should be reflected in the fusion system architecture. We employed the genetic algorithm with a novel coding to search the worst performing fusion strategy. The proposed algorithm tunes itself between feature and matching score levels, and improves the final performance over the original on two levels, and as a fusion method, it not only contains fusion strategy to combine the most relevant features so as to achieve adequate and optimized results, but also has the extensive ability to select the most discriminative features and their appropriate classifiers. Sparse Representation Classifier (SRC) and Nearest Neighbor classifier with euclidean distance, mahalanobis distance, cosine distance and correlation distance are exploited to calculate all the similarity measures. Experiments are provided on the FRGC database and show that the proposed method produces significantly better results than the baseline fusion methods.

Keywords

Genetic al gorithm, fusi on st rategy, feat ure l evel, score matching level, classifier selection, clas sifier fusion, Sparse Representation Classifier.

1 Introduction

In recent years, there are many research works and studies of m ultiple classifier system s. It h as been frequently demonstrated that combing classifiers can offer significant classification performance i mprovement for a n umber of non-trivial pattern recognition problems [1].

Fusion strategies can be roughly classified into three m ain categories: fusi on at a n early st age, fusion at a later stage and hy brid fusion. M any sy stems t hat i ntegrate information at an early stage are believe d to be m ore effective th an tho se t hat p erform in tegration at a later stage. T herefore, while it is relatively more d ifficult to achieve in practice [2], fu sion at early stage has drawn more atten tion in recen t year s. Th ere ex ist two typ es of early fu sion: fu sion at d ata lev el (fo r ex ample 3 D im age [3] or 3D/2D image [4]), and fusion at feature level [2]. In fact, at the fea ture le vel the concatenated feature vectors may cont ain noisy or redundant dat a, t hus l eading t o decreased pe rformances of t he cl assifier [5]. In this ca se, feature sel ection p rocedure i s an important step. It is essentially an o ptimization pr oblem that i nvolves searching within the space of possible feat ure s ubsets to find one s ubset that is opt imal (or near-optimal) with respect to a c ertain criterion. Se veral sea rch strate gies have bee n put forward a nd can be classifi ed int o three categories: optimal, heuristic, and randomized. Exhaustive search is the most straight forward a pproach t o o ptimal feature sel ection a nd i t i s guara nteed t o fi nd t he optimal subset. H owever, si nce t he num ber of possible s ubsets grows ex ponentially, exha ustive searc h becomes not feasible a nd im practical even for m oderate feature numbers. T he onl y opt imal feat ure sel ection m ethod, which avoids the exhaustive search, is based on the branch and bound al gorithm [5, 6]. B est i ndividual features, sequential forward selection (SFS) an d sequential backward sel ection (SBS) [5] are three well-known heuristic su boptimal feat ure sel ection sc hemes. Combining SFS and SBS gives birth to plus l-take away r

feature selection. A generalization of the plus l-take away r method is twofold: Sequential forward floating search [5, 7] and sequential backward floating search [5, 8], where l and r are determ ined automatically and update d dynamically. Evolutionary algorithms [1] are ra ndom search algorithms. Among them, genetic algorithms (GAs) include a s ubset of ev olutionary al gorithms foc using on the app lication o f selection , mutation, and recombination to a pop ulation of c ompeting problem sol utions. Obviously, GAs a re prime can didates f or random probabilistic search al gorithms with in the co ntext of feature selection.

In fusion at later stage, t here are three fusi on sub-le vels: score m atch l evel [9], ra nk l evel [10] a nd decision l evel [11]. Kittler and al. [12] presented and d eveloped a common theoretical fram ework fo r th ese co mbining classifiers. At the first level, similarity scores generated by classifiers are com bined by various t echniques [13], for example, Sum Rule, Product Rule, etc. Gabrys and al. [14] developed a weighted s oft com biners based on ge netic algorithms. At th e seco nd lev el, sorted lists co mputed b y classifiers are merged based on different approaches such as Borda Count and Logistic Regression [15]. At the third level, all the candidates of the classifiers are fuse d by adopting sev eral methods [16], i.e., Majo rity Vo te. Th e last category contains intermed iate fusion s chemes, such as serial fu sion an d m ultilevel fu sion. Th e m ain motivation of the serial architecture [16] is to filter out the most s imilar K classes usi ng a sim ple cla ssifier and the n to feed these K classes into a more complex and powerful second classifi er. On the ot her side, the re are few works that d escribe multilevel f usion. In [17], fusion is introduced in both feature level and confidence level.

In t his work, we de velop a c ommon fram ework for selecting features, classifiers and combining the latest; we confirm that many existing schemes can be considered as special cases of our ge neric fusion scheme. We s how that our f usion m ethod i s able t o obtain a global sub-optimal solution while lessen ing th e co mplexity o f calcu lation. Other cont ributions o f t his paper are: t he use o f genet ic algorithm wi th a novel c oding st rategy and se quential backward floating search for effective feature selection; at the sam e time an op timal fu sion strateg y sch eme is generated. F urthermore, a n appropriate classifier for each feature type was determined automatically.

The remainder of t his paper i s or ganized as f ollows: a framework for feature and score matching levels fusion is introduced i n section 2, and section 3 presents experimental results. Section 4 concludes the paper.

2 Optimal fusion sc heme selection framework

The proposed framework is shown in Fig. 1. It is based on genetic al gorithm, usi ng a novel co ding technique, t o search the optimal fusion scheme.

2.1 Framework Overview

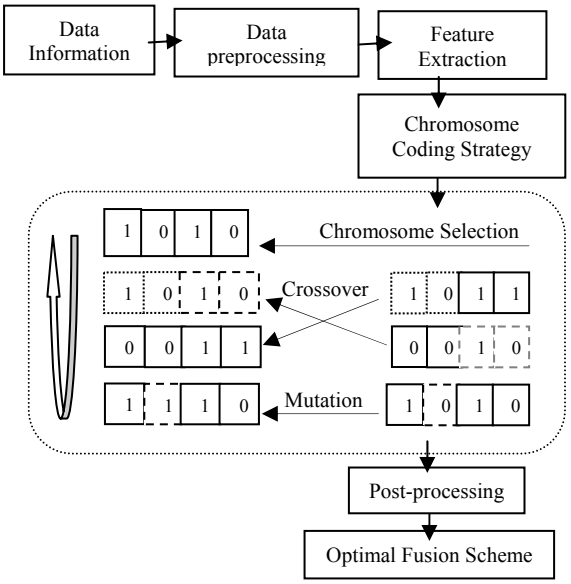


Figure 1 - *Algorithm Overview*

The proposed framework consists of four steps. In the first step, Data Prep rocessing, the d ata validity and in tegrity are chec ked a nd noisy data is rectified. T he sec ond ste p consists to features extraction. For m easurement cost and classification accuracy, Linear Disc riminant Analysis (LDA) is u sed to reduce t he d imensionality fo r each feature. The t hird st ep l ead t o: 1) fi nding one su bset of features th at is op timal with respect to t he corresponding fusion scheme and 2) det ermining, aut omatically, an appropriate classifier for each feature type. So, all features are coded to form individual “chromosomes” according to the model described in the section 2.3. Furthermore, these chromosomes are used by a ge netic algorith m [18] to encode the trial so lution for the current problem. Iterative selection, cr ossover, a nd m utation we re used t o m ake evolve a ne w population. At each ne w generation, a ne w set of chromosomes is produced, using the fittest g enes of the previous generation, for a b etter so lution. Assessment of t he sat isfactory de gree of t his sol ution, enc oded as individuals, is reflected in the fitn ess. In fact, fitn ess corresponds to perform ance ra te of eac h f usion st rategy represented by in dividual ch romosome. This fitn ess i s calculated according to eq . (2) taki ng into account different classifiers. Figure.2 illustrates this process. Also, the individuals with higher fitness have a high probability of being sel ected and producing offspring. The crossover operator produces better o ffspring by e xchanging t he characteristics o f th e parents. Th is en ables th e m ost efficient ch aracteristics to be con centrated in t he sam e individual. T he m utation o perator ra ndomly cha nges t he genetic representation of an individual and tends to inhibit the po ssibility o f conv erging to a lo cal optimum, rath er

than the global optimum. The evolution is carried out until a desired solution is arrived, or a pre-specified number of iterations are completed. The final solution with higher fitness represents the final fusion strategy.

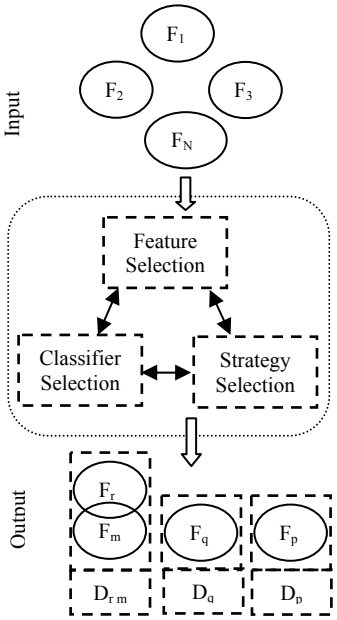


Figure 2 - *Genetic algorithm optimization for feature, classifier and strategy selection*

The last step can generate an optimal final fusion strategy. In fact, SBFS [5] is used to select the best features in new concatenated features (See Section 2.3). Furthermore, if the final strategy represents a matching score level strategy, we assign new different weights following eq.(3) to ameliorate the performance rate.

2.2 Performance Rate

The performance rate is calculated for each chromosome. So, all features output different scores. Min-Max Normalization [2] is used to map the matching scores to the range of [0, 1]. At the strategy selection stage, for each g in the gallery set, we compute a similarity score $S^{g,f}$ by classifier using each feature f with the probe. All these similarities $S^{g,f}$ are then sorted in a descending order. We assign to each score $S^{g,f}$ a weight $w^{g,f}$ which is a function of its ordered position $p^{g,f}$. Specifically, the weight $w^{g,f}$ is defined as:

$$w^{g,f} = f(p) \propto \ln (N_g / p^{g,f}). \quad (1)$$

where N_g is the number of the subjects in the gallery.
The matching score, in the strategy selection stage, between the g in the gallery and the probe is:

$$S(g) = \sum_{f \in features} w^{g,f} . S^{g,f} . \quad (2)$$

This weighting strategy gives more importance to the scores ranked at the first positions and aims to discard wrong matching of each feature in test by assigning a lower weight to its corresponding similarity with a gallery sample.
At the post-preprocessing stage, we use another Genetic Algorithm to assign a weight $P^{g,f}$ to scores of a particular feature. The final matching score between g in the gallery and the probe is:

$$S(g) = \sum_{f \in features} P^{g,f} . w^{g,f} . S^{g,f} . \quad (3)$$

The probe face is recognized as the one in the gallery which obtained the highest final score according to (3).

2.3 Feature subset and Strategy Selection

We propose a novel coding strategy to select simultaneously the efficient feature, the best classifier and the optimal fusion scheme. This coding strategy consists to divide the chromosome into two parts: Part A and Part B (See Figure.3). Given N features, Part A has N gene positions that correspond to each feature, and represented with integer values: 1 implies that the feature is active and used in feature level fusion, 0 implies that the feature is active and used in score level fusion, and -1 implies that the feature is inactive. Part B codes the fusion model that depends on the number N_F of active features at feature level fusion. In this model, we generate all possible combinations. However, we can't create a strategy that contains a single feature and we consider that combinations obtained by permutation are equivalent. Part B is also composed of two parts P1and P2: P1 refers to the model M and P2 associates the features in this model.

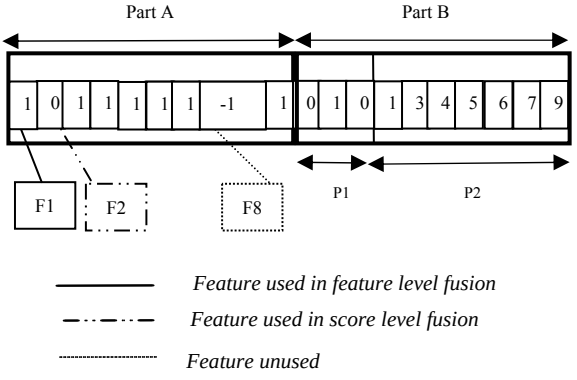


Figure 3 - *Example of chromosome coding strategy*

An example of this representation is illustrated in Figure 3. With a Part A as 1011111-11, we can generate 4 models M_i, with i in {1..4}: M_1=(7, 0), M_2= (2, 5), M_3= (3, 4), M_4=(2, 2, 3). The number of the selected model is represented in the chromosome by its binary code: the model M_2 is selected and represented by (010) and two vectors V1, V2 are created by concatenation, V1= [F1,

F3] and V2= [F4, F 5, F6, F7, F9]. The fusion strategy corresponds to a sco re matching lev el with V1 , V2 , an d F2. Th e fitn ess o f th is strateg y is calcu lated b ased o n performance rate described in previous section. Stochastic universal sa mpling [19] i s used t o sel ect best chromosomes "strategies". Uniform crossover is used only on Part A a nd ran dom mutation may occur on Part A or Part B o f c hromosome. Stop ping c riteria ch osen f or problem so lving is selected fro m th ese co nditions: 1) either the maximum number of iterations over the terminal number m ax of generations, 2) t he best fi tness val ue beyond the value of fitness limits.

3 Experimental Results

The proposed alg orithm is tested in a face recogn ition application, where th e objective is to find an op timal subset of features and their adequate fusion strategy.

3.1 Database, Experiment Settings and Feature Extracted

The FRGC [20] database was chosen for our experiments. Each fa ce dat a consists of one 3D face model and its registered 2D color im age. The 3D face models a re fi rst cropped with a sphe re of radius of 80 *mm* centered at the nose tip and preprocessed with techniques in [21]. In order to av oid t he i mpact of re gistration errors i n our a nalysis, we use d a manual re gistration m ethod, n amely R egion based Ite rative Closet Point (R-ICP) [22] for 3D face model. . As 2D c olor t exture i nformation i s densel y registered t o i ts co rresponding 3 D face dat a, t he previously cropped and registered 3D face model has also its 2D tex ture counterpart. The positions of both two eye inner corners are further use d fo r rot ation no rmalization. Finally, all th e 2D co lor faces are conve rted to the gray-level, and resized to 80×92 pixels.

FRGC v1.0 dataset is used for estimating LDA parameters while FRGC v2.0 is u tilized fo r train ing an d test; 1 16 subjects having eac h 4 face m odels were selected from FRGC v1.0 to train subs pace base d approaches s uch as estimating LDA parameters. One face scan (3D+2D) with a ne utral e xpression wa s sel ected from ea ch s ubject to make a gallery of 410 subjects (gallery database) and 3541 face sca ns we re treated as probes a nd were sepa rated to build training database (2332 face scan) and test database (1209 face sc an). The test database is di vided into two subsets acc ording t o t heir e xpression l abels. The fi rst subset contains face s cans with the ne utral expre ssion (713 probes), and the other one with face sc ans producing non-neutral expressions (496 pr obes). In the second step, we ha ve use d 2D an d 3 D features. Ge ometric feat ures include no rmal (Nor Vec), b inormal (BiN Vec), tang ent vector (Tang Vec) [23] and curvature. Four categories of curvature-based features are extracted. The first two types rely on m ain directions corresponding to maximum (Max Curv) and minimum (Min Curv) curvatures [24]. The last two are t heir derivatives, i.e ., the mean (Mean Curv) and Gaussian (Gauss C urv) curvat ures. We furt her investigated a nother t ype o f 3 D feature based o n t he anthropometric (An thr Mes) approac h which advocates extracting a sig nature from some an thropometric p oints considered th e most relev ant. Three different features are extracted from the 2D text ure images. The first one is the simple pi xel-based method that encodes grayscale intensity (In tensity) v alues in to a vector. The second one is a non-pa rametric feature na mely Lo cal Bin ary Patterns (LBP) [25]. The most important properties of LBP are the tolerance against the m onotonic illumination changes a nd its computational simplicity. The third feature is ex tracted by Gabor filters (Gabor) [26] which are spatially localized and selectiv e to sp atial orien tations and scales. Five different frequencies and eight equally spaced orientations are utilized to generate Gabor kernels.

For evaluating t he proposed app roach, experiments were designed in i dentification t ask with trai ning an d test stages. The t raining st age outputs t he o ptimal fusi on strategy. Th e sa me g allery d atabase is u sed in all stag es. In training stage, we achieve one experiment using gallery database an d training database. T he t raining database contains neutral and non neutral expressions. In test stage, three e xperiments we re car ried o ut: Ne utral vs. Neutral, Neutral vs. N on-Neutral, a nd Ne utral vs. All. In Ne utral vs. Neutral and Neutral vs. Non-Neutral, only the neutral and non-neutral probe subsets were used.

3.2 Classifiers

Two classifie rs, Sparse R epresentation Classifier and Nearest Neighbor Classifier were used in our experiments.

Sparse Repre sentation Classifier: Sp arse rep resentation for signal classification (SRSC) is proposed in [27]. SRSC incorporates reconstruction pr operties, di scriminative power an d sp arsity for robust classification . I n [28], a general cl assification ap proach f or (image-based) o bject recognition is pr oposed base d on a sparse rep resentation computed by L1-m inimization. T he m ethod base d o n sparse representation can oft en ac hieve high performance based on a data dictionary [29].

Nearest Neig hbor Classifier: Euclidean distance (4), Mahalanobis distance (5), cosi ne distance (6) a nd correlation distance (7) a re i ntroduced, an d t heir performances are als o c ompared in our e xperiments. A description of each of these metrics can be found below:

$$d(x,y)^2 = (x-y)(x-y)', \quad (4)$$

$$d(x,y)^2 = (x-y)C^{-1}(x-y)', \quad (5)$$

$$d(x,y) = 1 - \frac{(xy')}{(x'x)^{1/2}(y'y)^{1/2}}, \quad (6)$$

$$d(x,y) = 1 - \frac{(x-x_M)(y-y_M)'}{((x-x_M)(x-x_M)')^{1/2}((y-y_M)(y-y_M)')^{1/2}}. \quad (7)$$

where x and y are two rows vectors to compare, C is the covariance matrix, x_M is the mean value of x, and y_M is the mean value of y.

3.3 Results and Analysis

First, LDA is applied to reduce dimensionality of all features in FRGC v2 database. In order to use the genetic algorithm in training stage, we define some parameters. Part A of chromosome is organized as follows: {Tang Vec, BiN Vec, Nor Vec, Gauss Curv, Max Curv, Mean Curv, Min Curv, LBP, Gabor, Anthr Mes, Intensity}. Five similarity measure of each feature was computed with SRC classifier (SRC), 1-Nearest Neighbor with euclidean distance measure (1-NN-ED), 1-Nearest Neighbor with mahalanobis distance measure (1-NN-MD), 1-Nearest Neighbor with cosine distance measure (1-NN-CosD), 1-Nearest Neighbor with correlation distance measure (1-NN-CorrD). The selection algorithm used a population of 50 chromosomes. The mutation rate was set to 0.1 and the GA was stopped after 100 generations for experiment.

Table 1. Rank-one recognition rate of individual type of feature and classifiers selected by the GA for the best fusion strategy

Classifier	Features			
SRC	Gabor	81.97	Bin Vec	73.12
	LBP	77.01	Anthr Mes	58.97
	Intensity	53.35		
1-NN-CorrD	Tang Vec	82.22		
	Mean-Gauss	74.16		

The final fusion strategy generated by training stage is coded as follows. Part A: 0,0,-1,1,-1,1,-1,0,0,0,0, Part B: [P1: 001, P2: 4, 6]. It consists firstly to concatenate {Mean Curv, Gauss Curv} in vector V_1. Secondly, we use this optimal subset {V_1, Tang Vec, BiN Vec, LBP, Gabor, Anthr Mes, Intensity} in score level fusion. The post processing step is used to optimize the final fusion strategy. Firstly, SBFS is used to select the best feature in the new concatenated vector V1. Secondly, new weight processing eq. (3) is used. The final recognition rate is 95.67% in training stage. In test stage, we apply the final fusion strategy. In this case, the final recognition rate is 97.27% using Neutral vs All experiment. Table 2 compares the proposed fusion strategy with other fusion approaches (simple sum rule (baseline method), Gökberk and al. [7], Mian and al.[2]). The performance of each feature and classifiers selected by the GA is displayed in Table1. Others experiments (Neutral vs Neutral and Neutral vs Non Neutral) are presented in Table 3. In all experiments, our method improves rank-one recognition accuracy as compared with other methods in three aspects: selecting the most discriminative features, selecting appropriate classifier and proposing an optimized fusion strategy.

Table 2. Identification Results Using Neutral vs All

Method	Training Phase	Test Phase
Simple Sum Rule	_	94.13%
Gökberk and al. [7]	_	95.28%
Mian and al.[2]	_	94.71%
Our approach	95.67%	97.27%

Table 3. Identification Results (Rank-one) Using Neutral vs Neutral (N-N), Neutral vs NonNeutral (N-nonN)

Method	N-nonN	N- N
Simple Sum Rule	91.33%	97.76%
Gökberk and al. [7]	92.75%	97.88%
Mian and al.[2]	90.73%	97.76%
Our approach	95.16%	99.16%

4 Conclusions and Future Works

In this paper, we developed a common framework for optimal fusion strategy selection. The proposed framework, based on a genetic algorithm and a novel associated coding strategy, generates automatically a subset of best features, an appropriate classifier for each feature, and an optimal fusion strategy scheme. Experiments are provided on the FRGC database and show that the proposed method produces significantly better results.

In future works, we can integrate other features and classifiers to improve the potential of our method. We intend also to extend this fusion scheme in order to generate the best model for each application. We plan as well to analyze the impact of the quality of information on fusion strategy.

Acknowledgement

This work was partially carried out within the French FAR3D project supported by ANR under the grant ANR-07-SESU-004 FAR3D.

References

[1] Dymitr Ruta, Bogdan Gabrys. Classifier selection for majority voting. *J. Inf. Fusion*, 6 (1), 63–81, (2005).

[2] Ajmal S. Mian, Mohammed Bennamoun, Robyn Owens. Keypoint detection and local feature matching for texture d 3D face Recognition. International Journal of Computer Vision, 1-12, 2008.

[3] Theodoros Papatheodorou, Daniel Rueckert. Evaluation of Automatic 4D Face Recognition Using

Surface and Texture Registration. *Automatic Face and Gesture Recognition*, 321 – 326, May 2004.

[4] Gede Putra Kusuma and Chin-Seng Chua. Image Level Fusion Method for Multimodal 2D + 3D Face Recognition. *ICIAR* 2008, 984-992.

[5] Anil K. Jain, Robert P. W. Duin, Jian chang Mao, Statistical Pattern Recognition: A Review. IEEE Trans. Pattern Anal. Mach. Intell, 22(1), 4-37, 2000.

[6] Patrenahalli M. Narendra and Keinosuke Fukunaga. A Branch and Bound Algorithm for Feature Subset Selection. *IEEE Transactions on Computers*, vol. C-26, issue 9, 917-922, Sept 1977.

[7] Berk Gökberk, Helin Dutagacı, Lale Akarun, Bülent Sankur. Representation plurality and fusion for 3D face recognition. *IEEE Trans. on Systems Man and Cybernetics-Part B*, 38(1), 155–173, 2008.

[8] Ververidis, D., Kotropoulos, C.: Fast and accurate sequential floating forward feature selection with the Bayes classifier applied to speech emotion recognition. *Signal Processing* 88 (12), 29 56-2970, 2008.

[9] Jamie Cook, Mark Cox, Vinod Chandran, Sridha Sridharan. Robust 3D face recognition from expression categorization. *International Conference on Biometrics,* 271-280, 2007.

[10] Berk Gökberk, Albert Ali Salah and Lale Akarun. Rank-based decision fusion for 3D shape-based face recognition. *International Conference on Audio- and Video-Based Biometric Person Authentication*, 1019-1028, 2005.

[11] Timothy Faltemier, Kevin Bowyer, Patrick Flynn**.** 3D face recognition with region committee voting. *International Symposium on 3D Data Processing, Visualization, and Transmission*, 2006.

[12] Josef Kittler, Mohamad Hatef, Robert P.W. Duin, Jiri Matas. On Combining Classifiers. *IEEE Trans. Pattern Anal. Mach. Intell*, 226-239, 20(3) 1998.

[13] Afzal Godil, Sandy Ressler and Patrick Grother. Face recognition using 3D facial shape and color map information: comparison and combination. *Biometric Technology for Human Identification, SPIE*, 5 404, 351–361, 2005.

[14] Bogdan Gabrys Dymitr Ruta. Genetic algorithms in classifier fusion. *Appl. Soft Comput.* 6 (4), 337-347 (2006).

[15] Md. Maruf Monwar, Marina Gavrilova. FES: a system for combining face, ear, and signature biometrics using rank level fusion. *International Conference on Information Technology: New Generations*, 922-927, 2008.

[16] Berk Gökberk and Lale Akarun. Comparative analysis of decision level fusion algorithms for 3D face recognition. *ICPR,* 2006.

[17] Congcong Li, Guangda Su, Yan Shang, Yingchun Li, and Yan Xiang. Face Recognition Based on Pose-Variant Image Synthesis and Multi-level Multi-feature Fusion. *AMFG 2007*, 261-275, 2007.

[18] Mohammad Sedaaghi, Constantine Kotropoulos, Dimitrios Ververidis. Improving speech emotion recognition using adaptive genetic algorithms. *Proc. EUSIPCO,* Polland, 2007.

[19] James E. Baker. Reducing Bias and Inefficiency in the Selection Algorithm. *Proceedings of the Second International Conference on Genetic Algorithms and their Application,* 14-21, 1987.

[20] P. Jonathon Phillips, Patrick J. Flynn, W. Todd Scruggs, Kevin W. Bowyer, Jin Chang, Kevin J. Hoffman, Joe Marques, Jaesik Min, William J. Worek. Overview of the Face Recogniti on Grand Challenge, *CVPR*, 947-954, 2005.

[21] Przemyslaw Szeptycki, Mohsen Ardabilian, and Liming Chen. A coarse-to-fine curvature analysis-based rotation invariant 3D face landmarking. *International Conference on Biometrics: Theory, Applications and Systems,* 2009.

[22] Boulbaba Ben Amor, Mohsen Ardabilian, Liming Chen. New experiments on ICP-based 3D face recognition and authentication. *ICPR,* 2006.

[23] Steven W. Zucker. Differential geometry from the frenet point of view: boundary detection, stereo, texture and color. *Handbook of Mathematical Models in Computer Vision,* N. Paragios, Y. Chen, and O. Faugeras, eds., Springer, 2005.

[24] Hiromi T. Tanaka, Masaki Ikeda and Hisako Chiaki. Curvature-based face surface recognition using spherical correlation. *IEEE International Conference on Automatic Face and Gesture Recognition,* 372–377, 1998.

[25] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face Recognition with Local Binary Patterns. *ECCV,* 2004.

[26] Laurenz Wiskott, Jean-Marc Fellous, Norbert Krüger, Christopher von derMalsburg**.** Face recognition by elastic bunch graph matching. *IEEE Trans. on PAMI,* 1997.

[27] Ke Huang and Selin Aviyente. Sparse representation for signal classification. *International Conference on Neural Information Processing Systems*, 609-616, 2006.

[28] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Trans. on PAMI*, 2009.

[29] Zongbo Xie, Jiuchao Feng. KFCE: a dictionary generation algorithm for sparse representation. *Signal Processing*, 2009.

Mesure de qualité d'image par approche multiéchelle multidiectionnelle

Z. Haddad[1,2] A. Beghdadi[1] A. Serir[2] A.Mokraoui[1]

[1] L2TI (Laboratoire de Traitement et de Transport de l'Information)

Institut Galilée, Université Paris 13
99 avenue Jean Baptiste Clément, Villetaneuse – France

[2] LTIR (Laboratoire de Traitement d'Image et Rayonnement)

Institut d'électronique et d'informatique, USTHB
Bp 32, El Alia 16111 Bab ezzouar, Alger – Algérie

{zehira.haddad, beghdadi}@univ-paris13.fr

Résumé

Evaluer la qualité d'une image revient à estimer le changement de certaines caractéristiques de cette image. L'étape d'extraction des caractéristiques constitue donc l'étape la plus importante dans l'établissement d'une mesure de qualité d'images. La transformée en curvelet est actuellement l'une des transformées qui représente le mieux une image et ses principales caractéristiques. Ce document présente une nouvelle métrique d'évaluation de la qualité d'images basée sur la transformée en curvelet. La métrique proposée a été testée sur la base de données LIVE et a été comparée à d'autres mesures d'évaluation de la qualité d'images. Les résultats obtenus montrent l'intérêt et la fiabilité de cette mesure.

Mots clefs

Qualité d'images, Système visuel humain HVS, transformée en curvelet, DMOS, fonction logistique.

1 Introduction

L'objectif des méthodes d'évaluation de la qualité d'images est de proposer une mesure qui permet d'évaluer la qualité des images d'une manière cohérente tout en étant en accord avec les jugements subjectifs de l'œil humain. Dans de nombreuses applications de traitement d'image telles que la compression avec perte, le tatouage numérique ou le rehaussement d'images, il est nécessaire de pouvoir quantifier l'impact visuel de la dégradation engendrée.

Les deux plus anciennes méthodes, qui sont malheureusement les plus utilisées, pour mesurer la qualité d'une image sont le PSNR et l'erreur quadratique moyenne (EQM). Ces deux métriques sont très simples et faciles à calculer. Toutefois, ces métriques ne sont pas toujours en accord avec l'appréciation visuelle d'un opérateur humain. En effet, plusieurs études ont montré que des images peuvent avoir la même EQM ou le même PSNR que l'image originale tout en présentant des qualités perceptuelles très différentes. De même, les petites déformations géométriques peuvent facilement créer une EQM très élevée sans avoir de conséquences sur la qualité de l'image. De ce fait, plusieurs recherches ont été dédiées à l'élaboration de méthodes plus efficaces pour évaluer la qualité d'images. Le but étant de chercher des méthodes plus en accord avec le jugement humain.

Les mesures de qualité d'images sont répertoriées selon trois catégories principales: les mesures avec référence, les mesures sans référence et les mesures à référence réduite.

Les mesures avec référence [1], [2], utilisent toute l'information contenue dans l'image originale. Les premières métriques de cette catégorie ne sont pas en accord avec la perception visuelle de l'œil humain, tel est le cas du PSNR et de l'EQM. Lors de ces dernières années, de nombreuses études se sont orientées vers l'utilisation des principales propriétés du système visuel humain [17] pour l'établissement d'une mesure de qualité d'images. L'une des métriques les plus connues de cette catégorie est le SSIM, cette mesure de qualité structurelle reflète le fait que le système visuel humain soit sensible à l'extraction de l'information structurelle présente dans une image [3]. Les métriques qui ne nécessitent pas d'informations sur l'image originale sont répertoriées comme étant des métriques sans référence [4], [5], [6]. Notons toutefois que l'inconvénient principal de ces métriques est qu'elles sont généralement dédiées qu'à un certain type de dégradation. La dernière catégorie de métriques [7] n'a besoin que d'une partie des informations sur l'image original.

Ces dernières années ont vu l'émergence de différentes nouvelles transformées appelées transformées géométriques. Ces transformées représentent une

extension des ondelettes classiques dans un espace de représentation plus large. Leur principale caractéristique est qu'elles permettent de corriger certains inconvénients des ondelettes classiques [10] dans le cadre de la représentation d'objets anisotropes.

Des études assez récentes ont montré que l'une de ces transformées, la transformée en contourlet, a donné de très bons résultats dans le domaine de la qualité d'images. [20] et [21] offrent deux façons différentes d'utiliser cette transformée pour élaborer une mesure de qualité d'image. Dans le premier article, les auteurs utilisent une transformée en contourlet, tandis que dans le second les auteurs utilisent une transformée en contourlet basée sur la transformée en ondelettes. En étudiant de près ces nouvelles transformées, nous pensons que la transformée en curvelet peut être très intéressante et tout à fait appropriée au développement d'une mesure de qualité d'image. Les principales propriétés de cette transformée sont la richesse de sa représentation et sa pertinence pour la représentation des contours et des courbes dans une image. Etant donné que les dégradations les plus visibles dans une image se situent au niveau de ses contours, les propriétés des curvelets sont des atouts non négligeables pour l'élaboration d'une mesure de qualité d'image. Ce document vise à évaluer la qualité d'image en utilisant la représentation multi-échelle et multidirectionnelle des curvelets. Le fait d'évaluer la qualité d'image dans une représentation aussi riche permet d'extraire plus de caractéristiques représentatives de l'image pour mieux l'évaluer. De plus, la métrique proposée utilise différents paramètres reflétant certaines propriétés de l'image et introduisant des propriétés spécifiques du système visuel humain.

Ce papier est organisé comme suit. Dans la section 2, nous introduisons brièvement la transformée en curvelet. La métrique d'évaluation de la qualité d'image proposée est présentée dans la section 3. La section 4 est dédiée à la présentation et à la discussion des résultats obtenus. La section 5 conclut ce travail.

2 Transformée en curvelet

La transformée en ondelettes est un outil très intéressant qui a fait ses preuves en traitement d'images. Le succès de cette transformée a été tel que son utilisation s'est étendue dans différent domaines tels que le filtrage [11], le rehaussement, la reconnaissance de formes ou la compression, [12], [13]. Toutefois, il faut savoir que les ondelettes présentent quelques inconvénients. En effet, l'extension des ondelettes dans le domaine 2D est généralement effectuée par un produit tensoriel simple séparable. Ce qui génère plusieurs coefficients de fortes énergies le long des contours d'une image (figure 1). Pour remédier à cet inconvénient, de nouvelles transformées multi-échelles ont été développées [14]. On peut distinguer deux catégories différentes, des approches adaptative et non adaptatives, celles basées sur des bancs de filtres directionnels fixe, et celles qui utilisent un modèle géométrique indiquant une orientation d'analyse locale.

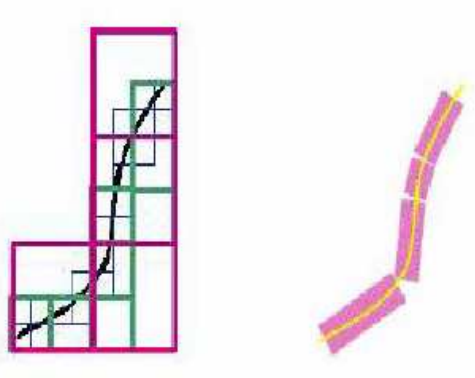


Figure 1 – *Différence entre une représentation en ondelettes et une représentation géométrique appropriée.*

La transformée en curvelet est une généralisation de la transformée en ridgelets. Donc, pour bien comprendre la transformée en curvelet il faut d'abord définir la transformée en ridgelet.

La transformée en Ridgelet [15] est définie comme une ondelette $\Psi_{a,\theta,b}$ construite le long d'une ligne orientée d'un angle θ et définie dans le plan cartésien (x_1, x_2) par l'équation:

$$\Psi_{a,\theta,b} = a^{-1/2}\Psi\left(\left(x_1 * \cos(\theta) + x_2 * \sin(\theta) - b\right)/a\right) \quad (1)$$

Les coefficients en ridgelet Rid_f d'une fonction f sont obtenus par projection sur cette base:

$$Rid_f(a,\theta,b) = \iint f(x_1,x_2)\Psi_{a,\theta,b}(x_1,x_2)\,dx_1dx_2 \quad (2)$$

Cette projection est étroitement liée à la transformée de Radon qui consiste à intégrer une image selon un ensemble de lignes:

$$Rad_f(t,\theta) = \int f(x_1,x_2)\ \delta(-x_1\sin\theta + x_2\cos\theta - t)\,dx_1dx_2 \quad (3)$$

Par conséquent, la transformée en ridgelet peut être considérée comme une transformée en ondelettes 1D de la transformée de Radon le long de l'axe de translation t :

$$Rid_f(a,\theta,b) = \int Rad_f(t,\theta)\ a^{-1/2}\Psi\left((t-b)/a\right)\ dt \quad (4)$$

La transformée en Ridgelet a été développée pour analyser des objets qui contiennent des discontinuités linéaires. Les contours d'une image sont dans la plupart du temps curvilignes et non rectilignes. Sachant qu'une courbe peut être représentée par plusieurs segments de droite. En se basant ainsi sur le fait qu'une image est supposée contenir localement des contours rectiligne, la transformée en ridgelet peut être généralisée du cas linéaire au cas curviligne. C'est l'idée de la transformée en curvelet. L'objectif de la transformée en curvelet [16] est de décrire l'image comme des petites parties d'une certaine taille et

d'orientation donnée. Pour cela, une analyse multirésolution est appliquée à l'image avant de lui appliquer la transformée en ridgelet localement sur des blocs dyadiques.
La transformée en curvelet effectue d'abord une analyse multi-échelles en différents niveaux *K*.

$$f = A_f^{[K]} + \sum_{k=1}^{K-1} HF_f^{[k]} \quad (5)$$

Puis, on applique à chaque image haute fréquence $HF_f^{[k]}$, $k = K-1,\cdots,1$ une transformée en ridgelet locale en passant par les différentes étapes suivantes:
- Initialisation de la taille du bloc $B_1 = B_{min}$.
- Pour chaque image haute fréquence $HF_f^{[k]}$, $k = K-1,\cdots,1$, appliquer localement la transformée en ridgelet.

La taille de chaque bloc peut changer d'une sous bande à l'autre selon les conditions suivantes :
-Si $B^{[k]} \bmod 2 = 1$, $B^{[k+1]} = 2B^{[k]}$,
-Sinon, $B^{[k+1]} = B^{[k]}$.
Un curvelet est définie comme une fonction $x = f(x_1, x_2)$ à l'échelle 2^{-j}, de l'orientation θ_l et à la position $x_k^{(j,l)} = R_{\theta,l}^{-1}\left(k_1 2^{-j}, k_2 2^{-j/2}\right)$ par:

$$\varphi_{j,l,k}(x) = \varphi_j\left(R_{\theta_l}\left(x - x_k^{(j,l)}\right)\right) \quad (6)$$

La transformée en curvelet est définie par [16]:

$$c(j,l,k) = \left\langle f, \varphi_{i,l,k} \right\rangle = \int_{\square^2} f(x)\, \varphi_{i,l,k}(x)\, dx \quad (7)$$

La figure 2 représente la décomposition en curvelet.

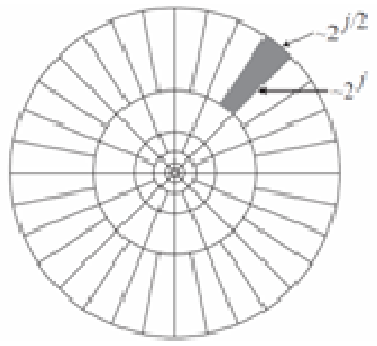


Figure 2 – *La décomposition en curvelet.*

3 La métrique de qualité d'image proposée

L'établissement d'une métrique de qualité d'image suit les deux étapes suivantes. Tour d'abord, nous décomposons l'image de référence et l'image dégradée en différentes sous-bandes et différentes orientations en utilisant la transformée en curvelet. Puis, nous calculons les dégradations, les variations des coefficients visuellement sensibles dans chaque sous-bande. L'utilisation de la transformée en curvelet pour la représentation d'images permet d'évaluer les distorsions dans un espace de représentation très riche. Par conséquent, elle permet de mieux évaluer la qualité d'image.
Soit f l'image originale et $\hat{f}$ l'image dégradée. La décomposition en curvelet donne plusieurs coefficients $c_k^{\theta}(x_k, y_k)$ correspondent à l'échelle k et à l'orientation θ.
La métrique de qualité d'image proposée est définie par la formule suivante:

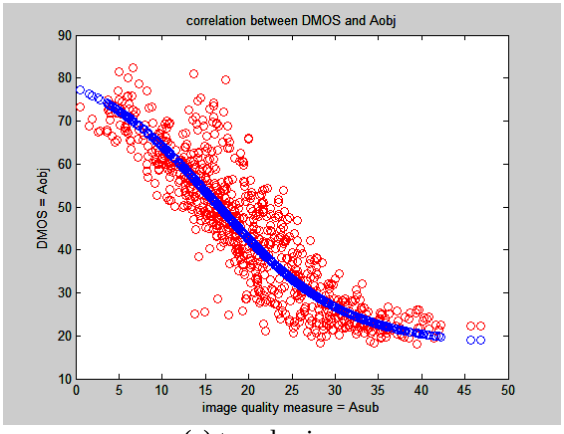
$$mcurv = 20\log_{10}\left(\frac{\sum_{x,y,\theta} \max_k 2^{-ksl} \left|c_k^{\theta}(x_k, y_k)\right|^l}{\sum_{x,y,\theta} \max_k 2^{-ksl} \left|c_k^{\theta}(x_k, y_k) - \hat{c}_k^{\theta}(x_k, y_k)\right|^l}\right)^{1/l} \quad (8)$$

Nous avons choisi, comme modèle de la métrique d'évaluation de la qualité d'images, une sorte de rapport signal sur bruit appliqué aux coefficients en curvelet pour chaque orientation et chaque échelle. Ce modèle comporte des paramètres différents qui utilisent nos connaissances sur le système visuel humain.
Dans le but d'avoir une sensibilité visuelle uniforme se rapprochant de la perception humaine pour chaque échelle et chaque orientation, nous introduisons un masquage perceptuel pour chaque sous-bande via le facteur 2^{-ksl}. De nombreuses études ont été faites sur la discernassion de la texture par le système visuel humain [18], nous simulons cet effet par le paramètre *s* qui est adapté à la structure de l'image à évaluer.

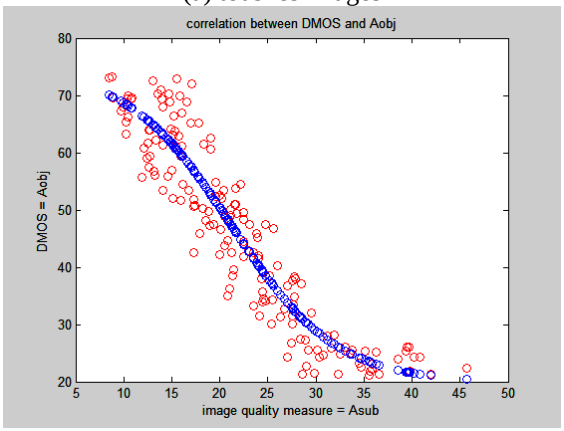
4 Les résultats expérimentaux obtenus

Nous testons la métrique de qualité d'images proposée sur la base de données LIVE [22]. Cette base de données se compose de 20 images originales et de 779 images dégradées. Cette base de données est notamment subdivisée en cinq classes. Chaque classe correspond à un type de distorsion particulier tout en comprenant des images originales et des images dégradées. La base de données comprend 227 images correspondant à des dégradations dues à la compression JPEG2000, 233 images correspondant au format JPEG, 174 images correspondant au flou gaussien, 174 images correspondant à un bruit blanc et 174 images correspondant à des erreurs de bit dans JPEG2000. Cette base de données fournit les évaluations subjectives (DMOS) d'environ 25000 individus. Nous jugeons qu'une mesure d'une qualité d'image est fiable si elle est bien corrélée à l'appréciation visuelle d'un opérateur humain.
Généralement, pour la validation d'une métrique de qualité d'image, une fonction logistique est utilisée pour ajuster les mesures de qualité d'image objectives en minimisant

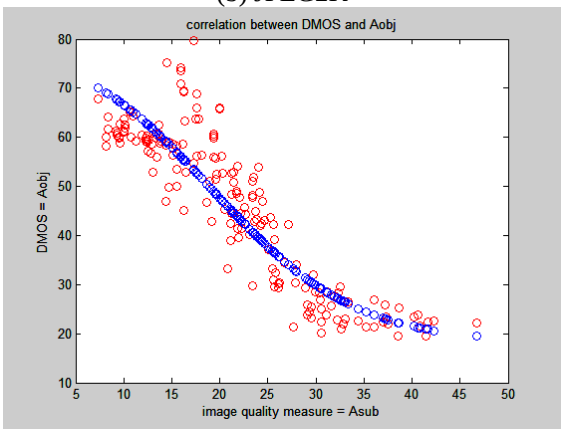
l'erreur quadratique moyenne entre les mesures objectives et subjectives. Dans cette étude, nous utilisons une fonction sigmoïde.
La figure 3 représente les résultats obtenus des évaluations objectives et subjectives sur la base de données LIVE. Les critères de validation choisis sont le coefficient de corrélation de Spearman, l'erreur quadratique moyenne et le coefficient de corrélation classique.



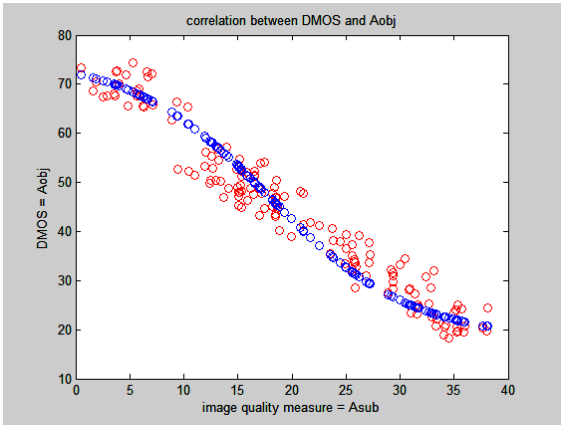
(a) tous les images



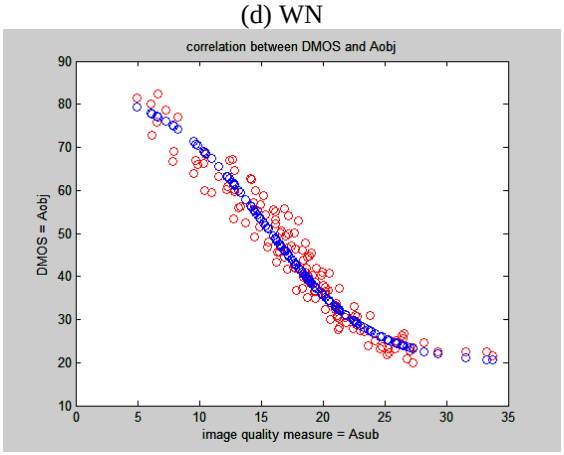
(b) JPEG2K



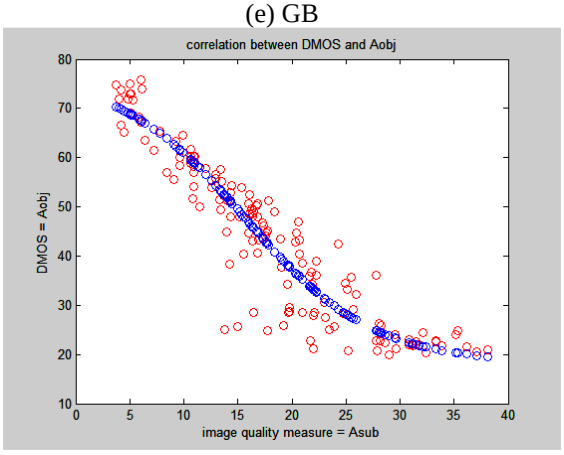
(c) JPEG



(d) WN



(e) GB



(f) FF

Figure 3 – *Evaluations subjectives et objectives de la distorsion perçue dans les différentes classes de la base LIVE. L'axe verticale représente le DMOS et l'axe horizontale représente la métrique proposée. Le tracé bleu représente la fonction logistique.*

Le tableau 1 représente les résultats de corrélation obtenus de la mesure proposée avec les mesures subjectives données par la base de données LIVE en utilisant la fonction logistique.

Dans le but de comparer les résultats obtenus de la métrique proposée avec d'autres métriques, nous donnons les résultats de corrélation des mesures de qualité d'images wssim [19], wbct [20] et msdd [21] avec la même base de données. Les résultats montrent que la métrique proposée est meilleure que les précédentes métriques pour les dégradations JPEG2K, GB et FF. On note également que pour toutes les dégradations confondues (la case ALL), la métrique proposée obtient les meilleurs résultats de corrélation.

Correlation coefficient						
	jpg2k	Jpeg	Wn	Gb	Ff	All
mcurv	0.941	0.898	0.968	0.969	0.929	0.893
wssim	0.940	0.935	0.962	0.952	0.953	0.884
wbct	0.914	0.880	0.970	0.377	0.812	0.674
msdd	0.914	0.925	0.958	0.950	0.923	0.890
Root Mean Squared Error						
mcurv	5.431	6.956	4.177	3.840	5.999	7.247
wssim	5.504	5.652	4.357	4.829	4.983	7.532
wbct	6.559	7.603	3.929	14.58	9.591	11.90
msdd	6.584	6.063	4.556	4.904	6.314	7.341
Rank order correlation coefficient						
mcurv	0.934	0.894	0.966	0.973	0.931	0.897
Wssim	0.931	0.899	0.957	0.960	0.962	0.879
Wbct	0.919	0.825	0.979	0.312	0.782	0.624
Msdd	0.899	0.883	0.946	0.948	0.923	0.890

Tableau 1 – *Résultats de corrélation de la métrique proposée avec la base de données LIVE.*

5 Conclusion

Ce document présente une métrique de qualité d'image multi-échelle multidirectionnelle. Cette métrique utilise les distorsions dans les différentes échelles et orientations de la décomposition en curvelet. Évaluer les distorsions dans une zone de représentation très riche permet de mieux évaluer la qualité d'image et de rendre la métrique de qualité très discriminante. Les tests de validation ont été effectués sur la base de donnée LIVE. Cette base de données a été spécialement conçue pour l'évaluation des performances des métriques de qualité d'images. Les résultats obtenus confirment l'efficacité de la métrique proposée. Le fait d'introduire dans l'établissement de la métrique, des propriétés connues du système visuel humain, rend cette métrique plus fiable et plus proche du jugement humain. Dans le cadre de travaux futurs, nous pensons que l'application de la métrique proposée peut être très intéressante pour des images qui contiennent beaucoup de contours ou des courbes comme dans le cas d'images d'empreintes digitales. Nous proposons également de tester la mesure sur d'autres bases de données et d'analyser en profondeur le fonctionnement de la métrique en s'attachant à l'aspect multi-résolution afin de tirer profit de la similarité entre la représentation en curvelets et représentation fréquentielle effectuée au niveau du cerveau.

Références

[1] I. Avcibas, B. Sankur, K. Sayood, "Statistical evaluation of image quality measures", *Journal of Electron Imaging* 11, 206–213, 2002.

[2] H.R. Sheikh, A.C. Bovik, G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics", *IEEE Trans. Image Process*. 14, 2117–2128, 2005.

[3] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, "Image quality assessment: from error visibility to structural similarity", *IEEE Trans. Image Process*. 13, 600–612, 2004.

[4] H. Sheikh, A. Bovik, L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000", *IEEE Trans. Image Process*. 14, 1918–1927, 2005.

[5] Z. Wang, A.C. Bovik, B.L. Evans, "Blind measurement of blocking artifacts in images", *ICIP*, Vancouver, Canada, pp. 981–984, 2000.

[6] H. Sheikh, Z. Wang, L. Cormack, A. Bovik, "Blind quality assessment for JPEG2000 compressed images", *ICIP*, New York, pp. 1735–1739, 2002.

[7] Z. Wang, E. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model", *Human Vision and Electronic Imaging* X, Proceedings, vol. 5666, pp. 149–159, 2005.

[8] A. Beghdadi, B. Pesquet-Popescu, "A new image distortion measure based on wavelet decomposition". *ISSPIT*, Vol. 1, pp. 485- 488, 2003.

[9] R. Eslami, H. Radha, "Wavelet-based contourlet transform and its application to image coding", *ICIP*, Singapore, pp. 3189–3192, 2004.

[10] E. Candes, D. Donoho, "Recovering edges iill-posed inverse problems: optimality of Curvelet ", *annals of statistics*, vol. 30, no. 3, pp. 784-842, 2002.

[11] Starck J, Candès E, Dohono D. "The curvelet transform for image denoising", *IEEE trans. on image processing*, 11, pp. 670-684, 2000.

[12] Hilton M, L., Jawerth B. D, et Sengupta A., "Compressing still and moving images with wavelets", *Multimedia systems*, vol. 2, pp.218-227, 1994.

[13] Marcellin M. W., Gormish M. J., Biling. A et Boliek M. P., "An overview of JPEG 2000", *Data Compression Conference*, pp. 523-544, 2000.

[14] H. Führ, L. Demaret, F. Friedrich, "Beyond wavelets: New image representation Paradigms ," *Survey article*, preprint version, 2005.

[15] E. Candes, D. Donoho, "Ridgelets: A key to higher-dimensional intermittency?" *Philosophical transactions Royal Society. Mathematical, physical and engineering sciences*, vol. 357, no. 1760, pp.2495- 2509, 1999.

[16] E. Candes, D. Donoho, “Curvelets - A surprisingly effective nonadaptive representation for objects with edges, curves and surfaces”, *Curves and Surfaces*, Vanderbilt University Press, Nashville, TN, 1999.

[17] “Special Issue on Image Quality Assessment”, *Signal Processing*, Vol 70, 1998.

[18] J. Malik and P. Perona, “Preattentive texture discrimination with early vision mechanisms”, *Journal Opt. Soc. Amer*. A, vol.7, pp.923–932, 1990.

[19] Z. Wang and EP Simoncelli. “Translation insensitive image similartity in complex wavelet domain. *ICASSP*, Vol. 2, pp. 573-576, March 2005

[20] X. Gao, W. Lu and D. Tao “Wavelet based contourlet in quallity evaluation of digital images” *Neurocomputing*, 72 (1-3) pp, 378-385, 2008.

[21] Mingna Liu and Xin Yang, “Image quality assessment using contourlet transform”, *optical engineering*, 48(10)107201, October 2009.

[22] http://live.ece.utexas.edu/

Efficacité énergétique d'une DCT zonale rapide dans le contexte de la compression d'image dans les réseaux de capteurs sans fil

L. Makkaoui[1] V. Lecuire[1] J-M. Moureaux[1]

[1] Centre de Recherche en Automatique de Nancy (CRAN)

Nancy-Université, CNRS
Campus Sciences, BP 70239
F-54506 Vandœuvre-les-Nancy Cedex, FRANCE

{leila.makkaoui, vincent.lecuire, jean-marie.moureaux}@cran.uhp-nancy.fr

Résumé

Cet article traite de la compression d'image dans le contexte d'application des réseaux de capteurs sans fil, où l'efficacité énergétique est un critère dominant pour la durée de vie des noeuds caméra, aussi bien que pour la durée de vie du réseau tout entier. Nous proposons d'intégrer dans une chaîne de compression de type JPEG une DCT zonale rapide, autrement dit de combiner une méthode de sélection zonale des coefficients avec une méthode de DCT rapide. Cela réduit le nombre de coefficients à calculer, à quantifier et à coder dans chaque bloc, entraînant mécaniquement des économies d'énergie sur toute la chaîne de compression. Les résultats obtenus sur le microcontrôleur de référence MSP430 (il a été adopté dans les noeuds Telos développés à Berkeley) montrent que des économies d'énergie importantes sont possibles, environ 7.5% à qualité d'image égale, et jusqu'à 40%, au prix d'une distorsion de l'image plus grande mais restant tout à fait acceptable pour beaucoup d'applications.

Mots clefs

Compression d'images fixes, transformée en cosinus discrète, réseaux de capteurs sans fil, consommation d'énergie.

1 Introduction

Les réseaux de capteurs sans fil représentent une révolution technologique qui change radicalement la façon de concevoir les systèmes de surveillance de très grandes échelles. En effet, les progrès croissants de la microélectronique et des communications par ondes radio permettent de fabriquer des *noeuds capteurs* de plus en plus petits, embarquant une unité de traitement de données, un module de transmission sans fil et une batterie. Ces capteurs abandonnés sur le terrain vont s'organiser en réseau de manière spontanée et collaborer entre eux pour récolter des données (grandeur physique mesurée) et les acheminer jusqu'à un point de collecte (noeud puits qui fait l'interface entre le réseau de capteurs et l'utilisateur consommateur des données). Les applications potentielles des réseaux de capteurs sans fil sont nombreuses. Citons par exemple les surveillances environnementales (activité volcanique, sismique, climatique, etc.) et militaires (détection des mouvements ennemis, localisation et traque de cibles). Un problème fondamental dans les réseaux de capteurs sans fil se rapporte à la consommation d'énergie, l'hypothèse communément retenue étant qu'il est impossible de recharger ou remplacer les batteries des noeuds capteurs une fois qu'ils ont été abandonnés sur le terrain. Par conséquent, tout noeud ayant épuisé sa batterie devient définitivement inutilisable, si bien que la couverture et la connectivité du réseau va diminuer inexorablement au cours du temps jusqu'à atteindre le point de rupture. L'efficacité énergétique doit donc être recherchée, pour optimiser la durée de vie des noeuds (critère local) aussi bien que la durée de vie du réseau tout entier (critère global).

Parmi toutes les applications potentielles des réseaux de capteurs sans fil, celles utilisant des capteurs d'image sont appréciables pour tout ce qui concerne la reconnaissance, la localisation et le dénombrement d'objets par la vision [1]. Des noeuds capteurs dotés d'une caméra existent déjà, comme Cyclops [2] qui est un prototype de laboratoire, et le capteur multimédia IMB vendu par Crossbow [3]. Mais les applications basées sur des capteurs d'image sont particulièrement gourmandes en énergie puisque la quantité de données nécessaire pour représenter une image diffère de plusieurs ordres de grandeur en comparaison d'une valeur scalaire classique (une mesure de température par exemple). Comme le transceiver radio est un des composants les plus gourmands en énergie, il semble évident que le coût d'énergie de la transmission de l'image peut être réduit significativement en compressant l'image à la source. Toutefois, les noeuds capteurs sont par nature très limités en capacité de calcul et de mémoire, c'est une conséquence directe des contraintes de consommation d'énergie, de miniaturisation et de coût de fabrication. Ils ne peuvent donc pas implanter des algorithmes de compression de trop grande complexité. De plus, des travaux

comme ceux de Ferrigno et al. [4] ont montré que des algorithmes comme JPEG2000, SPIHT et même JPEG sont inefficaces sur la plupart des plateformes logicielles utilisées dans les réseaux de capteurs, c'est-à-dire que l'opération de compression de l'image coûte plus d'énergie qu'une transmission de l'image non compressée.

Dans cet article, nous considérons une chaîne de compression d'image basée sur l'algorithme bien connu JPEG, et nous proposons de combiner une méthode de DCT rapide avec une méthode de sélection zonale des coefficients. Une telle combinaison, que nous désignons sous le terme de *DCT zonale rapide*, réduit le nombre de coefficients à calculer, et donc à quantifier et à encoder. Précisons que les méthodes de DCT rapides et de sélection zonale des coefficients ont déjà été largement étudiées dans la littérature [5, 6, 7, 8, 9, 10, 11]. Mais à notre connaissance, il n'y a pas de travaux qui vise à les associer et leur performance combinée reste à évaluer. Notre contribution se situe là, en particulier :

– Nous comparons deux modèles de sélection zonale des coefficients de la DCT, l'un utilisant une zone de forme carrée et l'autre une zone de forme traingulaire ;
– Nous évaluons la complexité de la DCT rapide de Loeffler en fonction de la dimension de la zone de sélection des coefficients ;
– Nous évaluons la consommation d'énergie de notre chaîne de compression pour le microcontrôleur MSP430 de Texas Instruments, qui fait référence dans les noeuds de capteurs sans fil.

2 Proposition

Les algorithmes de compression d'image basés sur la transformée en cosinus discrète (DCT) avec des blocs de 8×8 pixels sont très populaires mais cette transformée est coûteuse en calcul, et donc gourmande en énergie. Dans une chaîne de compression conventionnelle <Transformée - Quantification - Codage> implantée logiciellement, l'étape de la DCT représente à elle seule environ 60% du coût total en énergie [12]. De nombreux travaux traitent de la réduction du coût de calcul de la DCT. Une approche consiste à réécrire la transformée sous forme matricielle et à factoriser la décomposition de manière à réduire le nombre de multiplications scalaires nécessaires. Pour la DCT 1-D, l'algorithme de Loeffler-Ligtenberg-Moschytz (LLM) [5], avec 11 multiplications et 29 additions seulement, est le plus efficace (il a été prouvé qu'on ne peut pas faire moins que 11 multiplications). Le graphe des flux de l'algorithme LLM est présenté Figure 1. La DCT 2-D peut être obtenue en appliquant d'abord la DCT 1-D sur chaque ligne du bloc de pixels puis sur chacune des colonnes du bloc résultant. Avec l'algorithme LLM, la DCT 2-D nécessite 176 multiplications et 464 additions. Comme les multiplications sont des opérations coûteuses en énergie, quelques algorithmes tels que BinDCT [7], Cordic DCT [8] et Cordic Loeffler DCT [9] opèrent dans l'espace des valeurs entières et utilisent des additions et des décalages en lieu et place des multiplications. Cela coûte moins d'énergie mais les valeurs obtenues pour les coefficients de la DCT sont des approximations des valeurs exactes. Cela se fait donc au prix d'une plus grande distorsion de l'image. De toutes ces méthodes approchées, c'est l'algorithme de Cordic Loeffler DCT, avec 38 additions et 16 décalages (soit 608 additions et 256 décalages en 2-D), qui fournit le meilleur compromis entre la complexité de calcul et la distorsion de l'image. Dans le domaine de la DCT 2-D, Feig et Winograd ont proposé l'un des algorithmes les plus rapides pour des blocs de 8×8 pixels, qui nécessite seulement 94 multiplications et 454 additions. D'une manière générale, les approches considérant directement la DCT 2-D ont une complexité de calcul moindre que les approches 1-D par lignes puis colonnes mais elles exigent un espace mémoire important pour stocker les résultats intermédiaires. Les nombreux échanges mémoire qui sont sollicités par le microcontrôleur peut dégrader la vitesse d'exécution drastiquement.

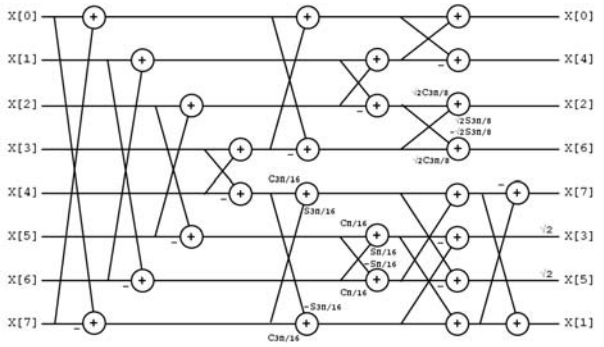
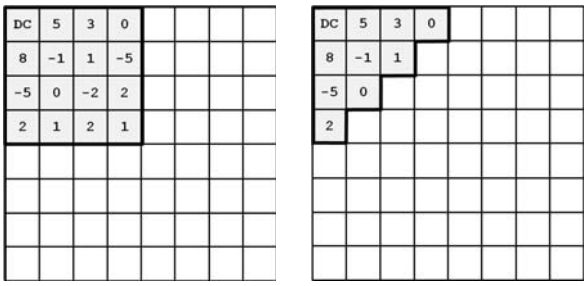


Figure 1 – *Graphe des flux de l'algorithme LLM.*

Une autre approche pour réduire le coût de calcul de la DCT et des étapes qui suivent dans la chaîne de compression est basée sur le codage zonal. Cela consiste à opérer sur un sous-ensemble de coefficients situés dans une zone du bloc 8×8 bien délimitée. Il s'agit de sélectionner les coefficients les plus importants, c'est-à-dire ceux de basses fréquences, pour calculer, quantifier, encoder et transmettre seulement ceux-là. Dans [10] par exemple, la zone de sélection des coefficients correspond au carré de longueur k, avec $k < 8$, situé dans la partie supérieure gauche du bloc 8×8 (voir Figure 2(a)). Il y a ici k^2 coefficients à traiter et les autres sont négligés (ils seront implicitement mis à zéro côté décodeur). Plus le paramètre k est petit, et plus le temps pour exécuter la DCT, puis quantifier et coder les coefficients sera rapide. Les effets de la variation du paramètre k et du niveau de quantification sur la consommation d'énergie, la latence, la qualité de l'image et le débit sont étudiés dans [11]. Dans [13], le même principe est repris, mais cette fois-ci, la zone de sélection des coefficients est définie par un triangle rectangle de cathète k, situé dans la partie supérieure gauche de la zone carrée précédemment mentionnée (voir Figure 2(b)). Dans cette

forme, il y a seulement $\frac{1}{2}k(k+1)$ coefficients significatifs au lieu de k^2. Remarquons que pour une valeur de k donnée, la forme triangulaire va permettre de diminuer le temps de calcul pour les étapes de quantification et de codage des coefficients comparée à la forme carrée, puisqu'il y a moins de coefficients à traiter, mais le coût de la DCT 2-D reste exactement le même. Que la forme soit carrée ou triangulaire en effet, une DCT 1-D sera appliquée d'abord sur les 8 lignes du bloc de pixels puis sur chacune des k colonnes du bloc résultant.



(a) Forme carrée (b) Forme triangulaire

Figure 2 – *Differentes formes de DCT zonale avec $k = 4$.*

Les deux approches présentées, approche par factorisation de la matrice et approche par sélection zonale des coefficients, sont complémentaires et peuvent être combinées. Nous désignerons une telle combinaison comme étant une *DCT zonale rapide*. Nous allons d'abord étudier les effets de la variation du paramètre k sur le coût de calcul des coefficients de la DCT, pour des zones des deux formes. Dans cet article, nous avons adopté l'algorithme LLM comme point de départ de notre analyse mais celle-ci peut être reproduite pour d'autres algorithmes de DCT rapide. Comme nous l'avons déjà indiqué, la DCT zonale rapide consiste à calculer seulement les coefficients les plus significatifs, ceux de basses fréquences. En considérant l'algorithme LLM, cela se traduit par réduire le nombre de ses sorties à k. Un exemple est montré Figure 3 pour le cas où $k = 4$.

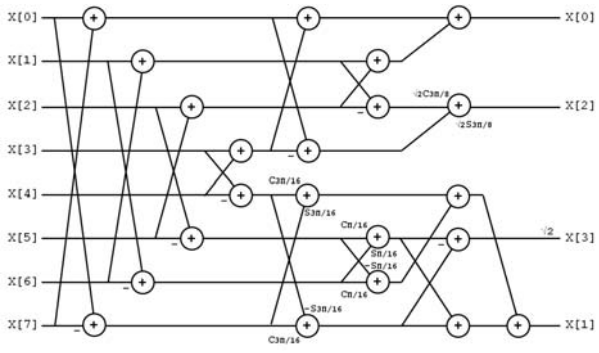


Figure 3 – *Graphe des flux de la DCT zonale avec l'algorithme LLM (avec $k = 4$).*

On constate qu'avec $k = 4$, le nombre d'opérations est réduit à 9 multiplications et 24 additions. Comparé à l'algorithme LLM original, les économies de calcul peuvent paraître modestes, 2 multiplications et 5 additions en moins seulement. Lorsqu'on projette les résultats à la DCT 2-D cependant, la réduction du coût de calcul devient franchement significative puisque on gagne 68 multiplications et 176 additions. Cela fait une économie d'environ 38% sur chaque type d'opération. La table 1 donne le nombre d'opérations nécessaires pour calculer la DCT zonale rapide, en 1-D et en 2-D, pour les différentes valeurs de k.

	Coût de la DCT 1-D		Coût de la DCT 2-D	
	mult.	add.	mult.	add.
LLM (référence)	11	29	176	464
LLM Zonal (k=7)	11	28	165	420
LLM Zonal (k=6)	10	26	140	364
LLM Zonal (k=5)	9	25	117	325
LLM Zonal (k=4)	9	24	108	288
LLM Zonal (k=3)	8	23	88	253
LLM Zonal (k=2)	6	20	60	200

Tableau 1 – *Coût de calcul de la DCT zonale rapide.*

Reste à comparer l'efficacité énergétique des deux formes de DCT zonale rapide. Pour une valeur de k donnée, la forme triangulaire est plus économique en énergie puisqu'il y a moins de coefficients à quantifier et à encoder. En contrepartie, la distorsion de l'image sera plus élevée, notamment pour des niveaux de quantification de moyenne et de haute qualité, puisqu'un plus grand nombre de coefficients sont négligés. Toutefois, l'efficacité énergétique se rapporte au rapport énergie distorsion pour un débit donné, ou encore au rapport énergie débit pour une distorsion donnée. Nous avons mené une série d'expériences sur plusieurs images tests, incluant Lena, Peppers et Baboon pour les débits de 0.5 et 1 bpp. Dans tous les cas, il apparaît que la forme carrée est la plus efficace. Cela s'explique facilement : Les résultats montrent en effet que la distorsion de l'image obtenue est pratiquement la même dans les deux formes lorsque le nombre de coefficients à traiter est similaire. Par exemple, pour la forme carrée avec $k = 4$, il y a 16 coefficients à traiter et pour la forme triangulaire avec $k = 5$, il y en a 15. Ces deux scénarios sont comparables à débit constant puisqu'ils donnent à peu près le même niveau de distorsion (pour l'image Lena à 0.5 bpp, le PSNR est de 33.52 dB pour la forme carrée et de 33.35 dB pour la forme triangulaire). Et au final, le coût de la DCT est en faveur de la forme carrée, alors que les coûts de la quantification et du codage des coefficients sont très proches. Prenons un autre exemple symbolique : Pour la forme carrée avec $k = 6$ et la forme triangulaire avec $k = 8$, il y a 36 coefficients à traiter. Toujours pour l'image Lena à 0.5 bpp, les PSNR sont très proches (34.26 dB pour la forme carrée et 34.28 dB pour l'autre). Et là encore, le coût énergetique est en faveur de la forme carrée.

3 Evaluation de performance

Les performances de la DCT zonale rapide ont été évaluées en considérant une chaîne de compression de type JPEG, et en utilisant les tables de quantification et du codage

de Huffman qui sont recommandées dans l'annexe de la norme JPEG [14]. Le premier critère de performance considéré était la distorsion de l'image compressée, à travers la mesure du PSNR. Les résultats sont présentés Figure 4 en prenant l'exemple de l'image Lena encodée au débit de 0.5 bpp. Les deux formes de zone, carrée et triangulaire, ont été utilisées avec plusieurs valeurs de k.

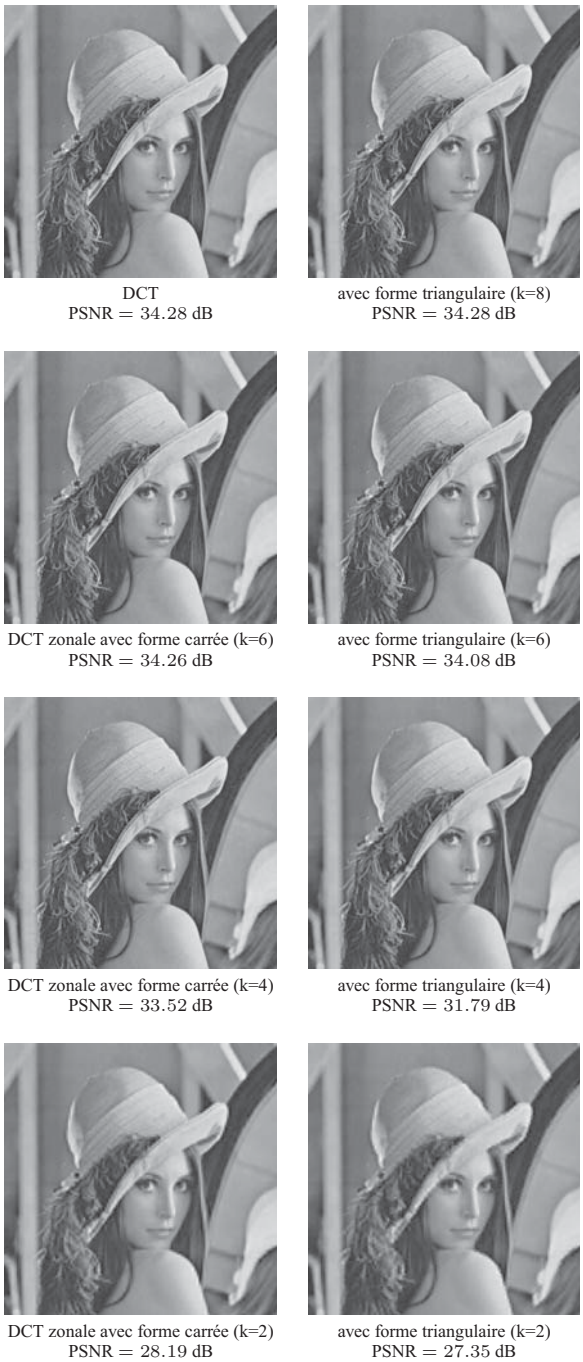


Figure 4 – *image Lena codée à 0.5 bpp, pour les deux formes zonales et différentes valeurs de k.*

Evidemment, la distorsion de l'image compressée augmente et le PSNR diminue quand k diminue. Pour des grandes valeurs de k (entre 6 et 8), et quelle que soit la forme adoptée, on constate aussi bien visuellement que dans les valeurs du PSNR que la qualité des images sont très proches de celle obtenue avec la DCT classique. Il y a donc matière à faire des économies d'énergie en utilisant une DCT zonale rapide avec un très faible impact sur la distorsion d'image. Pour des petites valeurs de k (4 et en dessous), la distorsion s'accroit rapidement lorsque k diminue. En fait, la qualité d'image reste visuellement acceptable pour beaucoup d'applications jusqu'à $k = 3$ pour la forme carrée et jusqu'à $k = 4$ pour la forme triangulaire. En dessous de ces valeurs, les effets de blocs classiques de la compression par DCT à bas débit deviennent trop marqués. Précisons que ces tendances ont été observées avec d'autres valeurs du débit et sur d'autres images tests.
Le deuxième critère de performance considéré était le temps d'exécution et le coût d'énergie de toute la chaîne de compression. Les résultats sont bien sûr dépendants de la plateforme utilisée pour les noeuds capteurs. Nous avons choisi la plateforme Telos [15] qui fait référence dans la communauté scientifique en réseaux de capteurs sans fil. Cette plateforme, qui a été développée par l'équipe du célèbre D. Culler à Berkeley (Université de Californie), est l'une des moins gourmandes en énergie qui existe actuellement. Elle est bâtie autour d'un microcontrôleur MSP430 qui peut fonctionner à 1.8V et consomme 3 mW en mode actif (l'horloge est cadencée à 8 MHz et le bus est de 16 bits). Telos est aussi équipé d'un circuit CC2420 qui est un transceiver radio compatible avec la norme 802.15.4. Le débit de transmission est de 250 kbps. La puissance consommée par ce circuit est de 35 mW en transmission à 0 dBm. Le coût d'énergie de la compression d'image est proportionnel au nombre de cycles exécutés par le microcontrôleur. Pour obtenir ce nombre de cycles, nous avons utilisé le simulateur WSim qui est développé par l'INRIA (disponible sur http ://gforge.inria.fr/projects/wsim/). Les résultats sont présentés Table 2 en considérant un débit de sortie de 0.5 bpp. Précisons qu'il s'agit des coûts de calcul par bloc de 8×8 pixels. Pour avoir les coûts correspondant au traitement d'une image entière, il suffit de multiplier les valeurs données par le nombre de blocs de l'image.

	DCT (cycles)	Quanti. (cycles)	Codage (cycles)	Total (cycles)	Temps (ms)	Energie (μJ)
LLM	246863	23980	79830	350673	43.8	132
Tri. 8	246863	14895	77058	338816	42.4	127
Carré 6	234303	14782	76926	326011	40.8	122
Tri. 6	234303	9898	75743	319944	40.0	120
Carré 4	217455	8136	75094	300685	37.6	113
Tri. 4	217455	6225	62667	286347	35.8	107
Carré 2	169503	4190	38167	211860	26.5	79
Tri. 2	169503	3876	33044	206423	25.8	77

Tableau 2 – *Temps d'exécution et consommation d'énergie pour compresser un bloc de 8×8 pixels à 0.5 bpp, pour différents scénarios de compression.*

Les valeurs de référence sont celles obtenues avec la chaîne de compression JPEG classique, c'est-à-dire lorsque les

64 coefficients de la DCT sont calculés, quantifiés et encodés. Il faut ainsi 43.8 ms au microcontrôleur MSP430 pour encoder un bloc, et la dépense énergétique est de 132 μJ. Comme prévu, c'est l'étape de la DCT qui est la plus coûteuse en calcul, elle prend environ 70% du total. Les valeurs associées à la forme carrée de la DCT zonale rapide sont les plus importantes à analyser puisque c'est la forme qui fournit la meilleure efficacité énergétique. Prenons d'abord le scénario avec $k = 6$. Ce scénario a un impact très faible sur la distorsion d'image comparé au scénario de référence (LLM). Mais les économies d'énergie sont déjà significatives puisqu'elles sont approximativement de 7.5%. Les applications qui tolèrent une distorsion plus importante de l'image peuvent utiliser des scénarios avec de plus petites valeurs de k. Les économies d'énergie s'avèrent alors très importantes : 14% dans le scénario avec $k = 4$ et 40% avec $k = 2$. De telles économies d'énergie vont permettre aux noeuds capteurs d'avoir une durée de vie beaucoup plus longue.

4 Conclusion

Dans cet article, nous avons étudié les performances d'une chaîne de compression de type JPEG qui intègre une DCT zonale rapide. Cette DCT zonale rapide réduit le nombre de coefficients à calculer, et donc à quantifier et à encoder. Elle entraîne mécaniquement une réduction de la complexité de calcul de la chaîne de compression, et par incidence une réduction de la consommation d'énergie sur le système hôte. Elle est particulièrement intéressante dans le contexte des réseaux de capteurs sans fil où le problème de la consommation d'énergie est dominant. Nous avons montré qu'une forme carrée fournit une meilleure efficacité énergétique qu'une forme triangulaire. Intégrée dans une chaîne de type JPEG, elle amène des économies d'énergie d'environ 7.5% à qualité d'image égale. Des économies plus importantes sont possibles au prix d'une distorsion d'image plus importante, mais qui reste acceptable pour de nombreux scénarios d'applications. Notre proposition prenait pour cible la DCT de Loeffler. Mais elle peut être reproduite pour d'autres méthodes de DCT rapide. Une comparaison des différentes variantes de DCT zonale rapide constitue la suite immédiate de nos travaux. Une comparaison de performance pour plusieurs plateformes de réseaux de capteurs sans fil doit aussi être réalisée.

Références

[1] S. Soro et W. Heinzelman. A survey of visual sensor networks. *Advances in Multimedia*, 2009 :Article ID 640386, 21 pages, 2009.

[2] M. Rahimi, R. Baer, O. I. Iroezi, J. C. Garcia, J. Warrior, D. Estrin, et M. Srivastava. Cyclops : In situ image sensing and interpretation in wireless sensor networks. Dans *ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pages 192–204, San Diego (CA), USA, November 2005.

[3] Crossbow Technology Inc.. http ://www.xbow.com/.

[4] L. Ferrigno, S. Marano, V. Paciello, et A. Pietrosanto. Balancing computational and transmission power consumption in wireless image sensor networks. Dans *IEEE Int. Conference on Virtual Environments, Human-Computer Interfaces, and Measures Systems (VECIMS)*, Giardini Naxos, Italy, July 2005. IEEE.

[5] C. Loeffler, A. Ligtenberg, et G. S. Moschytz. Practical fast 1-D DCT algorithms with 11 multiplications. Dans *Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 988–991, Glasgow, UK, May 1989.

[6] E. Feig et S. Winograd. Fast algorithms for the discrete cosine transform. *IEEE Transactions on Signal Processing*, 40(9) :2174–2193, September 1992.

[7] J. Liang et T. D. Tran. Fast multiplierless approximations of the dct with the lifting scheme. *IEEE Transactions on Signal Processing*, 49(12) :3032–3044, December 2001.

[8] H. Jeong, J. Kim, et W-K. Cho. Low-power multiplierless dct architecture using image correlation. *IEEE Transactions on Consumer Electronics*, 50(1) :262–267, February 2004.

[9] B. Heyne, C. C. Sun, J. Goetze, et S. J. Ruan. A computationally efficient high-quality cordic based dct. Dans *European Signal Processing Conference (EUSIPCO)*, Florence, Italy, September 2006.

[10] J. Bracamonte, M. Ansorge, et F. Pellandini. VLSI systems for image compression. a power-consumption/image-resolution trade-off approach. Dans *Conference on Digital Compression Technologies and Systems for Video Communications*, volume 2952, pages 591–596, Berlin, Germany, October 1996. SPIE.

[11] C. N. Taylor et S. Dey. Adaptive image compression for wireless multimedia communication. Dans *IEEE Int. Conference on Communications (ICC)*, volume 6, pages 1925–1929, Helsinki, Finland, June 2001.

[12] F. Marcelloni et M. Vecchio. A simple algorithm for data compression in wireless sensor networks. *IEEE Communications Letters*, 12(6) :411–413, June 2008.

[13] A. Mammeri, A. Khoumsi, D. Ziou, et B. Hadjou. Modeling and adapting JPEG to the energy requirements of VSN. Dans *Int. Conference on Computer Communications and Networks (ICCCN)*, pages 806–811, St. Thomas, Virgin Islands, USA, August 2008. IEEE.

[14] International Organization for Standardization. ITU-T recommendation T.81. http ://www.jpeg.org/jpeg/, 1992. ISO/IEC IS 10918-1.

[15] J. Polastre, R. Szewczyk, et D. Culler. Telos : enabling ultra-low power wireless research. Dans *Int. Symposium on Information Processing in Sensor Networks (IPSN)*, Los Angeles (CA), USA, April 2005. IEEE.

3D Objects Indexing and Retrieval Based On A New Efficient Optimal 2D Views Selection Method

M.A. Alaoui Mhamdi[1,4] A. Lachkar[2] S. El Alaoui Ouatik[3] D. Ziou[1]

[1] Département d'Informatique, Université de Sherbrooke, Canada

[2] Ecole Nationale des Sciences Appliquées de Fès, Morocco

[3] Faculté de Sciences Dhar El Mahraz, Fès, Morocco

[4] Laboratoire d'Informatique, Statistique et Qualité (LISQ), Morocco

Université sidi Mohamed Ben Abdellah,
Route d'Imouzzer B.P.2626 Fès – Morocco

ayoubalaoui@gmail.com, abdelmonaime_lachkar@yahoo.fr, S_ouatik@yahoo.com, Djemel.Ziou@usherbrooke.ca

Résumé

In this paper, we present a novel view-based approach for efficient 3D objects retrieval. A set of 2D images (multi-views) are automatically generated from the 3D object's views sphere approximated by a polyhedron subdivision loop scheme. We place the 3D object in it. To generate the initial views we place the camera on each of the triangle center of the polyhedron looking at the coordinate origin. For each 2D view associated with a triangle in the view sphere we apply the binarization to the 2D image and we extract the edge of the associated 2D shape. Afterwards, the most two similar adjacent triangles along edges are chosen. The Similarity among the 2D shapes is computed using our early proposed descriptor. Thus, we obtain a partitioned sphere into triangles regions. For each region in the views sphere; we place the camera at its associated center of mass (Local PCA) looking at the coordinate origin to take the most representative 2D view of all 2D views in it. The experimental results illustrate the efficiency of our proposed approach.

Mots clefs

Three-dimensional models, 3D Indexing and retrieval, Optimal 2D Views Selection, 2D shape descriptor,

1 Introduction

In recent years, with the significant advances in 3D acquisition and modeling, three-dimensional objects have become an important multimedia data type with many application possibilities. For example, 3D models can present complex information, and content-based searching problems in large 3D object repositories arise in many practical fields. Example application domains include CAD/CAM, Virtual Reality, medicine, molecular biology, military applications, and entertainment. In this context, content-based retrieval of 3D models has become an important subject of research. Several researchers have investigated the possibility of performing effective retrieval of 3D models from large archives, using shape properties instead of text. For efficient comparison and similarity estimation, 3D models can be represented with a set of meaningful descriptors that encode the salient geometric and topological characteristics of their shapes. The database objects are then ranked according to their distance to the descriptors of the query model. These descriptors can be global [1, 2], local [3, 4], structural [5, 6], transform-based approaches [7, 8] or by 2D/3D approach [9, 10]. This latter, consists to represent and describe a given 3D model by some 2D views. According to our investigation of the most well-known 2D/3D indexing methods proposed in the literature [9, 10, 11, 12, 13], we remark that they can be divided into two categories: in the first one, there is no automatic 2D views selection, in fact each 3D model is presented by a fixed number of 2D views [9, 10]. However, it is a major drawback that makes those methods inefficient. In the second one, the 3D model is presented by an "optimal" number of 2D views which are selected based on different used criteria's and different optimization algorithms [11, 12, 13]. Those methods will be detailed and criticized in the related works section.

The remainder of this paper is organized as follows. In Section II, the related works are presented and discussed. Our new proposed method is detailed and its robustness is discussed in Section III. Section IV presents the interpretation of the obtained results using our proposed

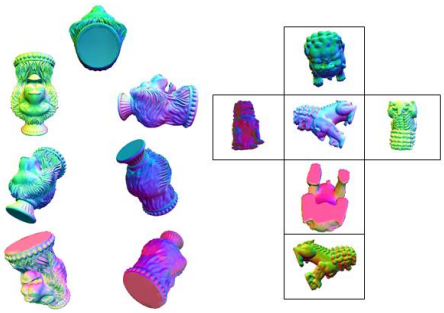
method compared to some well-known view-based methods. Finally, conclusion is presented in Section V.

2 Related Works

Within the pattern-recognition and Computer Vision communities, the problem of defining representative 2D views for recognition and representation of 3D objects has recently received significant attention [14, 15].
The 3D shape can indirectly be represented by various 2D shape descriptors associated with projection images. There exist two categories: with fixed associated 2D views and with optimal 2D views selection.

2.1 Fixed associated 2D views

This kind of category consists on attributing a fixed views' number. Mahmoudi et al. [10] introduced a new method based on 2D/3D silhouettes computation. Each 3D object is represented by a set of *seven views*: the *first three* directions are determined by the PCA applied on the 3D object and the other *four* are deducted from the principal views. To index the *seven* silhouettes (Figure 1- (a)) describing the 3D object, they utilized CSS (Curvature Scale Space) organized around an M-Tree index structure. This descriptor characterizes the contour by exploiting the maxima of curvature, identified through a multi-scale analysis.
The original work of Vranic [16] proposed the first descriptor based on depth images, Depth-Based image Descriptor (DBD) was introduced by Heczko et al. [17]. To ensure the geometric invariance behavior, each 3D object is first determined after a PCA and normalized according to a cube with parallel axes to those intrinsic references of the 3D object. By projecting the model on the *six* faces of the cube (Figure 1- (b)), depth images in gray level are calculated and then transformed into Fourier space using the 2D-FFT. The signature of the 3D object is then determined by storing, for each image Fourier processed, the low-frequency coefficients.
Ohbuchi et al. [9] proposed a new descriptor called Multiple Orientation Depth Fourier Descriptor (MODFD). The authors used two stages of the pre-processing with PCA to obtain independent representation of both translation and scale. Invariance against rotation is ensured by calculating the depth images of the 3D model taken from *forty two* different viewpoints equally spaced on the unit sphere. These images cover all possible views of the model by discretizing the space (θ, φ) of the unit ball. Using the method of Zhang [18], each depth image in the Cartesian coordinates system (x, y) is transformed into an image of depth in the system of polar coordinates (ρ, θ) system using a polar map. This latter was then transformed into a Fourier image. They considered only low frequencies to represent the corresponding view. Finally, they got a vector for each view and a set of vectors for each 3D object. The similarity between a pair of MODFD is calculated using the distances between all possible combinations of two sets of vectors. The calculation of similarity is very expensive.
However, presenting a 3D object with a fixed number of 2D views can lead to some major limitations that depend of the 3D shape complexity. The first one when the 3D model is complex and contains more information it leads to the problem of *under views estimation*. In the opposite case, if the 3D model has not a complex structure it can lead to another problem of *over views selection*.



(a) (b)

Figure 1 - *Fixed associated 2D views approach: (a) seven silhouettes, (b) six depth images.*

2.2 With optimal 2D views selection

To overcome those limitations of *under and over views estimation* leaded by the fixed 2D views number in the first category; the second one consist on automatic selection of the optimal 2D views.
Mokhtarian et al. addressed in [12] the issue of automatic selection of the best and the optimum number of 2D views for each 3D object. The object boundary of each 2D view is considered as a 2D shape and is represented effectively by less than ten pairs of integer values. These values include the locations of the maxima of its Curvature Scale Space (CSS) image contours. The CSS shape descriptor is expected to be selected for MPEG-7 standardization. They eliminated similar 2D views and selected a relatively small number of 2D views using an optimization algorithm. An unknown object is then recognized by a single image taken from an arbitrary viewpoint. Each object has been modeled using an optimized number of silhouette contours obtained from different viewpoints. This number varies from 5 to 25 depending on the complexity of the object and the measure of expected accuracy. However, it does not always give results aligned with human intuition. The main drawback of this method is its potentially high degree of ambiguity. The positions of zero-crossing point maxima for very deep and sharp concavities and for very long shallow concavities are almost identical. The convex parts of the curve are represented only implicitly by assuming that every concavity must be surrounded by two convexities.

Filali Ansary et al. [13] proposed a method for 3D model indexing based on 2D views, named AVC (Adaptive Views Clustering). The goal of this method is to provide an "optimal" selection of 2D views from a 3D model, and a probabilistic Bayesian method for 3D model retrieval from these 2D views. The characteristic views selection algorithm is based on an adaptive clustering algorithm (K-means) and used statistical model distribution scores to select the optimal number of 2D views. Starting from the fact that all views do not have equal importance, they introduced also a Bayesian approach to improve the retrieval. However, K-means and its variations present some limitations, when clusters have non-globular shapes or widely different sizes or densities. In addition, K-means is not efficient in the case of empty clusters. These factors decrease the accuracy of this 2D views selection method.
While most 3D object representations are complicated and inefficient, conventional multi-2D views representations are based on a large number of 2D views and cannot be used in many applications such as retrieval from large 3D objects databases. Multi-views representations have not yet successfully dealt with the following issues:

- What is the optimal number of 2D views?
- Do the extracted 2D views contain relevant information about the 3D object?

In our work, we consider that the problem of automatic 2D views selection can be divided into two underlying problems: the first one is the suitable used criteria and the second one is the use of an efficient optimization algorithm. Therefore, to enhance the performance of a given optimal 2D views selection; one can enhance both the performance of the used criteria and optimization algorithm or one of them.
For this aim, in this paper, we propose a new method based on two contributions. In the first one, we propose to use our early robust developed criteria [19] and in the second one, we suggest to use our new designed 2D optimal views selection algorithm.

3 The proposed method

The proposed method includes two contributions: the use of our robust developed criteria [19] and our new proposed 2D optimal views selection algorithm.

3.1 3D Pose Normalization

Note that, before applying our proposed 2D optimal views selection algorithm, a 3D PCA normalization must be performed in order to ensure invariance to the different geometric transforms. Indeed, 3D models have arbitrary position, orientation and scaling in 3D space. Since the extracted features are not invariant to position, orientation and scaling; to capture their invariant features, a feasible scheme is to place the model in a canonical coordinates frame to get the pose normalized. Then, a model is scaled, translated or rotated, the placing into the canonical frame. The pose normalization step is done through PCA [20, 21].

Let a 3D object defined by a triangular mesh M represented with a set T of n triangles $T = \{T_i, 1 \le i \le n\}$ and a set P of N vertices $P = \{P_i, 1 \le i \le N\}$. The covariance matrix C of the mesh M is approximated as follow:

$$C = \frac{1}{n}\sum_{i=1}^{n} S_i (g_i - m)(g_i - m)^T$$

Where S_i and g_i are the area of the i^{th} triangle of a shape and its center of gravity, m is the center of mass of the 3D model given by the formula:

$$m = \frac{\sum_{i=1}^{n} S_i g_i}{\sum_{i=1}^{n} S_i}$$

And n is the number of triangles of the 3D object. The process of scaling to a unit sphere (Figure 2 - (a)) is applied before the 3D alignment using the following formulas:

$$D = \max_{i=1,\dots,N} d(m, P_i)$$

$$P' = \left\{ P_i' \,|\, P_i' = \frac{1}{D} P_i, P_i \in P, i = 1, \dots, N \right\}$$

Where N is the number of vertices of the 3D object and P_i its i^{th} vertex.
The 3D alignment (Figure 2 - (c)) step must be performed after centering and scaling the 3D model in the centered unit sphere. Obviously the matrix C is a real symmetric one, therefore its eigenvalues are non-negative real numbers. Then we sort the eigenvalues in non-increasing order and find the corresponding eigenvectors. The eigenvectors are scaled to Euclidean unit length and we form the rotation matrix R which has the scaled eigenvectors as rows. We rotate all points in P'' and a new point set is formed:

$$P'' = \{ P_i'' \,|\, P_i'' = P_i' R, P_i' \in P', i = 1, \dots, N \}$$

(a) (b) (c)

Figure 2 - *(a) Centering and Scaling to the Unit Sphere, (b) and (c) are respectively 3D objects before and after PCA Alignment*

3.2 Our Robust 2D Shape Descriptor

Our robust 2D shape descriptor is based on multi-scale analysis. Let $f(u) = \langle (x(u), y(u)) | u \in [0, T] \rangle$ be the

parametric representation for a given curve of shape, where T is it arc-length, and u is the curvilinear abscise. And let $\langle g(u,\sigma)|\sigma > 0\rangle$ be set of Gaussians, where for a given σ, $g(u,\sigma)$ is given as follow : $g(u,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-u^2}{2\sigma^2}}$.

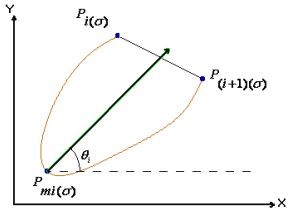


Figure 3 - *Token representation and orientation* θ_i.

The set of the smoothed curves $\langle f(u,\sigma)|\sigma \geq 0\rangle$, which are obtained by the convolution of $f(u)$ with the set of Gaussians $g(u,\sigma)$. For different value of σ is a multiscale representation of the curve of shape $f(u)$. A set of a multiscale curvature $K(u,\sigma)$ that corresponds to the set of curve of shape $\langle f(u,\sigma)|\sigma \geq 0\rangle$ is defined in [23] as follow:

$$K(u,\sigma) = \frac{x_t(u,\sigma)y_{tt}(u,\sigma) - x_{tt}(u,\sigma)y_t(u,\sigma)}{\left(x_t^2(u,\sigma) + y_t^2(u,\sigma)\right)^{3/2}}$$

Where x_t, y_t and x_{tt}, y_{tt} are respectively the first and second derivatives of x and y with respect to t.

Let $P = \{P_i(\sigma)\}_{i=1}^{N}$ be the set of minima that is the set of points such as $K(u,\sigma) = 0$. If we assume that the curvature $K(u,\sigma)$ is continuous, between two consecutive minima $P_i(\sigma)$ and $P_{i+1}(\sigma)$, there is always a maximum of $f(u)$, namely $m_i(\sigma)$, located at the point $P_{m_i}(\sigma)$. For each value of σ, a smoothed curve of shape is obtained, which is decomposed into portions or tokens according to the points P_i. Each token i of the curve $f(u,\sigma)$ is described by the vector $E_{(i\sigma)}(m_i(\sigma), O_i(\sigma))$ (Figure 3), with $m_i(\sigma)$ in [-180, 180] is the curvature at point $P_{mi}(\sigma)$, and $O_i(\sigma)$ in [0, 360] is the orientation defined in polar coordinates of the vector linking the median point of the segment $P_i(\sigma) - P_{i+1}(\sigma)$ with the point $P_{mi}(\sigma)$. Our descriptor is invariant to translation and scale; to assure its rotation invariance, we proposed the use of the principle of force equilibrium. Let $\Gamma_0 = \{\vec{f_i} | 1 \leq i \leq N\}$ be a set of features vectors of a given 2D shape where N is the number of vectors. If $\sum_{i=1}^{N} \vec{f_i} \neq \vec{0}$, then, there exists a vector $\overrightarrow{F_0}$ verifying $\sum_{i=1}^{N} \vec{f_i} + \overrightarrow{F_0} = \vec{0}$. The equilibrium vector is called the *principal vector* and its direction according the axis OX is called the *principal direction*. The principle consists on computing the principal direction θ and rotating all tokens' orientation vectors by $-\theta$ to let the principal vector of the features vectors of each 2D shape coincides with the OX axis.

3.3 The new proposed optimal 2D views selection algorithm

After performing the 3D normalization using PCA of the 3D targeted object, we construct a unit sphere with a regular mesh using two iterations of the Loop subdivision scheme on an initial tetrahedron to obtain 64 faces. We place the 3D object in it. To generate the initial 2D views, we place the camera on each of the triangle center looking at the coordinate origin. For each 2D view associated with a triangle in the view sphere we apply the binarization to the 2D image and we extract the edge of the associated 2D shape. Each triangle in the views sphere has an associated 2D view and has three adjacent triangles along its edges, thus, the similarity measure is computed between its 2D view and each of the three 2D views associated to its three adjacent triangles, the most two similar adjacent triangles along its edges are chosen. The Similarity among the 2D shapes associated to the 2D views is computed according to the chosen criteria (2D shape descriptor based Contours or Regions). In our case, it is based on our early developed robust shape descriptor [19]. Thus, we obtain a partitioned sphere into triangles regions. For each region in the view sphere; we place the camera at its associated center of mass (Local PCA) looking at the coordinate origin to take the most representative view of all views in it.

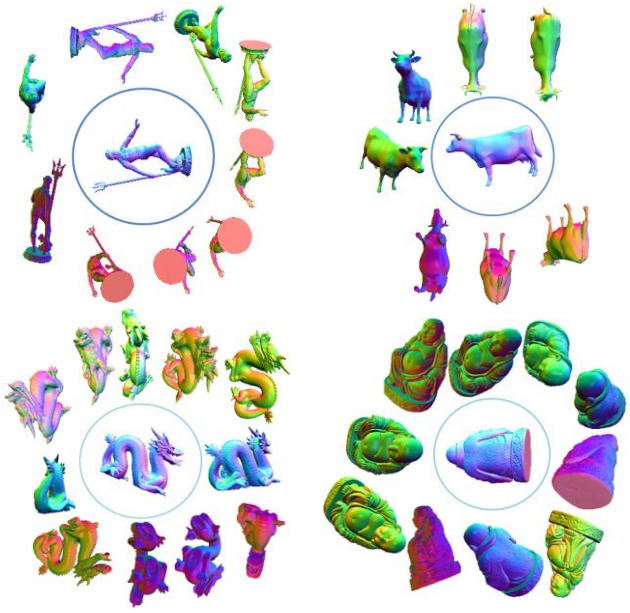


Figure 4 - *Some Results of the obtained optimal 2D views using our proposed method.*

4 3D/3D Matching

Let now A and B be two 3D models, with features vectors F_t^A and F_t^B respectively, and $F_t^A = \bigcup_{i=1}^{N_a} f_i^A$ and $F_t^B = \bigcup_{i=1}^{N_b} f_i^B$ where N_a and N_b are the numbers of the 2D

shapes associated to A and B respectively and f_i^A and f_i^B are the i^{th} shapes' descriptor of A and B respectively.
For an efficient matching procedure among the set of shapes of A and B; in the first step, we compute the distance between an i^{th} shape associated with A and every shape associated with the model B. The smallest of the computed distances is the distance $d(A_i, B)$ given by the following formula:

$$d(A_i, B) = \min_{1 \le j \le N_b} D(f_i^A, f_j^B)$$

$D(f_i^A, f_j^B)$ is the Minkowski distance computed between an i^{th} shape associated with A and a j^{th} shape associated with B according to the our early developed robust 2D shape descriptor [19].
Thus, the distance between the model A and B is given by the following formula:

$$d(A, B) = \frac{1}{N_a} \sum_{i=1}^{N_a} d(A_i, B)$$

5 Experimental Results

The proposed method was experimentally evaluated using the Princeton Benchmark Database [22]. It is one of standard databases of 3D models available on the Web by the team "Princeton Shape Retrieval and Analysis Group" to let researchers evaluating their 3D indexing algorithms on the same 3D database. The Princeton database contains 1814 3D models grouped in high-level semantics classes where the objects of the same class are heterogeneous. For example, the insects' class contains 3D models which represent insects of different shapes but with the same semantic.
To evaluate the proposed method, each 3D model was used as a query object. The retrieval performance was evaluated in terms of "precision" and "recall", where precision is the proportion of the retrieved models that are relevant to the query and recall is the proportion of relevant models in the entire database that are retrieved in the query.

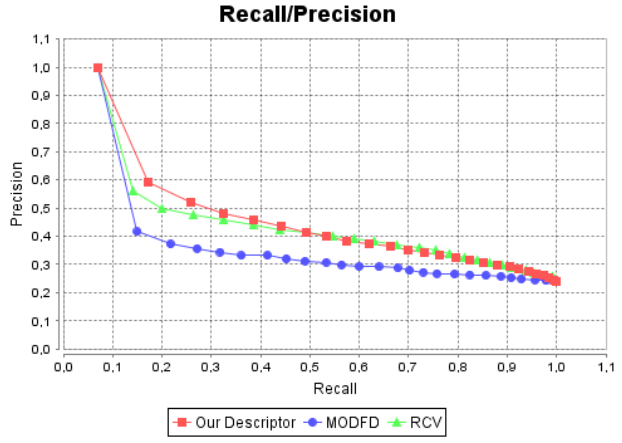


Figure 5 - *Recall/Precision curve evaluating the performance or our new proposed method.*

In order to evaluate our proposed approach, we compared it to the following two methods, which are based on a fixed 2D views number:

- Retrieving 3D shapes based on their appearance [9] that proposed Multiple Orientation Depth Fourier Descriptor (MODFD).
- Retrieval by shape using characteristic views (RCV) [10].

The retrieval experimental results (Figure 5) illustrate the efficiency of our proposed method over similar view-based methods. They showed the importance of local information and the efficiency of our early developed robust 2D shape descriptor.
From the figure 5, when the recall is between 0% and 50%; the most of the retrieved 3D models belong to the same class of the 3D query object in the case of our method compared with the descriptors MODFD and CSS. Which shows that the 2D extracted views using our proposed optimal views selection method, contains more relevant information about the 3D targeted object.
The robustness of our proposed method is from its invariance to the different geometric transforms (translation, scale and rotation). The limitations of our proposed method result in:

- The case of the 3D articulated objects, due to the tokens' orientation change.
- The high computational cost during the 2D views selection procedure, especially where there is a huge amount of local features to be extracted and the multi-scale analysis of each 2D shape worsens it more.
- Our proposed 2D views selection method doesn't take into account the human perception factor.

6 Conclusion and Future works

In this paper, we proposed a new method for efficient 3D model retrieval which contains two contributions: the first one is the use of our robust developed criteria and the second one is our new proposed 2D optimal views selection. Our new algorithm of optimal view selection showed it efficiency of extracting 2D views with more relevant information of 3D objects.
The retrieval experimental results showed efficiency and superiority of our method compared to other well-known methods based on fixed 2D views number. The strength of the proposed method is its robustness in terms of information relevance contained in the 2D extracted optimal views for a given 3D object.
In our future work, we are planning to improve our optimal views selection algorithm in order to take into account human perception factors. We are working on designing a well-suited index structure to improve the

speed response of our 3D developed Search Engine. In addition to using a 3D model (given a priori) as the query, we would like to add 2D sketch, in order to apply our method on various interactive applications, 3D face recognition, occlusion problem, classification of marine species.

Acknowledgements

The authors would like to acknowledge the Stanford Data Archives, AIM@SHAPE Shape Repository and the Princeton Shape Benchmark database [22] for the 3D models.
The authors are also grateful for the constructive and valuable comments from the reviewers of this paper.

References

[1] CHEN, D.-Y., TIAN, X.-P., SHEN, Y.-T., AND OUHYOUNG. On visual similarity based 3D model retrieval. *Computer Graphics Forum* 22, 3, 223–232, M.2003.

[2] KAZHDAN, M., FUNKHOUSER, T., AND RUSINKIEWICZ. Rotation invariant spherical harmonic representation of 3D shape descriptors. In *SGP '03: Proceedings of the 2003 Eurographics/ ACM SIGGRAPH symposium on Geometry processing*, 156–164, S. 2003.

[3] C. Chua and R. Jarvis. Point signatures: A new representation for 3D object recognition. *International Journal of Computer Vision*, 25(1):63–85, 1996.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, 2002.

[5] HILAGA, M., SHINAGAWA, Y., KOHMURA, T., AND KUNII. Topology matching for fully automatic similarity estimation of 3D shapes. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, ACM Press, 203–212, T. L. 2001.

[6] BIASOTTI, S., MARINI, S., SPAGNUOLO, M., AND FALCIDIENO. Sub-part correspondence by structural descriptors of 3D shapes. *Computer-Aided Design* 38, 9, 1002–1019, B. 2006.

[7] Novotni, M. Klein, R. Shape retrieval using 3D Zernike descriptors. *Computer Aided Design*, 36(11):1047–1062. 2004.

[8] Vranic, D. V. Saupe, D. 3D Shape Descriptor Based on 3D Fourier Transform. *EURASIP Conference on Digital Signal Processing for Multimedia Communications and Services (ECMCS '01)*, pages 271–274. Budapest, Hongrie. 2001.

[9] R. Ohbuchi, M. Nakazawa, and T. Takei. Retrieving 3D shapes based on their appearance. In *5th ACM SIGMM Workshop on Multimedia Information Retrieval (MIR 2003)*, Berkeley, California, USA, November 2003.

[10] S. Mahmoudi and M. Daoudi. Retrieval by shape using characteristic views. *The International Conference on Pattern Recognition ICPR02*, pages 457-460, August 11-15, 2002.

[11] J. Lee, B. Moghaddam, H. Pster, and R. Machiraju. Finding optimal views for 3D face shape modeling. In *International Conf. on Automatic Face and Gesture Recognition*, pages 31–36, 2004.

[12] F. Mokhtarian and S. Abbasi. Automatic selection of optimal views in multi-view object recognition. In *The British Machine Vision Conf. (BMVC'00)*, pages 272–281, 2000.

[13] T.Filali Ansary, M. Daoudi, J-P Vandeborre, A Bayesian 3D Search Engine using Adaptive Views Clustering. *IEEE Transactions On Multimedia*,Vol. (9), Issue (1) , Page 78-88, January 2007.

[14] T. Denton, M. Demirci, J. Abrahamson, A. Shokoufandeh, and S. Dickinson. Selecting canonical views for view-based 3-D object recognition. In *17th International Conf. on Pattern Recognition (ICPR'04)*, volume 2, pages 273–276, August 2004.

[15] P. M. Hall and M. J. Owen. Simple canonical views. In *The British Machine Vision Conf. (BMVC'05)*, volume 1, pages 7–16, 2005.

[16] D. Vranic, *3D Model Retrieval*, Ph.D. dissertation, University of Leipzig, 2004.

[17] M. Heczko, D. A. Keim, D. Saupe, and D. V. Vranic. Verfahren zur Ähnlichkeitssuche auf 3d objekten (methods for similarity search on 3D databases). *Datenbank-Spektrum*, 2(2):54-63, 2002.

[18] D. S. Zhang and G. Lu. Shape-based image retrieval using generic Fourier descriptor. *Signal Processing: Image Communication*, 17(10):825-848, November 2002.

[19] H. Silkan, A. Lachkar, S. E. Ouatik, and A. Elkharraz. Rotation Invariant and Robust Shape Descriptor for Content-Based Shape Retrieval. *Information and Communication Technologies International Symposium ICTIS'07 IEEE*. Morocco, Fez, 3-5 April 2007.

[20] Chen, D.Y. and Ouhyoung, M.: A 3D model alignment and retrieval system. *Proc. of International Computer Symposium*, 2002.

[21] Jolliffe, I.T. *Principal Component Analysis*. Springer, 1986.

[22] Shilane, P., Min, P., Kazhdan, M. et Funkhouser, T. The Princeton Shape Benchmark. *Shape Modeling International (SMI '05)*, pages 167–178. Genève, Italie. pages xii, 56, 57, 113. 2004.

[23] Farzin Mokhtarian, Alan K. Mackworth. "A Theory of Multiscale, Curvature-Based Shape Representation for Planar Curves". IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume 14, Issue 8. Pages: 789 – 805. August 1992.

Minimisation hiérarchique pour le suivi des mouvements de la main

O. Ben Henia[1] M. Hariti[1] S. Bouakaz[1]

[1] LIRIS (Laboratoire d'InfoRmatique en Image et Systèmes d'information)

Université Claude Bernard, Lyon 1
43, Boulevard du 11 novembre 1918
69622 Villeurbanne Cedex

{obenheni, mhariti, sbouakaz}@liris.cnrs.fr

Résumé

Les approches de suivi des mouvements de la main à base de modèle 3D peuvent être classifiées en deux catégories. La première catégorie utilise des filtres stochastiques comme le filtre de Kalman ou le filtre particulaire. La deuxième se base sur des méthodes déterministes définissant une fonction de dissimilarité qui compare les gestes de la main avec ceux du modèle 3D. La minimisation de cette fonction assure le suivi des mouvements de la main. Deux principaux problèmes surviennent lors de la minimisation. Le premier problème est celui des minimas locaux et le deuxième est celui du temps de calcul nécessaire pour se rapprocher suffisamment de la solution recherchée. Pour faire face à ces deux problèmes nous proposons une nouvelle fonction de dissimilarité qui est plus robuste face aux minimas locaux que d'autres fonctions très connues comme la fonction de Chanfrein[1]. Nous proposons aussi un algorithme de minimisation hiérarchique qui simplifie et améliore le suivi des mouvements de la main en diminuant les temps de calcul et en améliorant la robustesse face aux minimas locaux.

Mots clefs

Suivi des mouvements de la main, minimisation, modèle 3D.

1 Introduction

Le suivi des gestes de la main est un domaine en pleine expansion. Cela est dû aux nombreuses applications qui en découlent comme par exemple la création d'une interface Homme-Machine (IHM) où selon le geste de la main une action spécifique est réalisée. Les gants de données, appelés aussi gants instrumentés, sont couramment utilisés comme périphérique d'entrée pour saisir et suivre le mouvement de la main grâce à des capteurs. Malgré leur efficacité à capturer les mouvements de la main, les gants de données sont très couteûx, très fragiles et leurs câbles de liaison constituent une entrave les rendant encombrants.

De nombreux travaux de recherche s'intéressent à d'autres alternatives et notamment à la vision artificielle pour la capture des mouvements de la main [2][3][4]. En effet, les caméras vidéo sont plus accessibles en termes de coût et de simplicité d'utilisation. Cependant, le suivi des mouvements de la main à base de caméras reste encore complexe à cause des problèmes d'occultation et du nombre élevé des degrés de liberté de la main.

Dans ce papier nous proposons une méthode orientée modèle 3D paramétrique pour suivre des mouvements de la main dans une séquence vidéo. Nous définissons une nouvelle fonction de dissimilarité qui compare les gestes de la main avec ceux du modèle 3D. Cette fonction est ensuite minimisée pour chaque image de la séquence vidéo pour obtenir les paramètres du modèle permettant de reproduire les mouvements de la main observés dans la séquence d'images. En raison du nombre élevé des degrés de liberté de la main (aux alentours de 26), beaucoup de paramètres sont à estimer pendant la phase de minimisation. Cela rend le processus de minimisation sensible aux minimas locaux et plus couteux en temps de calcul. C'est pourquoi nous proposons une minimisation hiérarchique de la fonction de dissimilarité. Ceci nous a permis de simplifier et d'améliorer le suivi des mouvements de la main en diminuant les temps de calculs et en améliorant la robustesse face aux minimas locaux.

Dans la section suivante nous présentons un bref état de l'art des approches de suivi des mouvements de la main. La section 3 décrit le modèle 3D ainsi que la fonction de dissimilarité qui compare la projection du modèle avec l'image de la main. La section 4 détaille l'algorithme de suivi. Avant de conclure nous présentons dans la section 5 les résultats expérimentaux obtenus à partir de séquences d'images synthétiques et réelles.

2 Etat de l'art

Les approches de suivi des mouvements de la main dans une séquence vidéo peuvent être décomposées en deux classes. La première classe utilise une base de gestes à partir de laquelle on cherche le geste correspondant à celui observé dans une image de la séquence vidéo. Ces approches utilisent en général des techniques de classification ou de

régression [5][6].

En raison de la grandeur de l'espace des gestes que peut prendre une main, il est difficile voire impossible d'obtenir une base contenant tous les gestes possibles d'une main. Ainsi, ces approches sont bien adaptées pour la reconnaissance d'un nombre fini de poses prédéfinies pour des applications temps réel. Dans ce cas, le temps de calcul est privilégié sur la précision du suivi. C'est le cas des interfaces homme-machine. Dans cette optique, un système de reconnaissance de gestes utilisant un classifieur est proposé dans [7].

La deuxième classe regroupe des approches de suivi utilisant un modèle 3D paramétrique. Le problème du suivi est alors formalisé sous forme d'un problème d'estimation des paramètres du modèle 3D permettant de reproduire les gestes de la main observés dans une séquence vidéo. Les paramètres du modèle 3D peuvent être estimés en utilisant des méthodes stochastiques ou déterministes. Le premier type de méthodes utilise des filtres stochastiques comme le filtre de Kalman étendu utilisé dans [8] ou le filtre particulaire [9] [10]. Ce dernier donne de meilleurs résultats que le filtre de Kalman mais présente l'inconvénient d'être couteûx en temps de calcul. Outre les méthodes stochastiques, des méthodes déterministes ont aussi été utilisées pour réaliser le suivi des mouvements de la main. Dans ce cas, le problème de suivi est formalisé sous forme d'un problème de minimisation. En effet, une fonction de dissimilarité comparant les gestes de la main avec ceux du modèle 3D est définie. Cette dernière est minimisée afin d'estimer les paramètres du modèle 3D reproduisant les gestes de la main observés dans une séquence vidéo. Dans cette catégorie de méthodes, différentes fonctions de dissimilarité ont été proposées. Certaines se basent sur l'information de silhouette [11] tandis que d'autres définissent une distance au contour [1]. Delamarre et Faugeras [12] ont proposé un approche basée sur la stéréovision. Bray et al [13] ont défini une fonction qui utilise l'information contenue dans une carte de profondeur de la main. Cette carte de profondeur est obtenue grâce à des capteurs spécifiques.

Dans notre travail, nous définissons une nouvelle fonction de dissimilarité qui donne de meilleurs résultats que d'autres fonctions très connues comme la celle de Chanfrein [1] ou celle de la surface de non recouvrement utilisée dans [11]. Nous proposons par la suite de minimiser cette fonction en deux étapase en utilisant l'algorithme de Torczon[14]. La première étape donne les paramètres du modèle relatifs à la position et l'orientation de la paume de la main. La deuxième estime les angles d'articulations des doigts. Cette simplification nous a permis d'améliorer les temps de calculs et la robustesse face aux minimas locaux.

3 Modèle 3D et Fonction de Dissimilarité

3.1 Modèle 3D de la main

Le modèle 3D utilisé est un modèle paramétrique respectant la norme H-Anim. Ce modèle possède une partie cinématique et une partie apparence. Pour cette dernière nous utilisons des quadriques telles que des sphères, ellipsoïdes et cônes pour donner une forme au modèle 3D proche de celle de la main (Figure.1(b)).

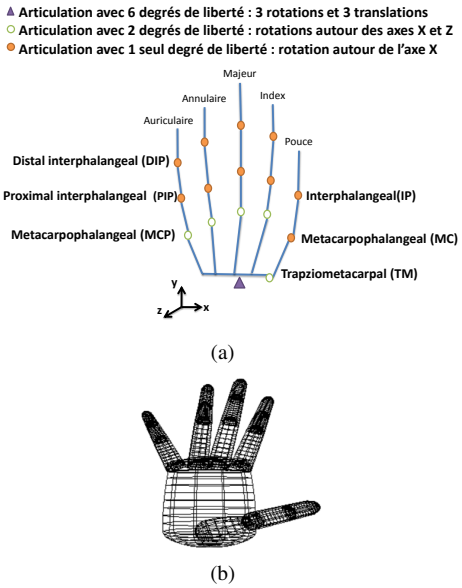


Figure 1 – *(a) Représentation squelettique du modèle 3D montrant ces différentes articulations (b) Apparence du modèle 3D à base quadriques*

La partie cinématique est constituée d'une hiérarchie de transformations 3D (rotations, translations) permettant d'animer le modèle 3D. On peut énumérer 26 paramètres correspondants aux degrés de liberté de la main. Les six premiers paramètres du modèle modélisent des mouvements globaux de la main : rotations et translations de la paume de la main. Les 20 paramètres restants modélisent des mouvements locaux ou plus fins de la main : les articulations des doigts. En effet, chaque doigt peut être modélisé par 4 degrés de libertés : deux pour les articulations MCP et leur abduction et deux correspondants aux articulations PIP et DIP. Nous exploitons la dépendance entre les angles DIP et PIP pour réduire la partie cinématique de notre modèle à 22 degrés de liberté. La formule utilisée est comme suit : $\theta_{DIP} = 2/3\theta_{PIP}$.

En utilisant ce modèle 3D nous allons générer des projections qui seront comparées avec les images de la main. Cette comparaison est réalisée grâce la fonction de dissimilarité présentée dans la sous-section suivante.

3.2 Fonction de dissimilarité

Parmi les fonctions les plus connues comparant les images de la main avec les projections du modèle 3D on peut citer les fonctions qui estiment une distance entre deux contours : celui extrait de l'image de la main (Figure.2(e)) et celui de la projection du modèle 3D. C'est le cas de la fonction de Chanfrein. Celle-ci estime une distance entre deux contours en utilisant la distance de Chanfrein. En effet, à partir de deux ensembles de pixels A et B représentant les contours extraits de deux images I_a et I_b une valeur de dissimilarité d_C est calculée. La fonction de Chanfrein d_C estimant une distance entre deux contours A et B peut être exprimée selon la formule suivante :

$$d_C(A,B) = \frac{1}{|A|}\sum_{a_i \in A} \min_{b_j \in B} d(a_i, b_j) \qquad (1)$$

où d est une approximation de la distance euclidienne calculée grâce à l'algorithme de Chanfrein [1].

La fonction de Hausdorff est elle aussi très connue et peut être considérée comme une variante de la fonction de Chanfrein. En reprenant les mêmes notations utilisées pour définir la fonction de Chanfrein, la fonction de Hausdorff peut être formulée de cette manière :

$$d_H(A,B) = \max_{a_i \in A}\{\min_{b_j \in B} d(a_i, b_j)\} \qquad (2)$$

Ces fonctions qui se basent sur les contours sont très sensibles au bruit présent dans les images. Une autre alternative est proposée par Ouhadi et Horrain[11] en calculant la surface de non recouvrement(Figure.2(d)) correspondant à la partie non commune aux deux surfaces : la silhouette de la main H_s (Figure.2(a)) et la projection du modèle M_p (Figure.2(b)). Notons SNR une image de dimension $N_l \times N_c$ contenant la surface de non recouvrement (Figure.2(d)), où N_l et N_c représentent le nombre de lignes et de colonnes respectivement. Un pixel (i,j) de l'image SNR est défini par :

$$SNR_{ij} = \begin{cases} 1 & \text{si le pixel (i,j) appartient à la surface} \\ & ((H_s \cup M_s) - (H_s \cap M_s)) \\ 0 & \text{sinon} \end{cases}$$

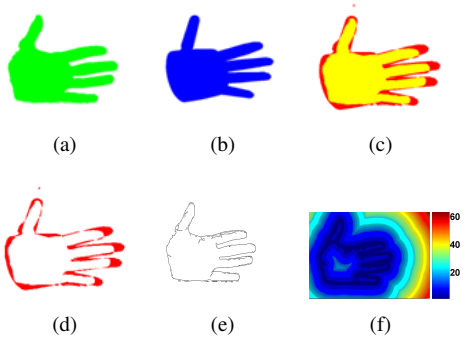


Figure 2 – *Différentes images utilisées pour le calcul de notre fonction de dissimilarité :(a) Silhouette de la main H (b) Projection du modèle M_p(c) Superposition de la silhouette de la main et de la projection du modèle (d) Surface de non recouvrement SNR (e) Contour de la main (f) Carte de distance D*

Pour rendre plus robsuste la fonction de non recouvrement, nous proposons d'ajouter une pondération à chaque pixel de la surface de non recouvrement SNR. Cette pondération est calculée à partir de la carte de distance D (Figure.2(f)) obtenue en appliquant l'algorithme de Chanfrein à l'image contenant le contour de la main(Figure.2(e)). Un élément D_{ij} de la carte D contient la distance d'un pixel (i,j) au contour de la main(Figure.2(e)). Ainsi, notre fonction de dissimilarité compare l'image de la main avec la projection du modèle associée aux paramètres R,T et θ, où R et T représentent le mouvement global de la paume de la main (3 rotations et 3 translations) et θ représente le mouvement local (angles d'articulation des doigts). Notre fonction de dissimilarité d_F est formalisée comme suit :

$$d_F(SNR, D) = \sum_{i=1,j=1}^{N_r,N_c} SNR_{ij} * D_{ij} \qquad (3)$$

4 Algorithme de Suivi

Le suivi des gestes de la main dans une séquence vidéo est réalisé en estimant les paramètres du modèle 3D permettant de reproduire le mouvement de la main observé dans la séquence vidéo. Cette estimation est réalisée en minimisant la fonction de dissimilarité pour chaque image de la séquence vidéo. Ceci permet de recaler la projection du modèle 3D sur la surface de la main pour chaque image et ainsi reproduire le mouvement observé dans la séquence vidéo.
Pour la première image de la séquence vidéo, nous supposons que les paramètres du modèle 3D sont proches de la solution recherchée. Pour le reste de la séquence vidéo, l'algorithme de minimisation est initialisé à partir des paramètres du modèle 3D de la main obtenus à l'image précédente. L'algorithme de minimisation utilisé est celui de Torczon[14] qui présente une amélioration de l'algorithme

du simplexe proposé par Nelder et Mead[15]. En effet, la méthode de Torczon[14] ne présente pas des problèmes de dégénérescence comme c'est le cas pour la méthode de descente du simplexe proposée par Nelder et Mead[15]. L'algorithme de Torczon[14] est un processus itératif explorant à chaque itération différentes directions pour en choisir celle qui minimise au mieux la fonction de dissimilarité. Une des particularités de cet algorithme est qu'il ne requiert pas la connaissance de la dérivée de la fonction à minimiser. Le deuxième argument justifiant notre choix de l'algorithme de Torczon[14] est lié au traitement de celui-ci qui explore différentes directions à chaque itération. Cette recherche multidirectionnelle peut être exécutée en parallèle afin d'améliorer les temps de calcul nécessaires pour atteindre la solution recherchée.
En raison de la grande dimensionnalité de l'espace de recherche, nous découpons l'algorithme de minimisation en deux étapes. La première étape estime les paramètres du modèle 3D relatifs au mouvement global de main : la translation et la rotation de la paume de la main. Ainsi, dans cette première étape, les paramètres qui représentent les angles des articulations des doigts sont fixes, tandis que ceux qui représentent la position et l'orientation de la main sont traités par l'algorithme de minimisation. Le processus est inversé lors de la deuxième étape, c'est-à-duire les paramètres d'orientation et de position sont tout d'abord fixés à ceux obtenus dans la première étape, et les angles des articulations des doigts sont ensuite estimés. Ce processus est résumé dans le schéma (Figure.3).

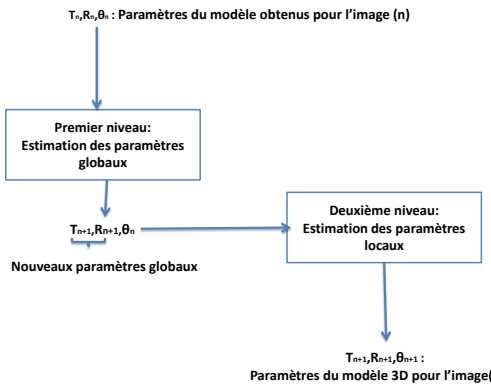


Figure 3 – *Processus d'estimation en deux étapes des paramètres du modèle 3D*

Outre la simplification du problème de minimisation, cette approche peut être justifiée par la variation lente du mouvement de la main entre deux images successives.

5 Résultats Expérimentaux

Les performances de notre algorithme de suivi des mouvements de la main sont évaluées sur des séquences d'images vidéo synthétiques et réelles. Nous évaluons notre fonction de dissimilarité en la comparant avec d'autres fonctions de comme celle de Chanfrein [1], celle de Hausdorff [16] ainsi que celle de non recouvrement [11].Pour cela, une séquence vidéo composée d'une centaine d'images de synthèse de dimension 320x240 est acquise à partir du mo-

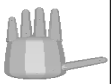
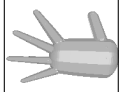
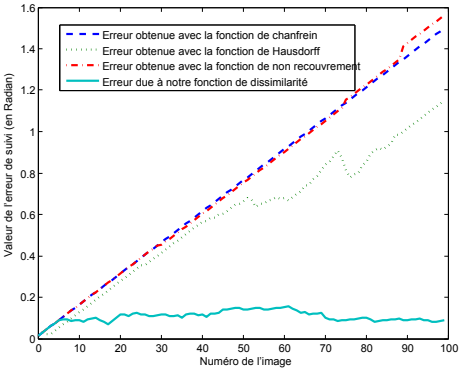
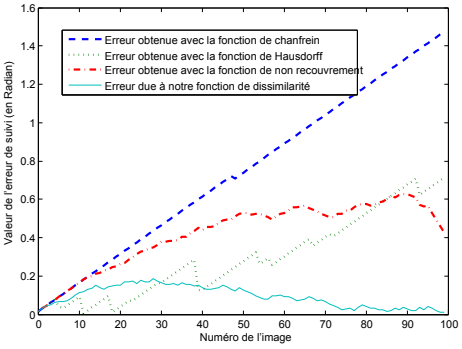
	Algorithme de suivi	Image 1	Image 50	Image 100
Poses à retrouver				
d_C	une étape			
	deux étapes			
d_H	une étape			
	deux étapes			
d_{SNR}	une étape			
	deux étapes			
d_F	une étape			
	deux étapes			

Tableau 1 – *Résultats du suivi obtenus par l'algorithme à une étape et celui à deux étapes en utilisant différentes fonctions de dissimilarité : fonction de Chanfrein(d_C), fonction de Hausdorff (d_H), fonction de non recouvrement(d_{SNR})et notre fonction proposée(d_F).*

dèle 3D de la main (Tableau 1). Pour obtenir cette séquence d'images, on fait varier trois (respectivement quatre) paramètres relatifs au mouvement global (respectivement local). Les paramètres du mouvement global sont la translation selon les axes X et Y ainsi que la rotation autour de l'axe des Z (Fig :1(a)). Les paramètres locaux sont les articulations MCP métacarpophalangienne (Fig :1(a)) et les abductions des doigts de la main excepté le pouce. Les résultats du suivi dans la vidéo de synthèse sont présentés dans le tableau 1.



(a)



(b)

Figure 4 – *Erreur de suivi de la rotation de la main autour de l'axe Z(a) Erreurr obtetnue en utilisant l'algorithme de minimisation à une étape(b) Erreur obtenue par l'algorithme de minimisation hiérarchique*

Dans le même tableau, nous pouvons voir que notre fonction de dissimilarité d_F fournit les meilleurs résultats comparés avec ceux des autres fonctions de dissimilarité : la fonction de Chanfrein d_C[1], la fonction de Hausdorff d_H [16] ou encore la fonction de non recouvrement (d_{SNR})[11].
Pour quantifier l'erreur de suivi associée à chaque fonction de dissimilarité, on calcule une différence entre les résultats du suivi obtenus et les valeurs exactes recherchées. L'erreur de suivi est alors tracée sous forme d'une courbe (Figure.4). Nous traçons seulement la courbe représentant l'erreur relative au suivi de la rotation autour de l'axe Z. Nous observons dans la figure4 que notre fonction de dissimilarité est plus efficace que les autres fonctions de dissimilarité. En effet, l'erreur de suivi est de 0,09 radian pour notre fonction de dissimilarité, alors que celle-ci peut être supérieure à 1 radian avec les autres fonctions, notamment pour la fonction Chanfrein. La même figure montre également que la minimisation hiérarchique est plus robuste qu'une minimisation en une étape. Plus précisément, dans le cas de la fonction de non recouvrement, la minimisation hiérarchique améliore considérablement les performances du suivi en diminuant l'erreur de suivi d'un rapport de 1/2 (figure4). Nos observations concernant l'erreur de suivi du mouvement de rotation autour de l'axe Z restent valables pour l'estimation des autres paramètres du modèle 3D. Outre la robustesse, la minimisation hiérarchique est plus rapide en temps de calcul qu'une minimisation en une étape. En effet, pour la séquence d'images de synthèse, la cadence de traitement est d'environ 8 images par seconde pour une minimisation en une étape alors que la minimisation hiérarchique a une cadence de 11 images par seconde. Ces résultats ont été obtenus en utilisant un processeur à 2.2GHZ. Les projections du modèle 3D sont calculées en utilisant la librairie graphique d'OpenGL.
Nous évaluons également notre algorithme de suivi sur une séquence d'images réelles (Figure.5).Dans la figure5 la ligne du haut montre des images de la vidéo traitée. La deuxième ligne montre les résultats du suivi obtenus par l'algorithme de minimisation à une étape. La dernière ligne montre les résultats du suivi obtenus par notre algorithme de minimisation hiérarchique. La figure 5 montre aussi la robustesse du suivi en utilisant une minimisation hiérarchique. En effet, nous pouvons souligner que la minimisation hiérarchique est plus efficace pour suivre les mouvements complexes des doigts de la main comme on peut le voir notamment dans l'image numéro 400, où le suivi des doigts se perd dans le cas de la minimisation en une étape.

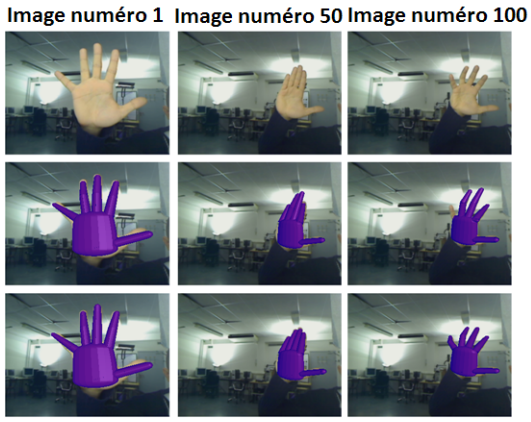


Figure 5 – *Résultats du suivi de la main dans une séquence d'images réelles*

6 Conclusion et Travaux Futurs

Dans cet article, nous avons proposé une méthode de suivi de suivi des mouvements de la main à partir d'une caméra,

sans marqueur et en utilisant un modèle 3D paramétrique de la main. Une nouvelle fonction de dissimilarité comparant des gestes de la main avec ceux du modèle 3D est proposée. Cette fonction est minimisée pour estimer les paramètres du modèle 3D reproduisant le mouvement de la main. Le grand nombre des degrés de liberté (environ 26) qui doivent être estimées rend la minimisation sensible aux minimas locaux et augmente le temps de calcul nécessaire pour atteindre la solution recherchée.
Une minimisation hiérarchique en deux étapes de la fonction de dissimilarité a permis de simplifier le problème de minimisation. La première étape de notre minimisation hiérarchique estime les degrés de libertés globaux de la main : position et orientation de la paume. La deuxième étape estime les degrés de libertés locaux de la main, c'est-à-dire les angles des articulations des doigts. D'après nos résultats expérimentaux, l'algorithme de minimisation hiérarchique est plsu robuste face aux minimas locaux qu'un algorithme classique de minimisation en une étape.
L'algorithme que l'on propose améliore également la rapidité du suivi des mouvements de la main. Dans de cadre de nos travaux de recherche, nous avons utilisé une seule caméra et il reste difficile de traiter, dans le cas mono-caméra, des mouvements complexes tels que des mouvements de torsion de la main. Nous sommes effectivement très vite confrontés au problème de l'auto-occultation. L'utilisation de plusieurs caméras ou encore d'autres technologies comme les caméras à temps de vol (exemple la swissranger [1]) peut être une solution pour suivre des mouvements plus complexes de la main. L'utilisation de plusieurs points de vue dans le cas multi-caméras ou celle de l'information 3D dans le cas des caméras à temps de vol peut résoudre certaines ambigüités. Des études en cours d'approfondissement tentent d'améliorer les temps de calcul en transférant une partie des traitements sur des unités de traitement graphiques appelées GPUs.

Références

[1] G. Borgefors. Hierarchical chamfer matching : A parametric edge matching algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6) :849–865, 1988.

[2] Ying Wu et Thomas S. Huang. Vision-based gesture recognition : A review. Dans *GW '99 : Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*, pages 103–115, London, UK, 1999. Springer-Verlag.

[3] Martin Tosas. *Visual Articulated hand tracking for Interactive Surfaces*. Thèse de doctorat, University of Nothingham, 2006.

[4] Bjorn Dietmar Rafael Stenger. *Model-Based Hand Tracking Using A Hierarchical Bayesian Filter*. Thèse de doctorat, University of Cambridge, 2004.

1. http ://www.mesa-imaging.ch/

[5] Rómer Rosales, Vassilis Athitsos, Leonid Sigal, et Stan Sclaroff. 3d hand pose reconstruction using specialized mappings. Dans *ICCV*, pages 378–385, 2001.

[6] Nobutaka Shimada, Kousuke Kimura, et Yoshiaki Shirai. Real-time 3-d hand posture estimation based on 2-d appearance retrieval using monocular camera. Dans *RATFG-RTS '01 : Proceedings of the IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems (RATFG-RTS'01)*, page 23, Washington, DC, USA, 2001. IEEE Computer Society.

[7] Tsukasa Ike, Nobuhisa Kishikawa, et Björn Stenger. A real-time hand gesture interface implemented on a multi-core processor. Dans *MVA*, pages 9–12, 2007.

[8] B. Stenger, P. R. S. Mendonca, et R. Cipolla. Model-based 3d tracking of an articulated hand. Dans *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 2, pages II–310–II–315 vol.2, 2001.

[9] Michael Isard et Andrew Blake. Condensation conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29 :5–28, 1998.

[10] Makoto Kato et Gang Xu. Occlusion-free hand motion tracking by multiple cameras and particle filtering with prediction. *IJCSNS International Journal of Computer Science and Network Security*, 6(10) :58–65, 2006.

[11] Hocine Ouhaddi et Patrick Horain. 3d hand gesture tracking by model registration. Dans *Proc.IWSNHC3DI99*, pages 70–73, 1999.

[12] Quentin Delamarre et Olivier Faugeras. Finding pose of hand in video images : a stereo-based approach. Dans *In IEEE Proc. of the third International Conference on Automatic Face and Gesture Recognition*, pages 585–590. IEEE Computer Society, 1998.

[13] Matthieu Bray, Esther Koller-Meier, Nicol N. Schraudolph, et Luc Van Gool. Stochastic meta-descent for tracking articulated structures. Dans *CVPRW '04 : Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 1*, page 7, Washington, DC, USA, 2004. IEEE Computer Society.

[14] J. E. Dennis, Jr., et Virginia Torczon. Direct search methods on parallel machines. *SIAM Journal on Optimization*, 1 :448–474, 1991.

[15] J. A. Nelder et R. Mead. A simplex method for function minimization. *The Computer Journal*, 7(4) :308–313, January 1965.

[16] Daniel P. Huttenlocher, Gregory A. Klanderman, Gregory A. Kl, et William J. Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15 :850–863, 1993.

Is it a face ? How to find and validate a face on 3D scans

Przemyslaw Szeptycki, Mohsen Ardabilian, Liming Chen

Ecole Centrale de Lyon, 36 av. Guy de Collongue, 69134 Lyon, France

{przemyslaw.szeptycki, mohsen.ardabilian, liming.chen}@ec-lyon.fr

Abstract

Rapid development of 3D scanning technologies lead researchers to use them for people recognition, cameras are faster, models are less noisy and with higher resolution. 3D facial models have been widely used for many biometrics applications. Nevertheless 3D face recognition topics mainly assume that scans contain a face and the face is mostly in a frontal position. In real world we might have a situation, where a scanned model is not sufficient for recognition.

In this article we propose a generic face model validation algorithm which can exclude non-face models from recognition query. The algorithm was tested on more than 1500 range scans including face and non-face models. Obtained results prove, that the generic model validation approach can be used to reject non-face models from the recognition pipeline.

Key words

3D face, Curvature, face validation.

1 Introduction

The use of 3D face models has emerged as a major face recognition solution in the last years to deal with unsolved issues, e.g. lighting and pose variations, for reliable 2D face recognition solutions [1]. Nevertheless for recognition purposes faces are generally detected manually and registered in standard position. The face detection/validation literature mostly concerns face detection on texture images of scanned scene. Such detection is dependent to the face rotation and the lighting conditions. Likewise challenges arise from the fact that the scanned persons are non-cooperative. To make 3D face recognition algorithms automatic and insensitive to the lighting and pose variations changes, a face has to be detected directly on a 3D model without reinforce from the texture. In many cases the face validation problem has been decomposed to a problem of face anchor points localization but with strong assumptions about the position and the orientation.

In this paper we present an algorithm for automatic face validation based on anchor points detection and distance between a generic face model and a query model, which let us to exclude non-frontal faces and non-face objects from a query. In order to ascertain the accuracy, the algorithm was tested on more than 1500 objects including faces and non-faces. The results prove that the method is stable and can reject with high accuracy non-face objects from the query.

The rest of this paper is organized as follows. Section II overviews the related work. Section III describes our generic model face validation algorithm. Description of a test data set is given in section IV. Experimental results and conclusion ends this paper.

2 Related Work

Face validation problem on 3D models is mainly decomposed to the problem of anchor points localization. To localize main points on the face researchers are using different tools and methods like geometrical analysis [2] or differential geometry [3].

To detect automatically face on a scanned scene, Mian et al. in [2] proposed a slice searching algorithm for the nose tip point. Searching for nose tip candidates is performed on each slice of the model. Authors center a circle on multiple horizontal intervals on the slice and inscribe a triangle using center of the circle and points of intersection of the slice and the circle. The point with maximum triangle altitude becomes the nose tip candidate on the slice. The point which has maximum confidence is taken as the nose tip of the face. Authors tested the algorithm on the FRGCv2 dataset, which is 2.5D face dataset. In this article the problem has been simplified to face detection on the face models without any non-face examples. Moreover considering only horizontal slices makes the algorithm dependent to rotations.

Curvature based approach has been proposed by Colombo et al. in [4], where authors describe how a face can be detected based on the curvatures analysis. Approach used in the article is based on HK-Classification which can portion a face surface to regions of convex and concave shapes. Their algorithm in comparison with previous one has a validation stage. In the validation stage authors are comparing segment from a face, which has been cropped based on anchor point candidates (the nose tip and the eyes corners). Such validation stage based on range image is exposed to holes and spikes in the model.

Researchers mainly search and detect a face on 3D models based on face anchor points, therefore rest of related work will be devoted to the anchor point localization on 3D face

models.
Many methods for landmark validation are not invariant to different type of rotations. Lu et al. in [5] proposed an approach which is based on the assumption that the nose tip has the highest Z value. Based on that, they rotate the face and in each rotation search for the largest value in the Z-axis. Verification stage, based on the vertical profile in the nose tip point gives them correct result which is not sufficient for the rotation along the Z-axis.
In this paper we propose generic face model validation and curvature based face main points searching algorithm. First stage in the algorithm is convex and concave regions searching and main point candidates localization based on maximum Gaussian curvature value in each region. Validation step relay on a scaled face generic model fitting to each combination of main points candidates. Such method is invariant to the rotation on all axes, invariant to the resolution and also do not needs help from texture images.

3 Differential geometry tools and generic face model for face validation

Our algorithm for automatic face models validation is based on a generic face model fitting to adequate face regions. Searching of the appropriate convex and concave regions is performed by Mean and Gaussian curvatures analysis on each vertex. Classification based on signs of the curvatures assigns the vertex to the convex or the concave regions. Searching is performed using large face area for curvatures calculation to estimate the most marked out regions (fig. 1). Having those regions and based on the generic face model (described in section 3.1) face can be validated. More details about the curvatures calculation can be found in [3].

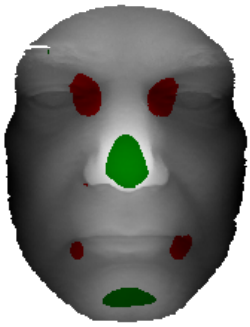


Figure 1 – *Main convex and concave regions on the face model (red - concave regions, green - convex regions).*

Points candidates extraction. Having Gaussian and Mean curvatures values calculated on each vertex, convex and concave regions extraction can be performed. To validate a query model, most marked out regions were chosen. Based on HK-Classification such regions can be easily extracted.
Models, face or non-face, can be more complicated having plenty of convex and concave regions. To reduce number of regions and to select only the most marked out, HK-Classification thresholding process can be performed. To localize correct regions which belong to the eyes and the nose Gaussian curvature was thresholded (figure 1 shows result for the main regions extraction from a face model).

Having those regions, main point searching can be performed in each region separately. To localize the nose tip and the inner eyes corners in each region simply maximum Gaussian curvature value was localized. Max Gaussian curvature value corresponds to the most convex/concave point in the region. Point with the maximum Gaussian curvature in **the convex regions will become the nose tip candidate**, while point with the maximum Gaussian curvature in **the concave regions will become the inner eye point candidate**.

3.1 Face generic model building

Our generic face model (figure 2) was built based on 40 models selected from the IV2 [1] dataset. The generic model is composed from 9 main face points (fig. 2) which positions were calculated based on selected 2.5D facial models. The models were firstly manually landmarked with 9 feature points. Next, all models were translated and rotated to a frontal position having the nose tip as the origin. Fusion of all models relay on mean main point position calculation in 3D space. The generic model was further normalized so that the distance between the two eye inner corners was equal 1 mm.

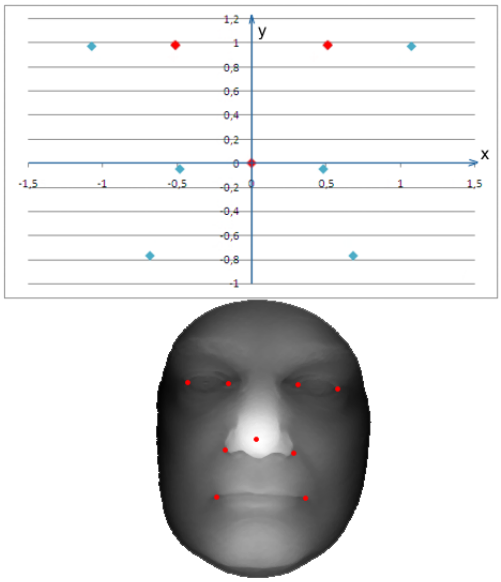


Figure 2 – *Generic model made based on 40 models from IV2 data set (x,y projection, red points - main three points - inner corners of the eyes and the nose tip).*

1. IV2 - French biometric data base created in cooperation of few laboratories (http ://lsc.univ-evry.fr/techno/iv2/PageWeb-IV2.html).

3.2 Face validation

The main part of the whole algorithm is based on the generic face model fitting to the query model. Well fitting of the generic model to the query model, means that the distance (equation 1) between points of the generic model and the closest points of the query model is small and less than some threshold.
The distance can be calculated based on equation :

$$dist = \sqrt{\sum_{i=0}^{n}(GP_i - CPQM(GP_i))^2}, \qquad (1)$$

where n is a number of points in the generic model (in our case 9), GP_i is a point in the generic model, $CPQM(GP_i)$ is the closest point on the query model to the point of the generic model.
To calculate distance between the generic model and the query model, first of all a correspondence between those two has to be established. The correspondence can be established based on the extracted concave and convex points from the query model, **related to the nose tip candidates and the inner eyes corners candidates on the face models**, (section 3) and the same points from the generic model (figure 3).
Query model can have numerous of convex and concave regions, where from each region in the previous section we have extracted the most convex and the most concave point. Having unknown number of the convex and the concave points (related on face models to the nose and the eyes) candidates and without any prior knowledge about the query model all combinations of points candidates have to be considered.
Heaving two sets of points, the nose tip candidates and the eyes inner corner candidates (section 3), each combination of three points (two concave points and one convex point) is considered to calculate translation and rotation between the generic face model and the query model. Figure 3 shows some correspondences between the Generic Model and a Query Model. Rotation and translation between two sets of points with known correlation can be calculated using Singular Value Decomposition algorithm [6, 7, 8], which is a matrix decomposition algorithm, used iteratively in the Iterative Closest Point algorithm. SVD let us to find fine translation and rotation between objects in correspondence based on their covariance matrix.
Having translation and rotation for each selected combination of points, generic model can be moved over the query model surface. To deal with scale changes, generic model was scaled based on the distance between the concave point candidates.
Now when scaled generic model is over the surface and anchored in the concave and the convex points, the distance can be calculated based on equation 1. This algorithm has to be repeated for all combinations of main point's candidates and the smallest distance from all distances between the generic model and the query model surface which is less than some threshold can validate face and also pick up correct anchor points on the model which will be the nose tip and the inner corners of the eyes.
Tests made on face models lead that the sum of distances between generic model and face model cannot be more than 70 mm which means that each point of generic model have to be in the distance less than 7.7 mm.

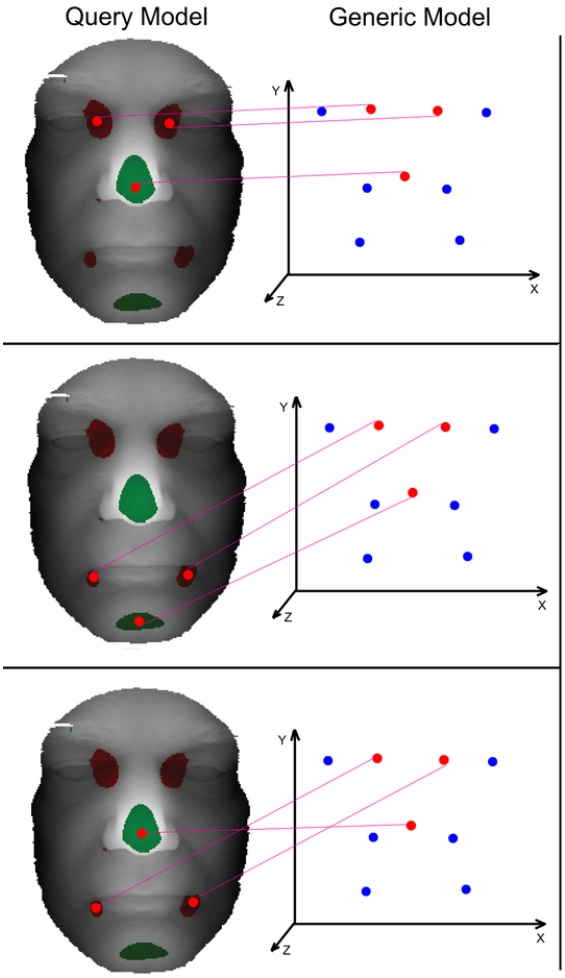


Figure 3 – *Few examples of different correspondence combinations of main points from the generic face model and the points candidates from the query model.*

4 Data set characteristic

The aim of this article is to deal with the problem of models validation for recognition purposes. In real world subjects might be non-cooperative, which means that can move during scanning process. This kind of situations causes many problems during acquisition and recognition.
The main goal building the test dataset was to simulate non-cooperative behavior of the subject. The whole test dataset contains three different datasets/subsets and can be divided based on their origin.

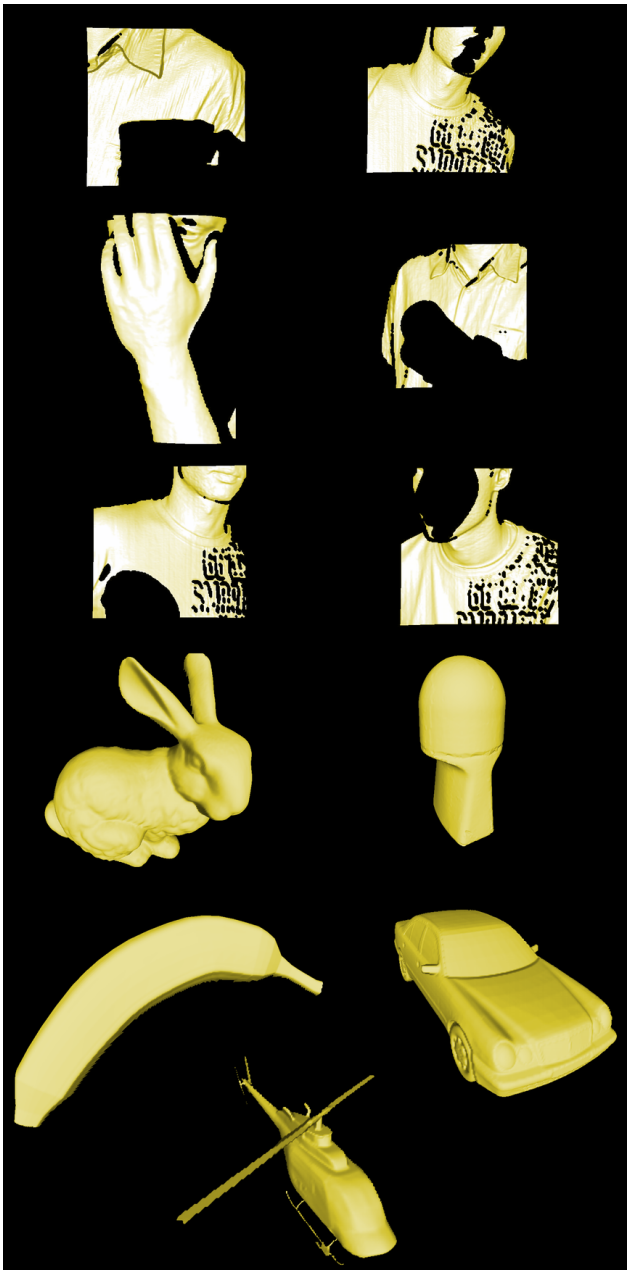


Figure 4 – *Examples of wrong query models.*

Figure 5 – *Examples of correct query models.*

The first part of the test dataset is a dataset called ”Unsupervised conditions” (tab. 1). It is our own data set, in which uncontrolled conditions were simulated. Data set contains 77 non-face models scanned during subject movement which causes scanning of some clothes or part of the face (upper part of figure 4) and correctly scanned faces (25 models) with some rotations and partial occlusions. All models were scanned using non-contact 3D digitizer Minolta Vi-300 with resolution 400x400 points (fig. 7).

To increase number of non-face models dataset has been supplemented by adding some models from Stuttgart Range Image Database [9] (tab. 1) like a bunny or a car (lower part of figure 4). Stuttgart Range Image Database contains a collection of synthetic range images taken from high-resolution polygonal models available on the web. Whole data set contains 42 models where each model has 258 range scans which give 10836 range models. For our purposes only part of this dataset was added to the experiments.

To validate algorithm ability to accept face models, test dataset has been supplemented by adding 933 models from FRGCv1.0 dataset [10]. This dataset is a frontal position face dataset which can be used to prove that algorithm is able to pass face models to the recognition pipe.

Un-supervised conditions	count
part of face/shoulders	77
Face-frontal	12
Face-occlusion	9
Face-rotation	4
SUM	102
FRGC v1.0	count
Frontal faces	933
Stuttgart Range Image Database	count
agfa	66
auto	66
banana	66
bunny	66
copter	66
bunny	66
deo	66
duck	66
SUM	528
SUM of all	**1563**

Table 1 – *Test data sets characteristic.*

5 Experiments

Based on division of the test dataset (tab. 1), algorithm ability to reject non-face models and to keep face models for future recognition purposes was performed.
I order to assess propriety of face models validation algorithm, whole test data set was processed. The threshold dividing test set to face and non-face models was set to 70 mm.
Test 1 : "Un-supervised conditions".
The first test was made on our own models scanned in the laboratory (fig. 7) (models simulate un-supervised conditions during acquisition). This test had to prove that algorithm is able to reject non-face models (part of a face or shoulders) from a recognition pipe while face models with some rotations or small occlusions should pass the conditions.
During this test all **face models** (25) were accepted : the minimum distance (equation 1) between the Generic Model and a face model was between 10.39 mm and 54.73 mm, much less than the face /non-face threshold. In case of **non-face models**, only one non-face model per 77 passed the distance condition and was accepted as a face model, rest of them were rejected as a non-face ones. The correct distance between this model and the Generic Model was 53.9 mm.
Test 2 : "FRGC v1.0".
The aim of the second test was to ascertain that the algorithm is able to labeled face models as correct ones. To ensure large variation of faces FRGCv1.0 dataset was chosen with 933 2.5D face models with frontal positions. All from 933 face models pass the test of a face validation with correct distances to the Generic Model between 9.14 mm and 60.40 mm.

Test 3 : "Stuttgart Range Image Database".
The last test was made to ascertain algorithm ability to reject non-face models. Test was made on the subset of "Stuttgart Range Image Database" with 528 2.5D non-face models. During this test all non-face models were rejected as not correct ones with distances to the Generic Face Model between 81.46 and 359.10

All tests results can be seen in figure 6, where data has been organized to show crossing part of face and non-face models, distance of face models to the Face Generic Model has been sorted in descending order while for non-face models in ascending order. Using 958 face models and 605 non-face models for validation purposes only one model did not pass the conditions and has been incorrectly accepted as a face model, all face models have passed validation conditions and was labeled as face models.

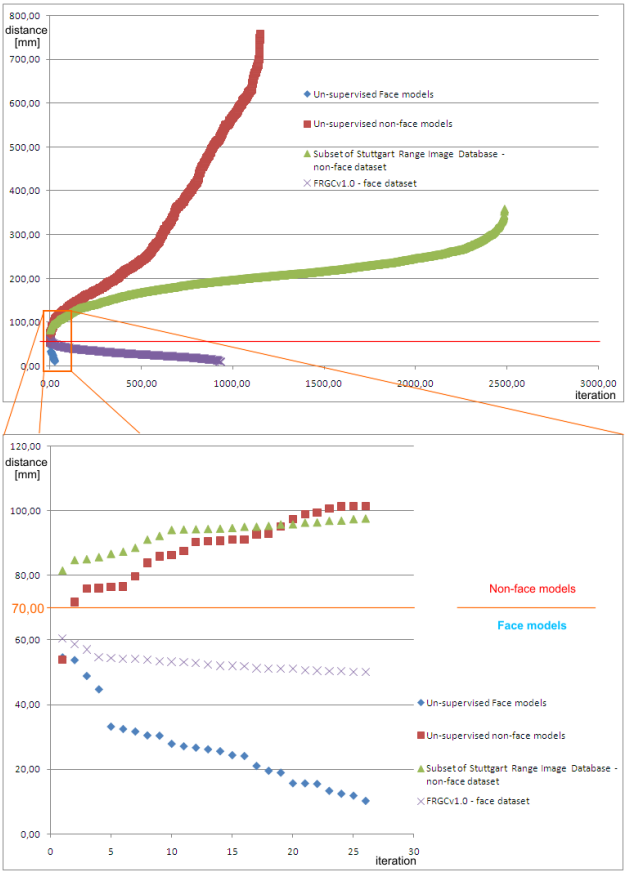


Figure 6 – *Results of face models validation organized to show crossing part between face and non-face models (vertical axis shows distance between the Generic Model anchored in selected convex and concave points and a query model, horizontal axis shows iteration, plot has been divided to different subsets in the data set).*

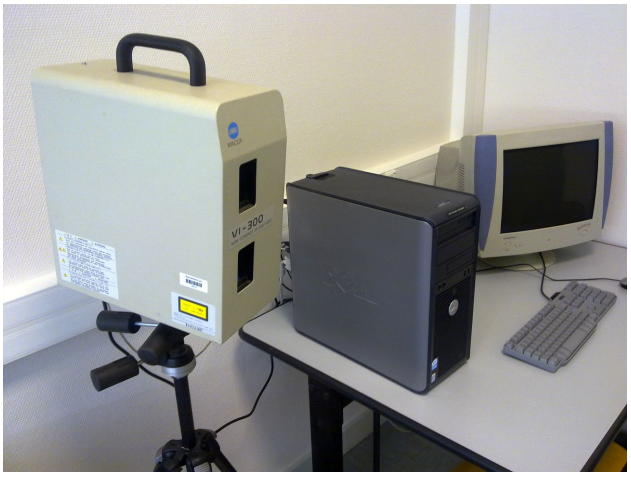


Figure 7 – *Scanning environment to simulate "Unsupervised conditions".*

6 Conclusion and future work

In this paper we presented an automatic algorithm for 3D face validation purposes. Recently 3D face recognition has been perceive as a major face recognition solution, while 3D face validation or searching on 3D scans is omitted.

Our solution for a query model validation/labeling as a face or a non-face model is based on main convex and concave points searching on the query model. Those points correspond to the inner corners of the eyes and the nose tip on face models. Based on those points correspondence between the Generic Face Model can be set up. The validation process is based on distance measurement between the query model surface and the Face Generic Model moved over the face surface and anchored in many combinations of the main convex and concave points from the query model. The smallest distance between the Generic Face Model anchored in one of main query model point's combination and the query model surface gives the measurement score. If measurement score is less than face/non-face threshold, query model is considered as a face model otherwise model is labeled as a non-face model and can be rejected from the recognition pipeline.

Presented results prove, that the Generic Face Model validation algorithm is stable (1/1563 models has been incorrectly labeled) and accepts with high accuracy face models (all, 958 face models were labeled as a face).

In our future work, we are moving to partial face models validation, to give more information to recognition algorithms where a big advantage will be to know, what part of model is missing.

7 Acknowledgment

This work was partially carried out within the French FAR3D project supported by ANR under the grant ANR-07-SESU-004 FAR3D.

Références

[1] Timothy C. Faltemier, Kevin W. Bowyer, et Patrick J. Flynn. Rotated profile signatures for robust 3d feature detection. *IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2008.

[2] Ajmal Mian, Mohammed Bennamoun, et Robyn Owens. An efficient multimodal 2d-3d hybrid approach to automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11) :1927–1943, 2007.

[3] Przemyslaw Szeptycki, Mohsen Ardabilian, et Liming Chen. A coarse-to-fine curvature analysis-based rotation invariant 3d face landmarking. *International Conference on Biometrics : Theory, Applications and Systems*, pages 1–6, 2009.

[4] Alessandro Colombo, Claudio Cusano, et Raimondo Schettini. 3d face detection using curvature analysis. *Pattern Recognition*, 39(3) :444–455, 2006.

[5] Xiaoguang Lu et Anil K. Jain. Automatic feature extraction for multiview 3d face recognition. *Automatic Face and Gesture Recognition*, pages 585–590, 2006.

[6] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *Transactions on Pattern Analysis and Machine Intelligence*, 13 :376–380, 1991.

[7] D.W. Eggert, A. Lorusso, et R.B. Fisher. Estimating 3-d rigid body transformations : a comparison of four major algorithms. *Machine Vision and Applications*, 9 :272–290, 1997.

[8] Gaojin Wen, Zhaoqi Wang, Shihong Xia, et Dengming Zhu. Least-squares fitting of multiple m-dimensional point sets. *The Visual Computer*, 22(6) :387–398, 2006.

[9] G. Hetzel, B. Leibe nad P. Levi, et B. Schiele. 3d object recognition from range images using local feature histograms. *CVPR*, 2001.

[10] P. Jonathon Phillips, Patrick J. Flynn, Todd Scruggs, Kevin W. Bowyer, Jin Chang, Kevin Hoffman, Joe Marques, Jaesik Min, et William Worek. Overview of the face recognition grand challenge. *Computer Vision and Pattern Recognition*, 1 :947–954, 2005.

Application du formalisme multiéchelles microcanonique pour la segmentation des signaux de parole

Vahid Khanagha[1] Khalid Daoudi[1] Oriol Pont[1] Hussein Yahia[1]

[1] INRIA Bordeaux Sud-Ouest (équipe GEOSTAT)

351 Cours de la Libération, Bât. A29, 33405 Talence, France

{vahid.khanagha, khalid.daoudi, oriol.pont, hussein.yahia}@inria.fr

Résumé

Dans la présente communication un cadre d'analyse nouveau, le formalisme multiéchelles microcanonique (FMM), est utilisé pour l'étude des signaux de parole. Le FMM est basé sur le calcul des paramètres géométriques et locaux – les exposants de singularité – qui permettent une analyse non-linéaire de la dynamique complexe et, en particulier, de caractériser la signature intermittente. Également, une mesure cumulative de ces exposants qui a la propriété de produire des changements clairs et distinctifs aux frontières des phonèmes est définie. Des expériences préliminaires sur la base de données TIMIT sont présentées. Elles montrent que les exposants de singularité apportent en effet des informations précises sur la dynamique locale de la parole. Ces expériences montrent également que la mesure proposée a un bon potentiel de fournir une méthode nouvelle et puissante pour la segmentation phonétique indépendante du texte.

Mots clefs

formalisme multiéchelles microcanonique, segmentation de signaux de parole.

1 Introduction

Il est théoriquement et expérimentalement démontré que la turbulence et des phénomènes fortement non-linéaires sont présents dans le processus de production de la parole [8, 2, 9, 10]. Toutefois, l'approche traditionnelle du traitement de la parole est basée sur des techniques linéaires qui s'appuient essentiellement sur le modèle source-filtre. L'approche linéaire ne peut pas prendre suffisamment en compte ou saisir complètement la dynamique complexe de la parole. Pour cette raison, le traitement non-linéaire de la parole a gagné une attention considérable au cours des dernières années.

Dans cet article, la dynamique non-linéaire de la parole en utilisant les concepts et les méthodes du cadre de systèmes turbulents est analysée. Notre approche est basée sur le formalisme multiéchelles microcanonique (FMM), qui est un nouveau cadre pour étudier les propriétés géométriques-statistiques des signaux complexes dans une perspective multiéchelles [12, 17]. Le FMM s'est avéré être une approche précise pour modéliser et analyser empiriquement des systèmes complexes et turbulents. Cela est particulièrement vrai pour systèmes invariants d'échelle [11].

Le FMM est une extension de son équivalent canonique qui est plus standard [5, 1]. La particularité du FMM est qu'il est basé sur des paramètres géométriques et locaux, plutôt que de s'appuyer sur des moyennes statistiques – telles que les fonctions de structure ou les fonctions de partition – comme c'est le cas dans le cadre canonique [11]. Par conséquent, cela rend possible l'étude de la dynamique locale des signaux complexes.

Nous montrons que les signaux de parole se situent dans le domaine d'applicabilité du FMM. Nous utilisons ensuite les paramètres locaux calculés par le FMM, appelés exposants de singularité [18], et montrons comment ils apportent des informations utiles pour l'identification des frontières de phonèmes.

1.1 Etat de l'art

La segmentation de la parole a des nombreuses applications potentielles dans la technologie de la parole, de la synthèse de la parole à la reconnaissance automatique de la parole (RAP). La segmentation est idéalement la première étape d'un système RAP, mais l'absence d'algorithmes de segmentation précise a conduit à une approche inverse : une large classe de méthodes de segmentation sont des versions adaptées des HMM à base de reconnaissance phonétique [15]. Ces méthodes de segmentation sont connus comme méthodes « dépendantes du texte », car ils reposent sur une base de données externe fournie de vocabulaire cible et ses transcriptions manuelles. D'autre part, il existe une catégorie de méthodes de segmentation, qui sont « indépendantes du texte ». Elles sont basées sur l'identification des variations des distances caractéristiques de la parole [3]. Les méthodes indépendantes du texte ne sont pas limitées à un corpus spécifique et elles s'appuient sur des paramètres acoustiques ou des mesures spectrales de base.

Dans ce papier, les changements temporels de la distribution des exposants de singularité sont exploités avec l'aide d'une mesure cumulative. Nous présentons des expériences

préliminaires qui montrent que cette mesure peut être utilisé facilement pour détecter les frontières de phonèmes. Le document est structuré comme suit : La section 2 introduit des concepts fondamentaux du FMM. D'abord, l'algorithme d'estimation des exposants de singularité est présenté, et ensuite les conditions de validité du FMM sont specifiées. La section 3 montre que la parole est un signal approprié pour ce formalisme et la section 4 discute l'utilisation des exposants de singularité pour la segmentation des signaux de parole. Enfin, dans la section 5, nous tirons nos conclusions.

2 Le formalisme multiéchelles microcanonique

Dans cette section, un bref aperçu sur les bases du FMM est présenté. Un examen plus approfondi de la théorie et les méthodes peut être trouvé dans [17]. Le FMM est basé sur le calcul des exposants d'échelle locaux d'un signal donné dont leur distribution est la quantité clé définissant sa dynamique intermittente. Ces exposants sont utiles pour l'étude des propriétés géométriques des signaux, et ils ont été utilisés dans une grande variété d'applications allant de la compression de données à l'inférence et la prévision de signaux [16, 13].

Comme avec tous les modèles, avant d'appliquer le FMM à un signal donné, la première étape consiste à étudier sa validité. La validité du FMM pour un signal repose sur l'existence d'une loi de puissance locale à chaque point du domaine du signal [17].

Formellement, pour au moins une fonctionnelle dépendante d'échelle Γ_r, la relation suivante doit être valide pour tout instant t :

$$\Gamma_r\left(s(t)\right) \;=\; \alpha(t)\, r^{h(t)} + \mathrm{o}\left(r^{h(t)}\right) \qquad r \to 0 \qquad (1)$$

où $h(t)$ est l'exposant de singularité du signal $s(t)$ [17]. Le facteur multiplicatif $\alpha(t)$ dépend de la fonctionnelle Γ_r choisie, mais pour certains systèmes invariants d'échelle, l'exposant $h(t)$ est indépendante d'elle. Cependant, il est hors du but de ce document d'étudier si la parole a des propriétés d'invariance d'échelle. Le terme $\mathrm{o}\left(r^{h(t)}\right)$ regroupe des perturbations additives qui, pour des échelles petites, sont négligeables par rapport au terme principal. Donc $h(t)$ quantifie la contribution dominante au « degré de régularité » de $s(t)$ à chaque instant du temps.

Si la fonctionnelle est choisie comme l'incrément linéaire, $\Gamma_r\left(s(t)\right) = s(t+r) - s(t)$, les exposants qui en résultent sont exposants Hölder et ils caractérisent des corrélations en loi de puissance. Lorsque des données empiriques sont analysées, il est typiquement difficile d'obtenir une bonne estimation des exposants Hölder à partir d'incréments linéaires : De la discrétisation, du bruit et des corrélations à longue portée entravent le calcul pratique de ces exposants avec l'Eq. (1).

Il existe une définition alternative et plus robuste pour la fonctionnelle Γ_r dans l'Eq. (1), qui est définie à partir de la quantification typique de l'intermittence en turbulence : la mesure « module du gradient ». Cette mesure décrit la dissipation d'énergie à l'échelle r du champ vitesse turbulent. Par conséquent, il est une quantité liée au transfert d'énergie d'une échelle à l'autre. Ainsi, l'exposant associé à la loi de puissance en termes d'échelle caractérise le contenu d'information et les transitions dynamiques du signal [5, 18]. Dans ce cas, la fonctionnelle Γ_r est définie comme la mesure module du gradient à échelle r divisé par le volume de la boule de rayon r :

$$\Gamma_r\left(s(t)\right) \;:=\; \frac{1}{\Lambda(B_r)} \int_{B_r(t)} \mathrm{d}\tau \; |s'(\tau)| \qquad (2)$$

où s' est la dérivée de s, B_r est la boule de rayon r et Λ désigne la mesure de Lebesgue sur l'axe réel. La mise en œuvre pratique pour éviter du bruit et des artefacts de discrétisation consiste à utiliser un support d'ondelettes pour la boule $B_r(t)$.

Il existe un ensemble particulier de points qui expriment le plus d'information sur la dynamique non linéaire du signal : la variété la plus singulière. En fait, pour un point donné, la plus petite est la valeur du coefficient de singularité, la plus grande est la prévisibilité future conditionnée à ce point [16]. Les transitions critiques du système surviennent aux points les plus singuliers, et ce fait a été utilisé avec succès dans nombreuses applications telles que la reconstruction de données perdues ou la prévision de valeurs futures [13].

2.1 Estimation des exposants de singularité

La méthode utilisée dans cette étude pour estimer les exposants de singularité est l'équivalent de l'Eq. (2) en termes de transformée en ondelettes continue. Ceci a l'avantage général de faire face aux particularités des données réelles telles que la discrétisation, le bruit d'acquisition et les corrélations à longue portée. En plus, la transformée en ondelettes annule des contributions polynômiales au terme additif $\mathrm{o}\left(r^{h(t)}\right)$, qui représentent un obstacle commun pour l'estimation précise des exposants de singularité. En général, nous examinons la loi de puissance à chaque instant :

$$\mathbb{T}_\Psi\left[|s'|\right](r,t) \;\propto\; r^{h(t)} \qquad (3)$$

où $\mathbb{T}_\Psi\left[x\right](r,t) := \left(\Psi_r * x\right)(t)$ représente la transformée en ondelettes continue, $\Psi_r(t) := r^{-1}\,\Psi(r/t)$ et Ψ est une fonction appelée ondelette mère.

Il convient de mentionner un autre avantage d'utiliser la transformée en ondelettes continue pour cette estimation : la possibilité de calculer la transformée sur un ensemble d'échelles non entières pour un signal discrétisé. La variable d'échelle r dans l'Eq. (3) peut être assignée des valeurs non entières, fournissant un schéma d'interpolation pour signaux en temps discret.

2.2 Validation du FMM

C'est facile de voir que prendre le logarithme aux deux côtés de l'Eq. (3) révèle une relation linéaire entre le loga-

rithme de la transformée en ondelettes et le logarithme de l'échelle. Il est donc possible d'estimer l'exposant de singularité $h(t)$, à chaque instant t en effectuant une régression log-log de la transformée en ondelettes par rapport à l'échelle. Par conséquent, l'Eq. (3) et donc l'Eq. (1) sont vérifiées pour un signal donné, si nous atteignons des coefficients de corrélation acceptables pour cette régression linéaire. Lorsque cela se produit, le FMM est applicable pour le signal en question.

3 Signaux de parole dans le cadre microcanonique

Dans cette section, nous étudions la validité du FMM pour les signaux de parole. Afin d'estimer les exposants de singularité, nous utilisons une légère modification de l'Eq. (3). En effet, dans notre expérience avec des signaux de parole, nous avons observé qu'une amélioration des coefficients de corrélation est obtenue lorsqu'on prend le logarithme aux deux côtés de l'Eq. (3) et divise par $\log(r)$, de sorte que nous avons la relation linéaire suivante :

$$\frac{\log \mathbb{T}_\Psi \left[s'\right](r,t)}{\log r} = \alpha_\Psi(t) \frac{1}{\log r} + h(t) \qquad (4)$$

Puis, en effectuant la régression linéaire, l'exposant de singularité $h(t)$ est estimé comme le biais de cette relation linéaire.

Nous utilisons l'ondelette lorentzienne comme ondelette mère. Cette ondelette définit une estimation précise pour les exposants les plus petits, au détriment d'une saturation de tous les exposants ≥ 1 [19]. Ceci est souhaitable car les exposants les plus petits sont les plus informatifs et, dans le cas présenté, les exposants obtenus sont loin de la saturation et donc ce fait n'apparaît pas comme une limitation effective. Pour faire la régression, nous avons choisi 10 échelles qui sont log-uniformément espacés entre 1 et 100 échantillons (qui correspondent à l'intervalle de 62.5 μs à 6.25 ms).

D'abord, nous vérifions l'existence des lois de puissance, l'Eq. (3), sur les phonèmes, car ils sont les unités acoustiques fondamentales de la parole. Toutes nos expériences sont effectuées sur la base de données TIMIT. Nous utilisons les transcriptions fournies par TIMIT pour construire une base de données d'étude de 3000 phonèmes : 500 instances différentes d'un phonème représentatif de chaque famille de phonèmes (voyelles, fricatives, occlusives, semi-voyelles et liquides, affriquées et nasales). L'estimation des exposants de singularité est effectuée en utilisant l'Eq. (4). La moyenne des coefficients de corrélation est présentée au Tableau 1.

Ensuite, nous avons effectué la même procédure sur des phrases entières. 500 signaux de parole d'une longueur approximative de 3 à 5 secondes ont été utilisés pour cette expérience. Nous avons obtenu un coefficient de corrélation moyen de 0.96. La perte faible par rapport à la moyenne des valeurs dans le Tableau 1 (qui est de 0.98) s'explique par la présence de longs segments de silence aux phrases entières.

Dans l'ensemble d'étude, cet expérience montre que la procédure d'estimation d'exposants utilisée obtient des coefficients de corrélation d'excellente qualité. Premièrement, ceci suggère que nous attendons une estimation très précise des exposants de singularité. Deuxièmement, ceci suggère aussi que le FMM est applicable aux signaux de parole. Nous pouvons donc procéder maintenant à étudier comment ces exposants apportent des informations utiles pour l'étude de la dynamique de la parole.

4 Application du FMM pour la segmentation du signal de parole

La parole est un signal non stationnaire qui est formé par la concaténation de petites unités acoustiques appelées phonèmes. La détection automatique des frontières entre les phonèmes est une tâche difficile et est encore un problème ouvert qui a de nombreuses applications dans la technologie de la parole. Dans la section 3, nous avons démontré la validité du FMM pour les signaux de parole. Ici, nous présentons nos observations sur les informations apportées par les exposants de singularité sur la dynamique variable des signaux de parole. Puisque des phonèmes différents sont fondamentalement signaux différents, chacun avec ses propriétés dynamiques caractéristiques, nous prévoyons que les exposants de singularité ont un comportement différent à l'intérieur des frontières de chaque phonème.

Afin de démontrer ces changements dynamiques, nous présentons les changements dans la distribution des exposants de singularité conditionnée au temps, $\rho(h|t)$, Figure 1.

Dans la Figure 1–haut, un signal de parole est montré et ses frontières de phonèmes sont représentés par des lignes verticales rouges. Ces limites sont extraites de la transcription manuelle de la base de données TIMIT. La Figure 1–milieu, affiche l'évolution temporelle de la distribution des exposants de singularité. Dans les axes verticaux, nous montrons le rang des exposants de singularité regroupés en classes de 5 centiles de largueur. À chaque instant du temps t, nous prenons une fenêtre de 30 ms, centrée autour de t et nous regroupons les exposants selon son rang global. Lorsque nous voulons représenter la probabilité conditionnelle, chaque colonne a été normalisée par sa norme-∞. Il est remarquable qu'il y a des changements dans la position du maximum et la variabilité de la distribution de h. En plus, la distribution alterne entre unimodale et multimodale, avec les cas unimodaux centrés au milieu du domaine et les cas multimodaux généralement avec deux modes : un à chaque extrême de la gamme.

Toutefois, ces changements de comportement de distribution sont visibles à l'œil, mais ils seraient extrêmement difficiles à détecter numériquement et automatiquement. Par conséquent, c'est nécessaire de définir une nouvelle mesure qui exploite ces changements de distribution. Dans ce but, nous remarquons que le principal changement dans les dis-

Type de phonème	Voyelles	Fricatives	Occlusives	Semi-voyelles et liquides	Affriquées	Nasales
Phonème	/aa/	/dh/	/b/	/el/	/ch/	/en/
Moyenne des coef. de corrélation	0.97	0.99	0.99	0.99	0.99	0.99

Tableau 1 – *La moyenne des coefficients de corrélation de la régression linéaire l'Eq. (3) pour un phonème représentatif de chaque famille.*

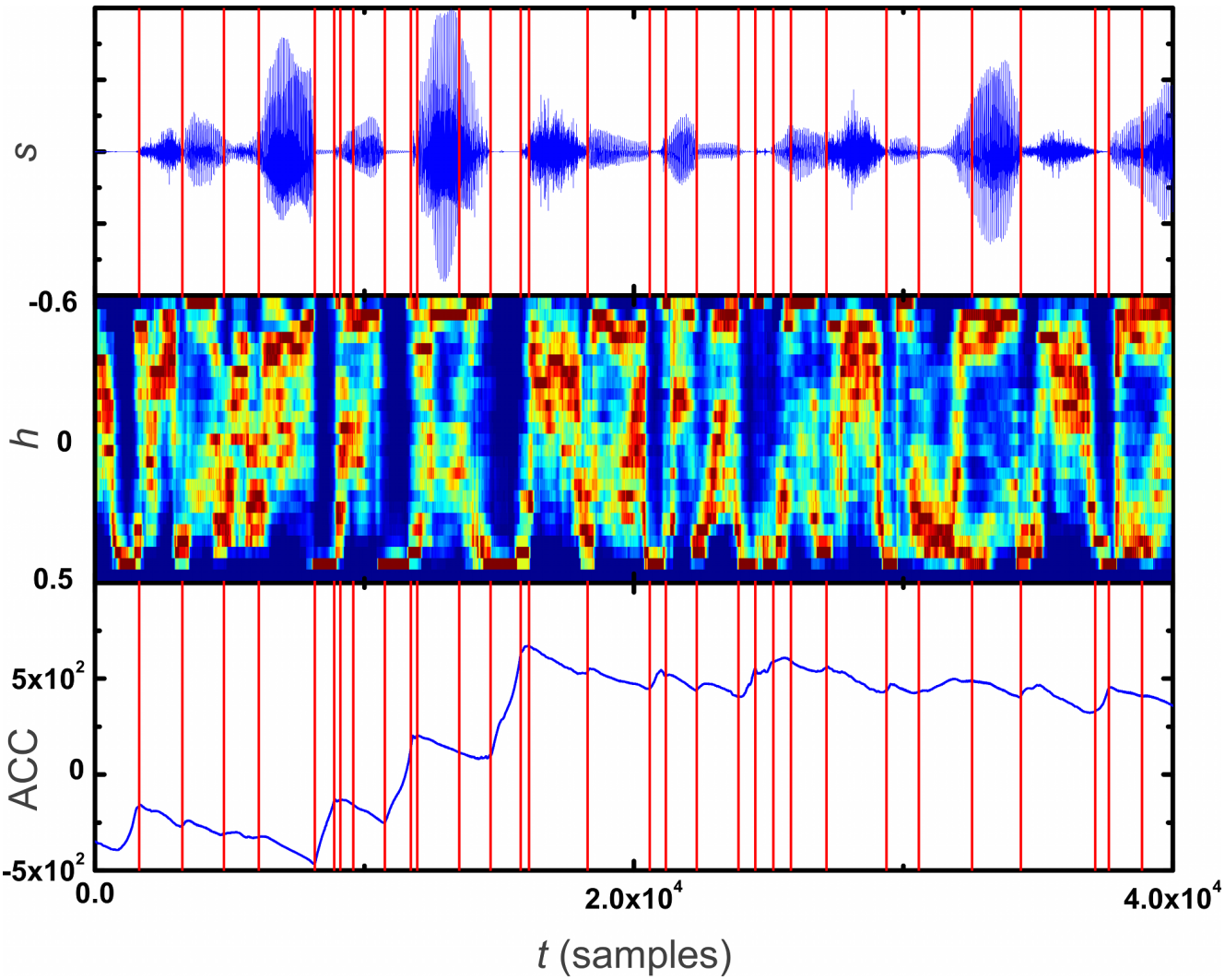


Figure 1 – **HAUT :** *Un signal de parole de la base de données TIMIT, consistant en une portion de 40.000-échantillons de la phrase «* She had your dark suit in greasy wash water all year *». Le signal est proportionnel à la pression d'air et il est présenté dans une échelle normalisée. Le signal est échantillonné à 16 kHz. Les frontières de phonèmes qui ont été manuellement positionnées, sont marquées par des lignes verticales rouges.* **MILIEU :** *Histogramme conjoint de la distribution des exposants de singularité (axe vertical) conditionnée à la fenêtre de temps (axe horizontal). Le rouge correspond à la probabilité maximale et le bleu foncé correspond à une probabilité nulle. L'axe horizontal est divisé en fenêtres de 30 ms. L'axe vertical est divisé en classes d'histogramme de largeur 5 centiles, l'échelle de l'axe est proportionnelle au rang des exposants de singularité, et non pas à leur valeur. Cela permet d'éviter les distorsions de probabilité faible.* **BAS :** *La fonctionnelle proposée pour l'identification des frontières de phonèmes. Il est remarquable que la plupart des frontières de phonèmes concordent avec des forts changements de pente. Pour améliorer la présentation, la tendance globale de la fonctionnelle ACC présentée dans l'Eq. (5) a été soustraite (pour la phrase entière, pas pour la partie présentée).*

tributions de la Figure 1–milieu, est le changement dans les moyennes. En autres termes, nous prévoyons que des phonèmes différents ont des moyennes des exposants de singularité différentes par rapport à leurs phonèmes voisins.

Afin de vérifier cela, nous utilisons la primitive de la fonction exposant de singularité comme estimateur de la moyenne instantanée. Formellement, nous définissons la nouvelle fonctionnelle comme :

$$ACC(t) = \int_{t_0}^{t} \mathrm{d}\tau \; h(\tau) \tag{5}$$

La fonctionnelle obtenue est tracée dans la Figure 1–bas pour le même signal de parole qu'avant. Pour améliorer la présentation des valeurs résultantes dans une fenêtre observable, nous soustrayons la moyenne globale des exposants de singularité de la valeur de l'exposant à chaque instant du temps.

Comme nous attendions, cette nouvelle fonctionnelle révèle les changements dans la distribution d'une manière très précise. En effet, a l'intérieur de chaque phonème la fonctionnelle ACC est presque linéaire (si nous négligeons les fluctuations à petite échelle). En outre, il y a des abruptes changements de pente aux frontières des phonèmes. Ces changements de pente permettent encore d'identifier les frontières entre les phonèmes extrêmement courts, telles que les occlusives. Des observations détaillées sur tout l'ensemble de signaux confirment ce comportement et, donc, la solidité de la fonctionnelle proposée, l'Eq. (5). Ces expériences suggèrent que les exposants de singularité liés au FMM en effet apportent des informations utiles sur les transitions critiques dans les signaux de parole. Elles suggèrent aussi que nous pouvons facilement utiliser ces exposants pour mettre au point une nouvelle méthode robuste pour la segmentation phonétique.

Une évaluation précise de cette mesure nécessite l'application d'un algorithme numérique non-supervisé pour l'identification de points de rupture à courbes bruyants qui sont linéaires par morceaux. C'est le but de nos recherches en cours. Au moment de la rédaction de cet article, nous étudions l'utilisation d'algorithmes Free Knot B-spline [14] pour atteindre cet objectif.

Les méthodes les plus proches de la nôtre que nous avons trouvé dans la littérature sont celles de [7] et [4]. Dans le premier cas, l'analyse de la trajectoire de la variation de la dimension fractale est utilisée pour la segmentation phonétique. Dans la deuxième étude, les auteurs proposent une approche basée sur la fractalité qui utilise les transitions de l'enveloppe de la dimension fractale locale pour déterminer les frontières entre les mots et les phonèmes.

Une comparaison précise entre ces méthodes et la nôtre est au-delà du but du présent document, car ces méthodes – comme la nôtre, pour le moment – seulement distinguent visuellement les phonèmes sans donner une procédure de segmentation automatique. Cependant, nous pouvons dire que la mesure que nous proposons semble beaucoup plus précise et plus facile à incorporer dans un algorithme de segmentation automatique que les mesures données dans [7] et [4].

5 Conclusions

Dans ce document, nous avons d'abord montré que le FMM est un cadre valable pour l'étude des signaux de parole. Dans cette perspective, nous avons ensuite analysé les propriétés locales des signaux de parole par les exposants de singularité en termes du FMM. Nous avons montré que ces exposants apportent des informations importantes sur la dynamique de la parole. Enfin, nous avons proposé une mesure géométrique de quantification qui produit des changements clairs et distincts aux frontières de phonèmes, et donc elle peut être utilisée pour une segmentation automatique et indépendante du texte. Nous soulignons que l'application complète du FMM pour le signal de parole exige des justifications plus précises pour faire face à toutes ses particularités. Pourtant, l'étude présentée dans le présent document révèle le caractère informatif des exposants de singularité, sans aucune manipulation supplémentaire.

Références

[1] A. Arneodo, F. Argoul, E. Bacry, J. Elezgaray, and J. F. Muzy. *Ondelettes, multifractales et turbulence*. Diderot Editeur, Paris, France, 1995.

[2] A. Barney, C. Shadle, and P. Davies. Fluid flow in a dynamical mechanical model of the vocal folds and tract : part 1 & 2. *J. Acoust. Soc. Amer.*, 105(1) :444–466, Nov. 1999.

[3] A. Esposito and G. Aversano. Text independent methods for speech segmentation. In *Summer School on Neural Networks 2004*, pages 261–290, 2004.

[4] P. C. Fantinato, R. C. Guido, S.-H. Chen, B. L. S. Santos, L. S. Vieira, S. B. J, L. C. Rodrigues, F. Sanchez, J. Escola, L. M. Souza, C. D. Maciel, P. R. Scalassara, and J. Pereira. A fractal-based approach for speech segmentation. In *Tenth IEEE International Symposium on Multimedia*, pages 551–555, Los Alamitos, CA, USA, 2008. IEEE Computer Society.

[5] U. Frisch. *Turbulence : The legacy of A.N. Kolmogorov*. Cambridge Univ. Press, Cambridge MA, 1995.

[6] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic-phonetic continuous speech corpus. Technical report, U.S. Dept. of Commerce, NIST, Gaithersburg, MD, 1993.

[7] W. Kinsner and W. Grieder. Speech segmentation using multifractal measures and amplification of signal features. In *Cognitive Informatics, 7th IEEE International Conference on*, Oct. 2008.

[8] I. Kokkinos and P. Maragos. Nonlinear speech analysis using models for chaotic systems. *IEEE Transactions on Speech and Audio Processing*, 13(6) :1098–1109, Jan. 2005.

[9] A. Kumar and S. Mullick. Nonlinear dynamical analysis of speech. *J. Acoust. Soc. Amer.*, 100(1) :615–629, 1996.

[10] M. Little. *Biomechanically Informed Nonlinear Speech Signal Processing*. PhD thesis, Oxford University, 2007.

[11] O. Pont, A. Turiel, and C. Perez-Vicente. Empirical evidences of a common multifractal signature in economic, biological and physical systems. *Physica A*, 388(10) :2025–2035, May 2009.

[12] O. Pont, A. Turiel, and C. J. Pérez-Vicente. Application of the microcanonical multifractal formalism to monofractal systems. *Physical Review E*, 74 :061110–061123, 2006.

[13] O. Pont, A. Turiel, and C. J. Pérez-Vicente. Description, modeling and forecasting of data with optimal wavelets. *Journal of Economic Interaction and Coordination*, 4(1) :39–54, June 2009.

[14] H. Schwetlick and T. Schutze. Least squares approximation by splines with free knots. *BIT Numerical Mathematics*, 35(3) :361–384, Sep. 1995.

[15] D. Torre-Toledano, L. Hernandez-Gomez, and L. Villarrubia-Grande. Automatic phonetic segmentation. *IEEE Transactions on Speech and Audio Processing*, 11(6) :617–625, 2003.

[16] A. Turiel and A. del Pozo. Reconstructing images from their most singular fractal manifold. *IEEE Trans. on Im. Proc.*, 11 :345–350, 2002.

[17] A. Turiel and C. P.-V. H. Yahia. Microcanonical multifractal formalism : a geometrical approach to multifractal systems. part 1 : singularity analysis. *J. Phys. A, Math. Theor.*, 41 :015501, 2008.

[18] A. Turiel and N. Parga. The multi-fractal structure of contrast changes innatural images : from sharp edges to textures. *Neural Computation*, 12 :763–793, 2000.

[19] A. Turiel and C. Pérez-Vicente. Multifractal measures : definition, description, synthesis and analysis. a detailed study. In J.-P. Nadal, A. Turiel, and H. Yahia, editors, *Proceedings of the "Journées d'étude sur les méthodes pour les signaux complexes en traitement d'image"*, pages 41–57, Rocquencourt, 2004. INRIA.

Comparaison de méthodes d'extraction fond/forme pour des scènes de circulation routière

N. Tronson[1] Y. Goyat[1] D. Gruyer[2]

[1] LCPC (Laboratoire Central des Ponts et Chaussées)
Route de Bouaye, BP 4129, 44341 Bouguenais – FRANCE
nicolas.tronson@lcpc.fr, yann.goyat@lcpc.fr

[2] LIVIC (Laboratoire sur les Interactions Véhicules-Infrastructure-Conducteurs)
14, route de la Minière - Bâtiment 824 - Satory 78000 Versailles – FRANCE
dominique.gruyer@inrets.fr

Résumé

Cet article compare trois méthodes pour isoler les objets en mouvement, souvent appelées méthodes d'extraction fond/forme. Nous nous penchons spécifiquement sur les problématiques liées aux scènes routières, dont les enjeux sont très importants en terme de surveillance et d'analyse du trafic. Pour tester indépendamment et avec qualité les influences d'éléments dégradants, nous proposons d'utiliser un logiciel permettant de générer des scènes virtuelles. L'avantage est d'obtenir des vérités terrain associées afin de mesurer précisément la qualité de l'extraction.

Mots clefs

Extraction fond/forme, vérités terrain, simulation, modèle virtuel, comparaison.

1 Introduction

Nous présenterons ici l'étude de trois méthodes utilisées pour l'extraction fond/forme sur une scène de circulation routière, selon différents réglages de paramètres propres à chaque méthode. Des scènes virtuelles permettant de créer la vérité terrain correspondante sont générées. Elles seront ensuite comparées avec les images extraites par chacun des algorithmes, qui pourront donc être notées à l'aide deux critères de qualité.

Nous allons tout d'abord décrire les différents problèmes typiques à une scène routière, puis, présenter une méthode très classique de la littérature et deux méthodes plus originales. Puis seront présentés, le principe utilisé pour générer les vidéos de test, les vidéos, et les critère de mesure utilisés. Enfin, les les tests avec la mesure de l'extraction permettront de comparer ces trois méthodes.

2 Problématiques de l'extraction dans une scène routière

Dans une scène routière, l'acquisition se fait à partir d'une caméra fixe et il existe de nombreuses causes pouvant altérer la qualité de l'extraction :

Le bruit sur une image se caractérise par des valeurs de pixel changeant légèrement. Étant donné que les méthodes d'extraction se basent sur la couleur du pixel, elles doivent donc avoir une certaine tolérance pour ne pas être trop sensible au bruit.

Le mouvement d'objets tel que les branches d'arbre est assez courant. Le but est de détecter ces objets comme appartenant au fond de l'image. Ce mouvement est de façon générale assez répétitif. La méthode d'extraction doit donc être capable d'apprendre les valeurs des pixels les plus récurrents sur un laps de temps à définir, et de considérer que ces valeurs correspondent au fond.

Le changement de luminosité est assez courant quand on observe des véhicules sur la route, ceci peut être lié à un passage de nuage. Les couleurs sur l'image varient donc et l'algorithme peut penser qu'il s'agit d'un objet. L'idéal serait d'être assez peu sensible au changement de luminosité, ou d'être capable de s'adapter rapidement pour limiter les mauvaises détections.

Les scènes sombres ont des niveaux de couleur assez faibles, les différences entre les couleurs sont moins marquées. Il peut donc être plus difficile de détecter le fond de la forme dans ce type de scène.

Les véhicules ayant une couleur proche du fond peuvent être plus difficilement détectés étant donné que les trois algorithmes se basent sur la couleur du pixel pour identifier le fond de la forme.

Les ombres sont des zones qui suivent les véhicules avec des changements de valeur de pixel par rapport au fond. Il est donc assez logique que les algorithmes puissent classer par erreur ces zones comme forme.

La densité de circulation peut poser problème. En effet, les méthodes de détection doivent apprendre le fond, et pour cela, elles considèrent que les valeurs des pixels les plus fréquentes sont des valeurs de fond. Ceci peut poser problème dans des situations urbaines à circulation très dense où le fond n'est pas souvent visible en raison du flux

dense et continu de véhicules.

3 Présentation des méthodes

Nous comparons ici l'extraction fond/forme à l'aide des trois méthodes :
– Mixture de gaussiennes
– Codebook 2 layers
– Vumètre
Le but est de comparer la qualité d'extraction avec chacune des méthodes sur une même scène, et avec une analyse de l'influence des causes altérant la qualité d'extraction décrites précédemment (voir section 2).

3.1 Mixture de gaussiennes (MOG)

Dans cette approche, chaque pixel est modélisé par une mixture de N gaussiennes, $2 \leq N \leq 5$ [1]. Pour $n = 1, \ldots, N$, un élément de la mixture de gaussiennes est représenté par une moyenne $\mu_{\mathbf{n}}$, un écart type σ_n, et un poids α_n ($\sum_n \alpha_n = 1$). On peut remarquer que σ_n est réduit à un scalaire, comme discuté dans [1].
Pour une nouvelle image traitée, la mixture de gaussiennes (pour tous les pixels) est mise à jour pour expliquer correctement les couleurs affichées par chaque pixel. Pour faire ceci, à un instant t, on considère que le modèle $\mathbf{M_t}$ généré pour chaque pixel à partir des mesures $\{\mathbf{Z_0}, \mathbf{Z_1}, ..., \mathbf{Z_{t-1}}\}$ est correct. La vraisemblance pour qu'un pixel appartienne au fond est :

$$P(\mathbf{Z_t}|\mathbf{M_t}) = \sum_{n=1}^{n=N} \alpha_n \mathcal{N}(\mu_{\mathbf{n}}, \mathbf{\Sigma_n}) \tag{1}$$

$$\mathcal{N}(\mu_{\mathbf{n}}, \mathbf{\Sigma_n}) = \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma_n}|^{1/2}} e^{-\frac{1}{2}(\mathbf{Z_t}-\mu_{\mathbf{n}})^T \mathbf{\Sigma}^{-1}(Z_t-\mu_{\mathbf{n}})} \tag{2}$$

avec d la dimension de l'espace de couleurs de la mesure $\mathbf{Z_t}$.
Pour mettre à jour le modèle, on associe d'abord la mesure $\mathbf{Z_t}$ à une gaussienne n' si

$$||\mathbf{Z_t} - \mu_{\mathbf{n}}|| < K\sigma_n \tag{3}$$

où K vaut 2 ou 3. L'opérateur $<$ est vrai si toutes les composantes du vecteur à gauche sont inférieures à $K\sigma_n$.
Cette mesure représente le fond si la gaussienne n' explique le fond de la scène. En fait, le poids $\alpha_{n'}$ est élevé. Cette gaussienne est alors mise à jour :

$$\alpha'_n \leftarrow (1-\delta)\alpha'_n + \delta \tag{4}$$
$$\mu'_{\mathbf{n}} \leftarrow (1-\delta)\mu'_{\mathbf{n}} + \delta\mathbf{Z_t} \tag{5}$$
$$\sigma^{2'}_n \leftarrow (1-\delta)\sigma^{2'}_n + \delta(\mathbf{Z_t} - \mu'_{\mathbf{n}})^T(\mathbf{Z_t} - \mu'_{\mathbf{n}}) \tag{6}$$

avec δ le coefficient d'apprentissage. Il représente la vitesse d'adaptation du modèle. Pour toutes les autres gaussiennes $n \neq n'$, la moyenne et la variance ne sont pas modifiées, mais :

$$\alpha_n \leftarrow (1-\delta)\alpha_n \tag{7}$$

Si le test 3 échoue, le pixel est associé au 1er plan. La gaussienne ayant le plus petit poids est réinitialisée avec la mesure actuelle :

$$\alpha_n = \delta \tag{8}$$
$$\mu_{\mathbf{n}} = \mathbf{Z_t} \tag{9}$$
$$\sigma_n^2 = \bar{\sigma}^2 \tag{10}$$

avec $\bar{\sigma}^2$ une variance élevée. Ces affectations sont aussi appliquées pour l'initialisation de la mixture.

3.2 Codebook 2 layers (CB2)

Cette méthode [4], est très largement inspirée du codebook [2]. Mais elle en diffère, en utilisant deux codebooks, bibliothèques de données pour chaque pixel contenant des informations pour modéliser le fond. Ceci a été réalisé de manière à pouvoir retenir des valeurs de pixels qui ont appartenu au fond, mais qui pourraient redevenir du fond, c'est typiquement le cas avec les mouvements de branches d'arbre.
Chaque codebook contient des éléments appelés codeword (CW) pour modéliser le fond de l'image, chacun des CW contient ces informations :
– v_i : valeur moyenne du pixel (R,V,B)
– I_{max} : limite maximale d'intensité du CW
– I_{min} : limite minimale d'intensité du CW
– f : fréquence du CW (nombre d'occurrences)
– λ : nombre maximal d'images où le CW ne correspond à aucun pixel
– p : première occurrence du CW
– q : dernière occurrence du CW
Le principe est le même qu'avec le codebook simple, mais avec deux codebooks par pixel : un principal appelé M, et un secondaire appelé H.
Le traitement se fait en 2 phases : une phase d'apprentissage qui sert à créer les codebooks principaux initiaux, et une phase de soustraction pour extraire le fond de la forme.
Pour chaque nouveau pixel $x_t = (R, V, B)$, son intensité I_t est calculée par $I_t = \sqrt{R^2 + V^2 + B^2}$
La distorsion de couleur δ entre ce pixel $x_t = (R, G, B)$ et un codeword c_i avec $v_i = (\bar{R}_i, \bar{V}_i, \bar{B}_i)$ peut être calculé par :

$$\langle x_t, v_i \rangle^2 = (\bar{R}_i R + \bar{V}_i V + \bar{B}_i B)^2 \tag{11}$$
$$\|v_i\|^2 = \bar{R}_i^2 + \bar{V}_i^2 + \bar{B}_i^2 \tag{12}$$
$$\|x_t\|^2 = R^2 + V^2 + B^2 \tag{13}$$
$$colordist(x_t, v_i) = \delta = \sqrt{\|x_t\|^2 - \frac{\langle x_t, v_i \rangle^2}{\|v_i\|^2}} \tag{14}$$

Un pixel x_t avec une intensité I_t correspond à un codeword c_i avec une valeur de pixel v_i et I_{min}, I_{max} si I_t est dans l'intervalle $[I_{min}, I_{max}]$ et la distorsion de couleur δ respecte $\delta < \epsilon$
En phase d'apprentissage, seule la couche M est construite, H reste vide. Pour un nouveau pixel x_t, on cherche un CW dans M correspondant à x_t. Si on en trouve un, il est mis à

jour avec x_t, sinon un nouveau CW est crée à partir de la valeur de x_t.
Un nouveau codeword est crée avec un pixel x_t de la façon suivante :

$$v_i \leftarrow (R, V, B) \quad (15)$$
$$I_{min} \leftarrow max\{0, I_t - \alpha\} \quad (16)$$
$$I_{max} \leftarrow min\{255, I_t + \alpha\} \quad (17)$$
$$f \leftarrow 1;\ \lambda \leftarrow t-1;\ p \leftarrow t;\ q \leftarrow t \quad (18)$$

avec α une valeur représentant une tolérance d'intensité.
Pendant la phase d'apprentissage, un codeword est mis à jour par un pixel x_t comme ceci :

$$\bar{R} \leftarrow \frac{\bar{R} \times f + R}{f+1} \text{ (de même pour V et B)} \quad (19)$$
$$I_{min} \leftarrow \frac{I - \alpha + f \times I_{min}}{f+1} \quad (20)$$
$$I_{max} \leftarrow \frac{I + \alpha + f \times I_{max}}{f+1} \quad (21)$$
$$f \leftarrow f+1;\ \lambda \leftarrow max\{\lambda, t-q\};\ p \leftarrow p;\ q \leftarrow t \quad (22)$$

En phase de soustraction, pour un nouveau pixel x_t, on cherche un CW dans M correspondant à x_t. Si on en trouve un, il est mis à jour avec x_t et le pixel est considéré comme appartenant au fond. Sinon, on cherche un CW correspondant à ce pixel dans H, si on un trouve un, on le met à jour avec x_t, sinon, on en crée un nouveau dans H avec la valeur de x_t.
Un CW est mis à jour comme précédemment dans la phase d'apprentissage à l'exception de I_{min} et I_{max} qui sont mis à jour de la façon suivante, avec β un coefficient pour changer la vitesse d'adaptation :

$$I_{min} \leftarrow (1-\beta)(I_t - \alpha) + \beta \cdot I_{min} \quad (23)$$
$$I_{max} \leftarrow (1-\beta)(I_t + \alpha) + \beta \cdot I_{max} \quad (24)$$

Ensuite, les modèles M et H sont affinés avec ces règles :
– Supprimer les CWs de H ayant $\lambda > T_H$
– Déplacer les CWs restant plus de T_{add} dans H vers M
– Supprimer les CWs de M n'apparaissant pas plus longtemps que T_{delete}

3.3 Vumètre (VUM)

Le vumètre par Goyat *et al.* [3] est un modèle non-paramétrique, basé sur une estimation discrète de la probabilité de distribution. Il s'agit d'une approche probabiliste pour définir le modèle du fond. Soit I_t une image à l'instant t, $y_t(u)$ donne les valeurs RVB du pixel u. Un pixel peut prendre deux états, (ω_1) s'il appartient au fond, (ω_2) s'il appartient au 1er plan. Cette méthode essaye d'estimer $p(\omega_1 \mid y_t(u))$. Avec 3 composantes de couleur i (Rouge, Vert, Bleu), la fonction de densité de probabilité peut être approximée par :

$$p(\omega_1 \mid y_t(u)) = \prod_{i=1}^{3} p(\omega_1 \mid y_t^i(u)) \quad (25)$$

avec

$$p(\omega_1 \mid y_t^i(u)) \approx K_i \sum_{j=1}^{N} \pi_t^{ij} \delta(b_t^i(u) - j) \quad (26)$$

où δ est le symbole de Kronecker, $b_t(u)$ donne le vecteur d'index de la classe associée à $y_t(u)$, j est un index de classe, et K_i est une constante de normalisation permettant de garder à chaque instant :

$$\sum_{j=1}^{N} \pi_t^{ij} = 1 \quad (27)$$

π_t^{ij} est une fonction de masse discrète représentée par une classe.
A la première image $(t = 0)$, les valeurs des classes sont mises à $\pi_0^{ij} = 1/N$ pour garder une somme à 1 comme dans l'équation 27. A chaque nouveau pixel, sa valeur correspond à une classe π_t^{ij}, son niveau est mis à jour de cette façon :

$$\pi_{t+1}^{ij} = \pi_t^{ij} + \alpha \cdot \delta(b_{t+1}^i(u) - j) \quad (28)$$

Après un certain nombre d'images, les classes modélisant le fond ont une valeur élevée. Pour pouvoir décider à quel moment un pixel appartient au fond ou non, un seuil T (voir figure 1) est défini.

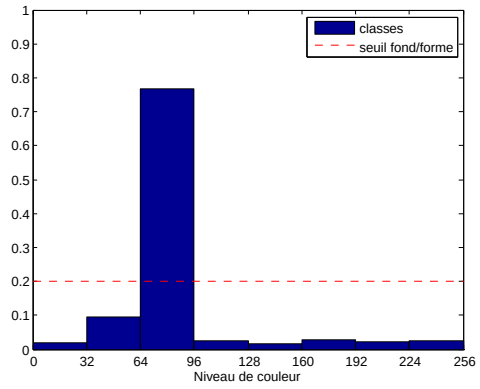


Figure 1 – *Vue des niveaux de classes et du seuil du vumètre*

Chaque nouveau pixel ayant une classe correspondante sous ce seuil sera détecté comme appartenant au fond.
En mode RVB, chaque pixel est modélisé par 3 vumètres (un par composante). Pour considérer un pixel comme fond, il doit être détecté comme fond par chaque vumètre.
Afin d'améliorer la détection et réduire des problèmes liés aux seuils entre deux classes, la valeur des classes au voisinage de la classe correspondante à un pixel est aussi mise à jour, mais de façon moindre.
Pour obtenir un bon apprentissage et une bonne adaptation de l'algorithme, il est nécessaire de bien choisir les paramètres (taux d'apprentissage α et seuil T). Ces valeurs peuvent être changées en fonction de la luminosité ou de la vitesse des véhicules suivis.

4 Génération des scènes et de leurs vérités terrains

Le principal problème pour mesurer la qualité de l'extraction provient de la difficulté à obtenir une vérité terrain que l'on pourra ensuite comparer à l'image extraite. La solution est d'identifier à l'aide d'un logiciel de dessin, de façon manuelle, les zones de formes et les zones de fond, et cela pour chacune des images de la séquence à analyser. Ce travail est assez fastidieux, et ne permet pas de tester les algorithmes avec beaucoup de séquences étant donné le temps nécessaire à la réalisation des vérités terrains.

La solution que nous proposons, consiste à générer des scènes virtuelles à l'aide d'un logiciel. Ceci a principalement deux gros avantages :

– La scène est paramétrée, donc on peut choisir d'isoler un seul facteur (exemple : le bruit), afin de voir le comportement de chaque algorithme face à celui-ci.
– Pour chaque séquence générée (figure 2), on obtient une séquence de vérités terrains associées (figure 3), permettant ainsi une mesure plus précise de la qualité d'extraction.

Figure 2 – *Scène générée*

Figure 3 – *Vérité terrain*

Les scènes sont réalisées grâce au logiciel SiVIC développé par le LIVIC [5], qui permet de générer à l'aide de scripts des scènes avec leurs vérités terrain associées. Les scripts permettent de contrôler chaque paramètre (luminosité, trajectoires des véhicules, bruit sur l'image, emplacement de la caméra, ...).

5 Méthodes de comparaison

Nous allons comparer les trois méthodes décrites précédemment sur différents critères. Le principal, celui dont nous allons parler dans cette partie, est bien entendu la qualité d'extraction. Nous tenons aussi compte de la vitesse d'exécution.

5.1 Classement des pixels

Afin d'analyser la qualité de l'extraction, l'image i obtenue après traitement par l'un des algorithmes est comparée avec l'image de vérité terrain correspondante. Les pixels sont donc classés suivant quatre catégories :

– VP_i (vrais positifs) : 1er plan détecté comme 1er plan
– FP_i (faux positifs) : fond détecté comme 1er plan
– VN_i (vrais négatifs) : fond détecté comme fond
– FN_i (faux négatifs) : 1er plan détecté comme fond

Pour chaque image i de la séquence, on compte le nombre de pixels dans chacune de ces quatre catégories.

5.2 Mesure Δ

Pour analyser la qualité de l'extraction à partir des pixels classés, on utilise une mesure appelée Δ. Le principe revient à calculer deux taux :

– La sensibilité (Se) : $Se_i = \frac{VP_i}{VP_i+FP_i}$
– La spécificité (Sp) : $Sp_i = \frac{VN_i}{VN_i+FP_i}$

La sensibilité reflète une bonne détection d'un objet, alors que la spécificité met plutôt en valeur la bonne détection du fond. L'idéal est d'avoir ces deux valeurs à 1.

On place ensuite les points Se_i en fonction de $1 - Sp_i$ pour toute une séquence (voir figure 4). La détection parfaite est caractérisée par le point de coordonnées (0,1). Plus on sera proche de ce point idéal, plus l'extraction pourra être considérée comme bonne. La droite de non discrimination passant par les points (0,0) et (1,1) montre que l'on ne parvient pas à différencier le fond de la forme.

Pour calculer la qualité de l'extraction, on mesure la distance Δ sur l'axe des ordonnées entre le point parfait (0,1) et la droite parallèle à la droite de non discrimination passant par le point à tester. On effectue cette mesure pour chaque point, puis on calcule la moyenne de ces distances, soit pour N points de coordonnées x_i et y_i, ce qui revient à calculer :

$$\Delta = \frac{1}{N} \sum_{n=1}^{n=N} 2 - Sp_i - Se_i \qquad (29)$$

L'idéal est d'avoir Δ qui tend vers 0.

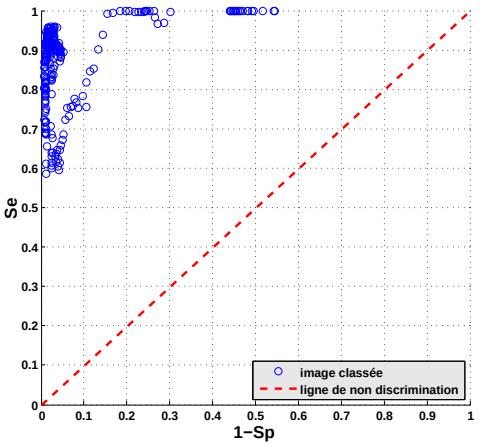


Figure 4 – *images dans un repère sensibilité/spécificité*

5.3 Mesure F

Un autre critère très intéressant est la mesure F. Pour cela on calcule la précision et le rappel d'après les pixels classés suivant les quatre catégories décrites précédemment. On a

donc

$$Prec_i(P) = VP_i/(VP_i + FP_i) \quad (30)$$
$$Prec_i(N) = VN_i/(VN_i + FN_i) \quad (31)$$
$$Rap_i(P) = VP_i/(VP_i + FN_i) \quad (32)$$
$$Rap_i(N) = VN_i/(VN_i + FP_i) \quad (33)$$
$$Prec_i = (Prec_i(P) + Prec_i(N))/2 \quad (34)$$
$$Rap_i = (Rap_i(P) + Rap_i(N))/2 \quad (35)$$
$$F_i = \frac{Prec_i \times Rap_i}{Prec_i + Rap_i} \quad (36)$$

L'idéal sera donc d'obtenir une mesure proche de 1, ce qui montrerait une extraction parfaite. La note F attribué sera la moyenne de tous les F_i sur la séquence.

6 Tests

6.1 Séquences

Pour analyser ces algorithmes, nous utilisons différentes scènes avec leur vérités terrain associées. La première est réelle, il s'agit de la séquence data3 venant de l'IPPR contest 2006 [1]. Les scènes suivantes, sont simulées à l'aide du logiciel SiVIC.

- **vidéo 1** (fig. 5) : Scène réelle, rue avec piétons et bus, scène assez sombre.
- **vidéo 2** (fig. 6) : carrefour, véhicules sur différentes voies, 3 voitures et piétons, bruit assez fort, changement de luminosité faible.
- **vidéo 3** (fig. 7) : rue, scène assez sombre, 1 bus, véhicules variés et des piétons, changement brusque de luminosité à mi-séquence.
- **vidéo 4** (fig. 8) : rond-point, un seul véhicule qui ne génère pas d'ombre, un arbre au milieu du rond-point avec ombre, l'arbre bouge, mouvement du soleil.
- **vidéo 5** (fig. 9) : circulation dense, voitures de tailles et couleurs différentes et 1 bus, ombres venant des véhicules et des bâtiments, masquages entre véhicules.
- **vidéo 6** (fig. 10) : identique à la vidéo 5 mais filmée sous un angle différent : de l'autre côté de la route.
- **vidéo 7** (fig. 11) : vue de dessus, bruit assez fort, automobiles de couleurs et tailles variées et 1 bus, ombres des véhicules.
- **vidéo 8** (fig. 11) : identique à la vidéo 7, bruit faible.

Figure 5 – *vidéo 1*

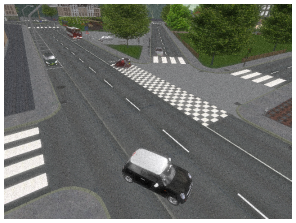


Figure 6 – *vidéo 2*

1. http ://media.ee.ntu.edu.tw/Archer_contest

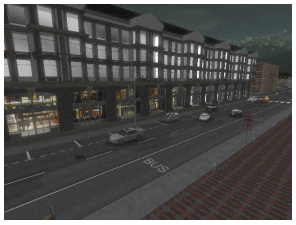


Figure 7 – *vidéo 3*

Figure 8 – *vidéo 4*

Figure 9 – *vidéo 5*

Figure 10 – *vidéo 6*

Figure 11 – *vidéos 7 & 8*

Vidéo	Durée	Images	Taille	img/sec
1	00 : 59, 80	299	320×240	5
2	00 : 28, 96	724	640×480	25
3	00 : 29, 00	725	640×480	25
4	00 : 39, 24	981	640×480	25
5	00 : 23, 04	576	640×480	25
6	00 : 23, 04	576	640×480	25
7	00 : 26, 84	671	640×480	25
8	00 : 26, 48	662	640×480	25

Tableau 1 – *Caractéristiques des séquences de test*

6.2 Paramètres des tests

Étant donné que le paramétrage de l'algorithme lors du test aura toute son importance pour la qualité de l'extraction, chaque séquence est testée avec les mêmes paramètres. On définit des configurations possibles pour chacune des méthodes (voir tableau 2). Et on choisit les paramètres qui obtiennent les meilleurs résultats.

7 Analyse des résultats

7.1 Qualité d'extraction

On calcule la qualité de l'extraction comme décrit précédemment. Cette mesure est faite pour chacune des

	MOG		CB2		VUM	
config.	δ	K	α	β	α	T
1	0,9	0,01	0,4	1,1	0,01	0,2
2	0,8	0,01	0,4	1,3	0,01	0,4
3	0,5	0,01	0,4	1,5	0,01	0,1
4	0,9	0,005	0,55	1,1	0,005	0,2
5	0,8	0,005	0,55	1,3	0,005	0,4
6	0,5	0,005	0,55	1,5	0,005	0,1
7	0,9	0,02	0,7	1,1	0,02	0,2
8	0,8	0,02	0,7	1,3	0,02	0,4
9	0,5	0,02	0,7	1,5	0,02	0,1

Tableau 2 – *Paramètres des tests avec les trois méthodes*

séquences avec chaque méthode et chaque configuration (cf. tableau 2). On garde la configuration qui obtiendra les meilleurs résultats globalement pour toutes les méthodes. Les résultats obtenus avec ce réglage de paramètres optimums pour les différentes vidéos sont rapportés dans le tableau suivant :

	MOG *(config 3)*		CB2 *(config 9)*		VUM *(config 6)*	
vid	Δ	F	Δ	F	Δ	F
1	0,317	0,813	0,152	0,884	0,188	0,914
2	0,517	0,615	0,528	0,643	0,517	0,613
3	0,570	0,724	0,393	0,796	0,501	0,828
4	0,366	0,705	0,315	0,766	0,317	0,786
5	0,364	0,850	0,396	0,854	0,298	0,883
6	0,433	0,816	0,475	0,822	0,357	0,859
7	0,293	0,798	0,195	0,870	0,266	0,728
8	0,244	0,904	0,144	0,928	0,103	0,919
moy	0,388	0,778	0,325	0,820	0,318	0,816

Tableau 3 – *Mesure de Δ et F*

Étant donné que l'idéal est d'avoir Δ=0 et F=1, on remarque que la qualité est globalement plus mauvaise avec la mixture de gaussiennes, mais en revanche, la qualité d'extraction est à peu près identique de façon générale entre le vumètre et le codebook 2 layers. Le vumètre est meilleur sauf dans les deux scènes où le bruit y est très exagéré. Ceci est dû à la largeur d'une classe qui ne peut pas contenir les variations liées au bruit.

7.2 Vitesse d'exécution

Les tests ont été réalisés sur un processeur Intel Xeon® E5520 cadencé à 2,26 GHz, la vitesse d'exécution dépend donc de ces caractéristiques. Pour que ces vitesses soient comparables, tous les algorithmes testés ont été codés en langage C.

Le temps moyen de calcul pour chaque séquence avec les différents paramètres possibles a été mesuré. On peut en déduire la vitesse moyenne de traitement d'une image avec chacune des méthodes (cf. tableau 4).

La vitesse d'exécution est plus rapide avec le vumètre, puis le codebook 2 layers et enfin la mixture de gaussiennes. Mais dans nos conditions de tests, aucune de ces trois méthodes ne permet de faire un traitement en temps réel à 25 img/s sur des images de taille 640×480.

	fps		
Séquence	MOG	CB2	VUM
vidéo 2	5,08	10,31	11,90
vidéo 3	5,21	10,31	13,70
vidéo 4	5,18	11,24	13,51
vidéo 5	5,24	10,64	13,51
vidéo 6	5,21	8,40	12,66
vidéo 7	5,13	9,80	13,33
vidéo 8	5,18	10,64	12,20
moyenne	5,13	9,90	12,66

Tableau 4 – *Comparaison des temps d'exécution*

8 Conclusion

Nous avons montré une comparaison de trois méthodes d'extraction fond/forme. Celles-ci ont été testées avec des scènes simulant les problématiques de circulation routière. On remarque qu'en règle générale, la méthode du vumètre de Goyat *et al.* [3] est meilleure aussi bien sur la qualité d'extraction que sur la vitesse d'exécution. Elle possède néanmoins ses limites avec des bruits extrêmement forts. Le codebook 2 layers [4] donne également de très bons résultats. La mixture de gaussiennes [1] est la moins bonne des trois méthodes testées. Il pourrait être intéressant, par la suite, de pouvoir mettre à disposition ces vidéos avec leurs vérités terrains, afin que chacun puisse comparer ses propres méthodes avec les mêmes critères. Une réflexion est en cours pour réaliser un site de partage de données.

Références

[1] C. Stauffer et W. E. L. Grimson, Adaptative background mixture models for a real-time tracking. Conference on Computer Vision and Patern Recognition, 1999, vol. II, pp. 246–252.

[2] K. Kim, T. H. Chalidabhongse, D. Harwood et L. Davis. Real-time foreground-background segmentation using codebook model. *Real-time Imaging*, 11(3) :167–256, 2005.

[3] Y. Goyat, T. Chateau, L. Malaterre et L. Trassoudaine. Vehicle trajectories evaluation by static video sensors. *9th IEEE International Conference on Intelligent Transportation Systems*, 2006.

[4] M. H. Sigari et M. Fathy. Real-time Background Modeling/Subtraction using Two-Layer Codebook Model. *International MultiConference of Engineers and Computer Scientists*, 2008.

[5] D. Gruyer, C. Royere, N. du Lac, G. Michel et J.-M. Blosseville. SiVIC and RTMaps, interconnected platforms for the conception and the evaluation of driving assistance systems. *World Congress and Exhibition on Intelligent Transport Systems and Services*, 2006.

Moments Disque-Harmoniques basés sur l'échantillonnage Healpix pour une description rapide et robuste des formes 2D

N. Ennahnahi[1] M. Oumsis[1] M. Meknassi[1]

[1] LISQ (Laboratoire d'Informatique-Statistiques-Qualité)

Département Informatique,
Faculté des Sciences Dhar Mahraz, Université USMBA, Fès
B.P. 1796 (Atlas) Fès MAROC

nahnnourd@yahoo.fr; oumsis@yahoo.com; m.meknassi@gmail.com

Résumé

Ce papier présente une amélioration effective de notre descripteur de forme basé-région (Moments Disque-Harmoniques), nommé DHMD. Ce descripteur est basé sur des harmoniques définies sur le disque unité. Nous proposons un nouveau échantillonnage du disque unité en exploitant l'outil HEALPIX (Hierachical Equal Area iso-Latitude Pixelization). L'adoption de HEALPIX est justifiée par ses propriétés géométriques qui assurent le traitement rapide de très larges volumes de données. HEALPIX a été initialement développé pour répondre aux besoins exprimés, en traitement et analyse de données, par les nouvelles missions de recherches relatives au Fond Diffus Cosmologique (ou dans le vocable anglais : Cosmic Microwave Background CMB). La version basique du descripteur DHMD a démontré sa supériorité par rapport à ses compétiteurs en termes de qualité de recherche. L'amélioration que nous proposons concerne principalement le gain en temps d'extraction du vecteur descripteur. La validation des améliorations est élaborée sur la base MPEG-7 CE-1-B. La version améliorée du descripteur DHMD affiche une réelle supériorité Qualité/Temps et encourage son emploi dans les systèmes de recherche supportant de larges bases d'images.

Mots clefs

Recherche d'images par le contenu, Descripteurs de formes, harmoniques sphériques, Polynômes de Legendre, Moments Disque-Harmoniques, HEALPIX.

1 Introduction

La recherche d'images par le contenu (SRIC ou CBIR dans le vocable anglais) est une des plus importantes applications en vision par ordinateur. La demande d'une meilleure qualité délivrée à un délai raisonnable a incité une vaste quantité d'activités scientifiques. Toutes les propositions ont contribué à apporter des améliorations en matière de qualité de représentation d'images et de temps de réponse. La principale composante d'un processus CBIR réside dans la tâche d'extraction d'une représentation fiable permettant une mesure de similarité entre l'image requête et les images de la base stockée dans le système. Les descripteurs de formes peuvent capter des propriétés pertinentes encapsulées dans un objet au sein de l'image. Ce type de descripteurs est dédié pour représenter un objet segmenté dans une image plutôt que l'image toute entière.

Deux catégories de descripteurs de forme existent : ceux basés-contours et ceux basés-régions. On peut citer le descripteur de Fourier FD [1] et les représentations multi-échelles CSS [2,3] comme descripteurs de formes performants basés sur les contours. Ils ne prennent en compte que l'information résidant dans le contour de la forme étudiée. La transformation Angulo-Radiale ART [4], les moments géométriques [5], les moments de Legendre [6], les moments de Zernike [7] et les moments pseudo-Zernike [6,8] sont quelques instances de la classe des descripteurs de formes à base des régions. Ce sont des représentations qui intègrent également l'information située dans la région occupée par la forme. Des études comparatives intéressantes ont été proposées, dans [1,9,10], pour évaluer les diverses représentations de formes 2D.

Récemment, les travaux intéressés par la recherche de modèles tridimensionnels ont proposé des descripteurs de formes 3D robustes possédant un pouvoir discriminatoire élevé. Cette récente activité a engendré une nouvelle voie de recherches qui s'intéresse à la réutilisation et l'adaptation de quelques descripteurs de formes 3D pour la représentation et la recherche des formes 2D. Dans cette optique, quelques auteurs se sont inspirés par la transformation en Harmoniques Sphériques SHT [11,12,13,14].

Dans [15] nous avons proposé un mapping disque-sphère. Ce mapping nous a servi à formuler un nouvel ensemble de fonctions de base orthogonales, nommées les fonctions Disque-Harmoniques DHF. Ces fonctions harmoniques nous ont permis de proposer un descripteur de formes

basé-région qui peut représenter les objets simples ou complexes par un ensemble de moments harmoniques. Ce descripteur est nommé DHMD. Des résultats d'évaluation intéressants ont été présentés dans [15].
Nous exploitons, dans le présent travail, le mécanisme interne de l'outil HEALPIX [16,17] pour réaliser la transformée rapide en harmoniques sur le disque unité. L'outil HEALPIX a été développé, à l'origine, pour répondre aux besoins accrus en traitement et analyse de larges volumes de données par les nouvelles missions en CMB. Cet outil permet une pixellisation sphérique à égale aire et iso-latitude et assure une analyse rapide en harmoniques sphériques même à très grandes résolutions.
Etant donné que la qualité d'un descripteur de forme ne réside pas seulement dans la précision de ses résultats, mais également dans le temps qu'il met pour donner réponse à une requête, nous présentons dans ce papier, une amélioration majeure en termes de temps d'extraction du descripteur DHMD sans perdre en qualité et en pouvoir discriminatoire.
Le reste du papier est structuré comme suit. Les moments Disque-Harmoniques sont rappelés dans la section suivante. Nous introduisons l'outil HEALPIX dans la section 3. Le nouveau paramétrage du disque unitaire est détaillé dans la section 4. L'extraction du descripteur DHMD basé sur HEALPIX est exposée dans la section 5. Les comparaisons de performance et de complexité sont détaillées dans la section 6. La dernière partie est consacrée aux conclusions et perspectives.

2 Les moments Disque-Harmoniques

Les fonctions disque-harmoniques, dénotées $H_l^m(r,\varphi)$, que nous avons proposées dans [15] sont une généralisation des harmoniques sphériques sur le disque unitaire. Elles s'écrivent en coordonnées polaires (r,φ) comme le montre la formule suivante :

$$H_l^m(r,\varphi) = N_l^m P_l^m(r) e^{jm\varphi} \tag{1}$$

Où P_l^m dénote le polynôme de Legendre associé d'ordres *l* et *m*, et N_l^m désigne le facteur de normalisation avec $l \geq 0$ et $-l \leq m \leq l$:

$$N_l^m = \sqrt{\frac{2l+l}{4\pi}\frac{(l-m)!}{(l+m)!}} \tag{2}$$

Les fonctions harmoniques $H_l^m(r,\varphi)$ respectent la formule de symétrie suivante :

$$H_l^{-m}(r,\varphi) = (-1)^m H_l^m(r,\varphi)^* \tag{3}$$

Avec $[H_l^m(r,\varphi)^*]$ est le conjugué de la fonction complexe $H_l^m(r,\varphi)$.

Elles ont également la propriété d'être complètes orthogonales sur le disque unitaire et sont séparables en coordonnées polaires, avec une exponentielle complexe pour la partie angulaire.

Les moments harmoniques C_l^m issus de ces fonctions de base peuvent être extraits par calcul de l'intégrale suivante :

$$C_l^m = \int_0^1 \int_0^{2\pi} H_l^{m^*}(r,\varphi) f(r,\varphi) r\,dr\,d\varphi \tag{4}$$

Où $f(r,\varphi)$ est une fonction image définie sur le disque unitaire.
Dans la version primaire de notre descripteur DHMD, nous avons calculé les moments harmoniques par une intégration numérique simple, où nous nous sommes contentés de sommer sur l'ensemble des points du disque unité. La transformée en Disque-Harmoniques DHT s'écrivait :

$$C_l^m = \sum_r \sum_\varphi f(r,\varphi) H_l^{m^*}(r,\varphi) \tag{5}$$

$0 \leq r \leq 1$ et $0 \leq \varphi \leq 2\pi$

L'inconvénient de cette simplicité vient du fait que dans ce cas nous devons calculer les polynômes de Legendre associés et la partie angulaire complexe pour chaque point du disque unitaire, et ce pour toutes les valeurs souhaitées des ordres *l* et *m*. Pour réduire le temps de calcul, nous avons fait recours au stockage préalable de ces polynômes pré-calculés pour chaque ordre *l* et *m* à chaque point du disque unitaire. Cette solution, bien qu'elle réduise le temps d'extraction du vecteur descripteur, elle s'avère gourmande en matière de ressources mémoire. Une discussion détaillée est présentée dans la section des résultats expérimentaux.
Alors qu'en fait, le calcul des harmoniques sphériques peut être rapide si on fait recours à un échantillonnage sphérique adéquat. Le problème d'échantillonnage sur la sphère n'est pas trivial. La contrainte d'une localisation iso-latitudinale des échantillons est indispensable pour assurer la transformée discrète rapide en harmoniques sphériques FSHT. Les partitions icosaédriques offrent une pixellisation bien uniforme mais non alignée sur des anneaux iso-latitudinaux, ce qui prévient l'application de la transformée rapide en harmoniques sphériques.
Dans la méthode FSHT classique [18,19], le calcul de cette transformée se base sur une grille équiangle.

Cette disposition iso-longitudinale des échantillons souffre de la présence d'une affinité entre les pixels au niveau des régions polaires de la sphère. En plus, Cette partition bien qu'elle offre la possibilité de l'application des théorèmes d'échantillonnage et une quadrature exacte pour l'intégration numérique, ses pixels sont à tailles largement variées. Dans cette partition classique, un ensemble de segments radiaux, partant du centre du disque unitaire, sont générés à pas angulaire régulier. Les points échantillonnés se placent sur les intersections entre ces segments et des cercles concentriques de rayons $(r = \sin\theta)$ en faisant varier l'angle θ à pas fixe également.

Un inconvénient supplémentaire de cette méthode vient du fait que la forme étoilée des segments radiaux peut ne pas prendre en compte quelques détails localisés dans les secteurs compris entre segments consécutifs. Et l'ensemble de points échantillons ne peut être dans ce cas qu'une représentation plus ou moins infidèle à la forme d'origine.

3 HEALPIX

L'outil HEALPIX (Hiearchical Equal Area Iso-Latitude Pixelization) [16,17] est basé sur une première division de la sphère en 12 large pixels qui peuvent être ensuite subdivisés de manière dyadique à la résolution souhaitée, en générant une carte de N_{pix}= $12xN_{side}^2$ pixels à la résolution N_{side} (Figure 1).

Une carte HEALPIX est soigneusement construite afin d'avoir une pixellisation d'égale aire et les pixels se trouvent également alignés sur des anneaux iso-latitudes.

La propriété d'égalité d'aire des pixels permet de surmonter le problème d'affinité des pixels polaires sur la sphère et donner un poids égal aux pixels.

Les positions sur la sphère sont définies par $(z = \cos\theta, \varphi)$ avec $\theta \in [0, \pi]$ est la Co-latitude en radians mesurée à partir du pole Nord et $\varphi \in [0, 2\pi]$ représente la longitude. Pour une résolution Nside, les pixels sont répartis sur ($4xN_{side}$-1) anneaux iso-latitudes, et peuvent être ordonnés par l'index $p \in [0, Npix]$ qui parcourt les anneaux du pôle Nord vers le pôle Sud.

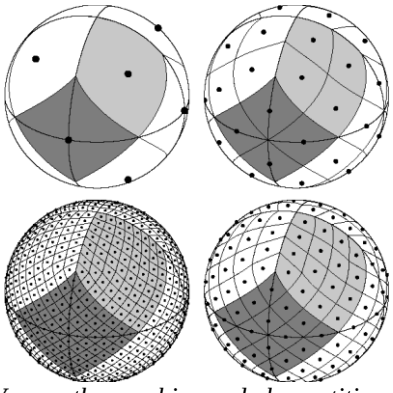


Figure 1- *Vues orthographiques de la partition HEALPIX de la sphère[16] pour Nside=1,2,4 et 8.*

Sur l'hémisphère Nord, les positions des pixels sphériques sont données par les équations suivantes :

Pour les pixels de la calotte polaire nord $(z > \frac{2}{3})$:

avec $p_h = \frac{(p+1)}{2}$, $(1 \le i \le Nside)$ l'index de l'anneau et $(1 \le j \le 4i)$ l'index du pixel au sein de l'anneau i :

$$\begin{aligned} i &= I(\sqrt{p_h - \sqrt{I(p_h)}}) + 1 \\ j &= p + 1 - 2i(i-1) \\ z &= 1 - \frac{i^2}{3N_{side}^2} \\ \varphi &= \frac{\pi}{2i}(j - \frac{s}{2}) \\ s &= 1 \end{aligned} \qquad (6)$$

Pour les pixels de la ceinture équatoriale Nord $(0 \le z \le \frac{2}{3})$:

Avec $p' = p - 2N_{side}(N_{side} - 1)$, $(N_{side} \le i \le 2N_{side})$ et $(1 \le j \le 4N_{side})$:

$$\begin{aligned} i &= I(\frac{p'}{4N_{side}}) + N_{side} \\ j &= (p' \bmod 4N_{side}) + 1 \\ z &= \frac{4}{3} - \frac{2i}{3N_{side}^2} \\ \varphi &= \frac{\pi}{2N_{side}}(j - \frac{s}{2}) \\ s &= (i - N_{side} + 1) \bmod 2 \end{aligned} \qquad (7)$$

Où l'index auxiliaire(*s*) décrit le déphasage longitudinal sur les arcs, et *I(x)* représente le plus grand nombre entier inférieur à *x*.

Les pixels de l'hémisphère Sud peuvent être obtenus par une symétrie miroir par rapport au plan équatorial.

4 Le paramétrage disque-sphère

Soit *D(x, y)* un point du disque unitaire. En coordonnées polaires (r, φ), le point *M(x, y)* vérifie:

$$\begin{aligned} x &= r\cos\varphi \\ y &= r\sin\varphi \end{aligned} \qquad (8)$$

Soit *S(X, Y, Z)* un point de la sphère unitaire. En coordonnées sphériques, le point *S(X, Y, Z)* vérifie:

$$X = \sin\theta\cos\varphi$$
$$Y = \sin\theta\sin\varphi \quad (9)$$
$$Z = \cos\theta = \sqrt{1-(X^2+Y^2)}$$

Pour mettre en correspondance les points *D(x, y)* et *S(X, Y, Z)* nous proposons les relations suivantes :

$$X = x$$
$$Y = y \quad (10)$$
$$Z = \sqrt{1-r^2}$$

Nous déduisons de (8), (9) et (10), le passage entre un point *D(r, φ)* du disque unitaire et son correspondant *S (θ, φ*) dans le système d'échantillonnage HEALPIX:

$$\cos\theta = \sqrt{1-r^2} \text{ et } \varphi_{disque} = \varphi_{sphère} \quad (11)$$

Avec $\varphi_{sphère}$ désigne la co-latitude du pixel sphérique, φ_{disque}dénote la coordonnée angulaire du système polaire.

Cette mise en correspondance entre sphère et disque ne concerne que l'hémisphère Nord, du moment où la forme est entièrement mappée sur cette région et l'autre hémisphère se voit remplie par des valeurs nulles.

5 Extraction du descripteur DHMD

Pour extraire le vecteur descripteur d'une forme binaire, nous prévoyons une étape de prétraitement qui consiste en un centrage et une normalisation de l'échelle. Ceci permet d'englober entièrement la forme par un disque de rayon unité, centré sur le centre de masse de la forme.

Après l'échantillonnage du disque unitaire contenant la forme par le biais des équations (6), (7) et (11), nous effectuons une Transformée Discrète Rapide en Disque-Harmoniques FDHT. Le noyau de cette transformation se compose d'une branche de FFT régulière, suivie d'une branche de transformations discrètes de Legendre :

$$C_l^m = N_l^m \sum_r w P_l^m(r) \sum_\varphi f(r,\varphi) e^{-jm\varphi} \quad (12)$$

Où *w* désigne une quadrature d'intégration proposée par HEALPIX.

La forme du vecteur descripteur est identique à celle retenue dans [15], et consiste en les amplitudes normalisées de ces coefficients.

Le nombre des coefficients composant le vecteur descripteur est 66 pour un ordre maximal *l=10*. Nous dénotons, dans tout ce qui suit, la version améliorée du descripteur DHMD par l'abréviation DHMHD : DHMD-basé sur HEALPIX).

6 Résultats expérimentaux

Une comparaison détaillée des performances qualité/temps est proposée dans [15] et concerne sept autres descripteurs de formes assez connus dans la littérature. Une supériorité satisfaisante de notre descripteur DHMD a été validée lors de ladite comparaison et l'invariance rotationnelle a été également vérifiée pour le descripteur DHMD.

La partie B de la base de formes MPEG-7-CE-1 est retenue pour l'évaluation des performances des descripteurs. Cette base abrite 1400 images distribuées sur 70 classes à égal effectif [20]. Chaque image de la base contient une seule forme (Figure 2).

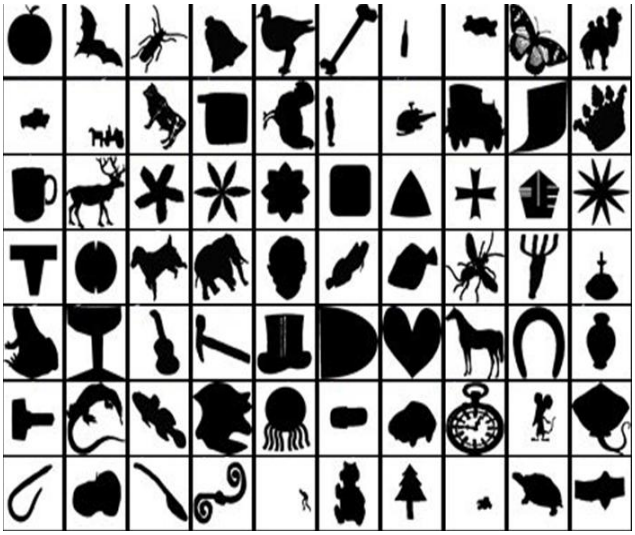


Figure 2- *Des exemples de la base de test MPEG7 CE 1 Partie B. Cette base contient 70 classes rassemblant chacune 20 formes similaires.*

Nous nous contentons dans le présent papier de présenter la nouvelle version du descripteur DHMHD en comparaison avec l'ancienne variante DHMD.

Les courbes Rappel-Précision constituent l'instrument de mesure de la qualité des descripteurs comparés. Le nouveau descripteur DHMHD affiche la même performance que l'ancienne version du descripteur (Figure. 3).

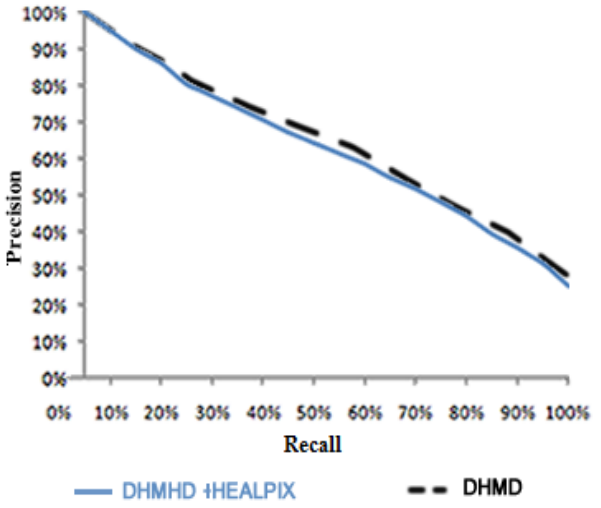


Figure 3- *Les courbes Rappel-Précision de DHMD et DHMHD basé sur HEALPIX.*

Nous présentons par la suite une analyse de complexité pour mettre en valeur le gain perçu en mémoire et en temps de calcul des polynômes de Legendre.
Dans un disque de diamètre 65 pixels on y compte 3209 pixels approximativement. Pour chacun de ces points on calcule 66 moments C_{lm} jusqu'à l'ordre $l=10$. Nous avons dans ce cas 211794 valeurs réelles à calculer. Alors qu'avec HEALPIX, si $N_{side}=32$ est la résolution de base, on aura 12288 pixels au total et le nombre total des anneaux iso-latitudinaux vaut la moitié de ($4N_{side}$ -1), soit 127 anneaux sur la sphère. Il en résulte donc que le nombre total de polynômes de Legendre à calculer ne dépasse pas 8448 polynômes au total. Le rapport de gain en mémoire et en complexité est remarquablement persuasif.
Gorski et al. proposent dans [16] une comparaison de complexité, sous forme graphique, entre différents types de partitions et rendent compte la convenance de HELPAIX pour le calcul rapide des harmoniques sphériques. Ils soulignent également que l'extra temps de calcul dissipé pour la génération non-optimale des fonctions de Legendre associées, résulte typiquement en une complexité de l'ordre de $O(N_{pix}^2)$, alors que cette complexité se réduit à l'ordre $O(\sqrt[3]{N_{pix}})$ si on opte pour la pixellisation HEALPIX.
Cette amélioration est d'origine purement géométrique : les fonctions de Legendre membres des harmoniques sphériques, qui doivent absolument être générées via des itérations et des récurrences, ne sont évaluées qu'une seule fois pour chaque anneau.
Nous soulignons également une autre amélioration pesante qui se traduit par l'introduction de la transformée rapide de Fourier FFT lors du calcul des moments Disque-Harmoniques DHM avec la formule (12).
Pour illustrer la discussion précédente, nous présentons par la suite la figure qui regroupe les temps de traitement pour l'extraction des vecteurs descripteurs de DHMD et DHMHD sans pré-calcul des polynômes de Legendre.
Les temps de (Figure.4) sont exprimés en secondes et représentent les temps moyennes respectives calculées sur la totalité de la base de formes MPEG7-CE-1-B (1400 formes) par une machine CELERON-D 1.7GHz à 256 Mo.
Le temps consacré à la phase de prétraitement, qui comprend un centrage et une mise à l'échelle de la forme, ne dépend guerre de la méthode d'échantillonnage mais surtout de la base de formes étudiée. Ceci explique que ce temps reste inchangé pour les deux méthodes comparées. L'apport de la nouvelle approche basée sur le paramétrage du disque unité par le mécanisme interne d'HEALPIX est remarquable puisque la phase de calcul des moments harmoniques est à peu prés 12 fois plus rapide qu'avec la méthode basique. Si on compare les temps moyens du processus d'extraction des vecteurs descripteurs DHMD et DHMHD, tout entier, on peut constater que le gain est de l'ordre de 6 fois en faveur de la nouvelle approche proposée dans ce papier.

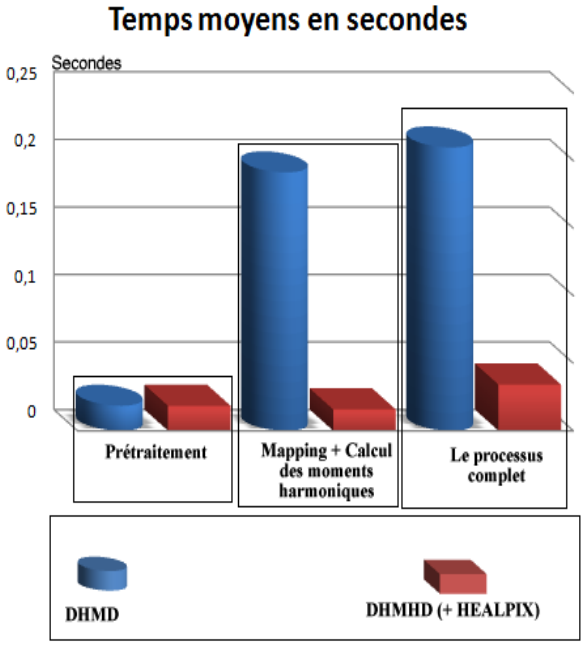


Figure 4- *Les temps moyens des différentes phases du processus d'extraction des vecteurs descripteurs DHMD et DHMHD basé sur HEALPIX.*

7 Conclusion et perspectives

Nous avons apporté une amélioration majeure en temps d'extraction du vecteur descripteur DHMD inspiré par les harmoniques sphériques. Cette contribution est due à l'introduction de l'outil HEALPIX qui connaît déjà ses succès avec les missions spatiales exigeant le traitement rapide de quantités de données volumineuses (ex. La mission Planck de l'agence spatiale Européenne ESA). Le caractère iso-latitudinal des pixels HEALPIX nous a permis d'exploiter la Transformée de Fourier Rapide FFT et un calcul optimisé des polynômes de Legendre associés. La qualité des résultats, obtenus par le nouveau descripteur DHMHD, reste satisfaisante et se traduit par une courbe rappel-précision étroitement voisine à la version basique.

Nous prévoyons dans le futur proche, étudier l'effet de la résolution $N_{side,}$ de la carte HEALPIX, sur le temps de traitement et la qualité des résultats.

Références

[1] Zhang D.S et Lu G. A comparative study of Fourier descriptors for shape représentation and retrieval. Dans *Fifth Asian Conference on Computer Vision (ACCV02)*, pages 646-651, 2002.

[2] Abbasi S. et Mokhtarian F. et Kittler J. Curvature scale space image in shape similarity retrieval. *Multimedia Systems*. 7(6) : 467-476, 1999.

[3] Abbasi S. et Mokhtarian F. et Kittler J.. Enhancing CSS-based shape retrieval for objects with shallow concavities. *Image and vision computing*, 18(3): 199-211, 2000.

[4] Whoi-Yul Kim et Young-Sung Kim. A new region-based shape descriptor. Dans *ISO/IEC MPEG99/M5472*, Maui, Décembre 1999.

[5] Hu M. Visual pattern recognition by moment invariants. *IRE trans. Inf. Theory*, IT-8 : 179-187, 1962.

[6] Teh C.-H. et Chin R.T. On image analysis by the methods of moments. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 10(4) : 496-513, 1988.

[7] Chong C-W. et Raveendran P et Mukunda R. A comparative analysis of algorithms for fast computation of Zernike moment. *Pattern Recognition*, 3 : 890-901, 2003.

[8] Haddadnia I. et Ahmadi M. et faez K. An efficient feature extraction method with pseudo-Zernike moment in RBF neural network-based human face recognition system. *Eurasip Journal on Applied Signal processing*, 2003(9) : 731-742, 2003.

[9] Zhang D.S. et Lu G. A comparison of shape rtrieval using Fourier descriptors and short-time Fourier descriptors. Dans *Second IEEE Pacific-Rim Conference on Multimedia PCM01* , pages 855-860, 2001.

[10] Zhang D.S. et Lu G. Evaluation of MPEG-7 shape descriptors against other shape descriptors, *Multimedia Systems*, 9(1) : 15-30, 2003.

[11] Thomas Funkhouser et Patrick Min et Misha Kazhdan et Joyce Chen et Alex Halderman et David Dobkin et David Jacobs. A search engine for 3D models. *ACM Transactions on Graphics*, 22(1): 83-105, 2003.

[12] Pu J.T. et Karthik R. On visual similarity based on 2D drawing retrieval. *Computer Aided Design*. 38(3) : 249-259, 2006.

[13] Sajjanhar A. et Lu G. et Zhang D. et Hou J. et Chen Y. Spherical Harmonics and Distance Transform for Image Representation and Retrieval. *Lecture Notes in Computer Science (IDEAL09)*, 5788: 309-316, 2009.

[14] Sajjanhart A. et Lu G. et Zhang D.S. Spherical harmonics descriptor for 2D-image retrieval. Dans *IEEE International Conference on Multimedia and Expo ICME'05*, pages 105-108, 2005.

[15] Ennahnahi N. et Oumsis et M. Bouhouch A. et Meknassi M. Fast shape description based on a set of moments defined on the unit disc and inspired by three-dimensional spherical harmonics. *Image processing IET, 4(2)* :120-131, April 2010.

[16] Górski K.M. et Eric Hivon et Banday A.J. et Wandelt B.D. et Hansen F.K. et Reinecke M. et Bartelmann M. HEALPIX- a frame work for high resolution discretization an fast analysis of data distributed on the sphere. *The Astrophysical Journal*, 622(2) : 759-771, Avril 2005.

[17] HEALPIX. http://healpix.jpl.nasa.gov/

[18] Healy Jr. et Rockmore D. et Kostelec P. et Moore S. FFTs for the 2-sphere-Improvements and Variations. *The journal of Fourier Analysis and Applications*, 9(4) : 341-385, 2003.

[19] http://www.cs.dartmouth.edu/~geelong/sphere/

[20] Latecki L.J. et Lakämper R. et Eckhardt U. Shape Descriptors for Non-rigid Shapes with a Single Closed Contour. Dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 424-429, 2000.

Autocorrélation basée sur les transformations pour la détection de régions affines covariantes

S. Khoualed A. Bartoli T. Collins

Clermont Université, France

{khoualed.samir, adrien.bartoli, toby.collins}@gmail.com

Résumé

Dans cette contribution nous introduisons une nouvelle approche pour détecter les éléments saillants dans des images. Elle est basée sur ce que nous avons appelé *warp-based autocorrelation*. Nous combinons cette méthode avec une paramétrisation à base de points de contrôle, afin de détecter les régions covariantes. L'algorithme commence par une recherche exhaustive sur un ensemble de paramètres affines. Afin de rejeter les régions instables, un seuillage est appliqué. Les résultats obtenus montrent que notre approche surpasse les autres méthodes en terme de répétabilité et de précision.

1 Introduction

La détection des éléments saillants (c.-à-d., caractéristiques) [1] est une étape préliminaire pour beaucoup de tâches en vision par ordinateur. Par conséquent, différents détecteurs d'éléments saillants ont été proposés, qui diffèrent largement par le type d'éléments détectés. Ils peuvent être divisés en différentes catégories. Les détecteurs de régions covariantes surpassent les autres détecteurs en terme de répétabilité. Parmi les détecteurs existants, les détecteurs Harris-Affine, Hessian-Affine [1, 2] et MSER [3] sont les meilleurs en terme de répétabilité. Harris-Affine et Hessian-Affine sont deux méthodes similaires basées sur le principe du détecteur de coin de Harris-Stephen [4]. Les deux approches sont basées sur l'adaptation de forme affine (*affine shape adaptation*). Leurs algorithmes peuvent être résumés en deux étapes : 1- identification de la localisation spatiale initiale des régions à l'aide de détecteur de Scale-Invariant Harris-Laplace (pour Harris-Affine) et à base de la matrice Hessienne pour Hessian-Affine [5], 2- pour chaque point initial, une normalisation itérative à base de l'adaptation de forme affine est effectuée afin de rendre la région invariante par rapport à la transformation affine. La méthode MSER (Maximally Stable Extremal Regions) est basée sur une approche différente. Elle repose sur l'extraction de composantes connexes en utilisant un seuillage de l'image. Les régions extrémales de MSER, sont constituées de tous pixels dont l'intensité est plus (ou moins) élevée que celle des pixels de son extérieur. Une autre méthode exploitant les arêtes présentes dans l'image a été proposée dans [6, 7]. Le concept d'EBR est que les arêtes sont des éléments caractéristiques stables et qu'elles peuvent être facilement détectées malgré de nombreuses transformations (échelle, angle de vue, luminosité, etc). Les auteurs de [6, 7] ont également proposé une autre approche, appelée IBR. L'algorithme IBR commence par la détection des extrémas de l'intensité à plusieurs échelles. Le voisinage de chaque extrémum est ensuite inspecté de manière radicale afin de marquer le contour de la région. Les régions avec des formes arbitraires sont ensuite remplacées par des ellipses. Le problème avec les détecteurs cités ci-dessus est constaté par les auteurs de [2]. Ils ont effectué une comparaison complète entre ces méthodes. Cette comparaison a montrée qu'il n'existe pas de détecteur qui surpasse absolument les autres détecteurs pour tous types de déformations. Par exemple, MSER surpasse Harris-Affine et Hessian-Affine par rapport aux changements d'angle de vue. En revanche , pour le changement d'échelle, il vient en second rang après le détecteur de Hessian-Affine. Hessian-Affine et Harris-Affine ont les scores les plus élevés de répétabilités par rapport au flou. Toutefois, MSER est le plus sensible à ce type de changement. Par conséquent, il semble utile de développer un détecteur de régions qui garantie explicitement une bonne robustesse par rapport à tous les types de déformation (échelle, angle de vue, illumination, etc). Ceci est l'objectif de notre contribution, qui généralise l'approche de Harris et Stephen [4] pour inclure non seulement la détection des points d'intérêts mais aussi la détection des régions, pour aboutir à une meilleure résistance aux différents types de déformations.

Notre approche est basée sur la détection des transformations locales qui ont des maximas locaux sur la fonction de saillance. La fonction de saillance est l'optimisation de ce que nous avons nommé le *warp-based autocorrelation* [2]. Les régions d'images qui maximisent la fonction de saillance sont qualifiées de covariantes.

Contenu. §2 présente un aperçu sur la méthode de Harris et

1. Il n'existe pas une définition standard de l'élément saillant. Souvent il dépend du contexte de problème ou du type d'application. Cependant, la définition générale est qu'un élément saillant est une partie saillante d'une image : les arêtes, coins/points d'intérêts, gouttes/régions d'intérêts, etc.

2. *warp-based autocorrelation* ou **autocorrélation basée sur les transformations pour la détection de régions affines covariantes**

Stephens qui est la base de notre approche. §3 décrit notre approche pour la détection de régions affines covariantes. §4 présente des résultats expérimentaux.

Notation. $g(.;\boldsymbol{\sigma})$ désigne une Gaussienne centrée de covariance $\boldsymbol{\sigma}$. Dans la suite, $\boldsymbol{\sigma}_0$ désigne une matrice de covariance isotrope. Le vecteur $\boldsymbol{\gamma} = (s_C\ h_1\ h_2)^\top$ est le vecteur de paramètres affines : l'échelle s_C, le cisaillement horizontale h_1 et le cisaillement verticale h_2 (le paramètre de rotation est omis car il est indéterminée par la transformation d'une ellipse vers un cercle). Les notations I_0, I et $\tilde{I}$ désignent l'image d'entrée (de référence), l'image transformée de la référence par $\boldsymbol{A}(\boldsymbol{\gamma})$ et la version filtrée de l'image transformée avec la Gaussienne g respectivement. La matrice affine $\boldsymbol{A}(\boldsymbol{\gamma})$ est définie comme la transformation locale d'une ellipse définie dans l'image de référence I_0 vers un cercle fixe C définit dans l'image transformée I. Sauf indication contraire, nous supposons que l'image est continue et nous ignorons les effets d'échantillonnage.

2 Travaux antérieurs

La méthode de Harris et Stephens [4] est une approche communément utilisée pour détecter les points d'intérêts dans l'image. Elle sert de base à beaucoup d'autres approches. Plusieurs de ces méthodes se basent sur la maximisation de la fonction d'autocorrélation locale proposée par Harris et Stephens.

Le principe de la méthode de Harris et Stephen est basé sur la maximisation de l'autocorrélation locale utilisant un décalage local à 2-dimensions d'une région (patch) pondérée isotropiquement. Ensuite, une recherche exhaustive à 2-dimensions est lancée sur la position afin de sélectionner les maximas locaux de l'autocorrélation locale. Considérons une région circulaire C dans l'image I, centrée sur le point $\boldsymbol{q} \in \mathbb{R}^2$. La variation en intensité de pixel correspondante à une variation élémentaire $\delta\boldsymbol{q}$ autour de $\boldsymbol{q}$ dans C peut être écrite comme :

$$c(\boldsymbol{q};\delta\boldsymbol{q}) = \int_C g(\boldsymbol{\tau};\boldsymbol{\sigma}_0)\, \|(I(\boldsymbol{\tau}+\boldsymbol{q}+\delta\boldsymbol{q}) - I(\boldsymbol{\tau}+\boldsymbol{q}))\|^2\, d\boldsymbol{\tau}, \quad (1)$$

où $\boldsymbol{\tau}$ est la coordonnée locale du pixel appartenant à la région circulaire centrée C. $g(.;\boldsymbol{\sigma}_0)$ définit une gaussienne circulaire isotrope dans C, centrée sur $\boldsymbol{q}$ et avec une covariance isotrope $\boldsymbol{\sigma}_0$. Pour la variation élémentaire $\delta\boldsymbol{q}$ autour de $\boldsymbol{q}$, $I(\boldsymbol{\tau}+\boldsymbol{q}+\delta\boldsymbol{q})$ peut être approché avec une approximation de Gauss-Newton (c.-à-d., expansion de Taylor d'ordre 1) comme suit :

$$I(\boldsymbol{\tau}+\boldsymbol{q}+\delta\boldsymbol{q}) \approx I(\boldsymbol{\tau}+\boldsymbol{q}) + \left(\frac{\partial I}{\partial \boldsymbol{q}}(\boldsymbol{\tau}+\boldsymbol{q},)\right)^\top \delta\boldsymbol{q}. \quad (2)$$

Ceci mène à l'approximation :

$$c(\boldsymbol{q};\delta\boldsymbol{q}) \approx \delta\boldsymbol{q}^\top \left(\int_C g(\boldsymbol{\tau};\boldsymbol{\sigma}_0)\nabla I(\boldsymbol{\tau}+\boldsymbol{q})\nabla I^\top(\boldsymbol{\tau}+\boldsymbol{q})d\boldsymbol{\tau}\right)\delta\boldsymbol{q}, \quad (3)$$

ce qui donne en notation matricielle :

$$c(\boldsymbol{q};\delta\boldsymbol{q}) \approx \delta\boldsymbol{q}^\top \boldsymbol{M}(\boldsymbol{q})\delta\boldsymbol{q}, \quad (4)$$

où $\boldsymbol{M}(\boldsymbol{q})$ est la matrice d'autocorrélation :

$$\boldsymbol{M}(\boldsymbol{q}) = \int_C g(\boldsymbol{\tau};\boldsymbol{\sigma}_0)\nabla I(\boldsymbol{\tau}+\boldsymbol{q})\nabla I^\top(\boldsymbol{\tau}+\boldsymbol{q})d\boldsymbol{\tau}. \quad (5)$$

La fonction de saillance est définie par :

$$s : \mathbb{R}^2 \to \mathbb{R};\ \ s(\boldsymbol{q}) = \min_{\delta\boldsymbol{q}\in\mathbb{R}^2} c(\boldsymbol{q};\delta\boldsymbol{q}).$$

En se basant sur l'analyse spectrale, le minimum de $c(\boldsymbol{q},\delta\boldsymbol{q})$ sur $\delta\boldsymbol{q}$ sous la contrainte $\delta\boldsymbol{q}^\top\delta\boldsymbol{q} = 1$ est donné par la plus petite valeur propre de $\boldsymbol{M}$: $s(\boldsymbol{q}) = \lambda_{min}(\boldsymbol{M}(\boldsymbol{q}))$. Enfin, les points saillants (c.-à-d., les points d'intérêts) $\tilde{\boldsymbol{q}}$ sont sélectionnés avec :

$$\tilde{\boldsymbol{q}} = \arg\operatorname*{local\,max}_{\boldsymbol{q}\in I} s(\boldsymbol{q}). \quad (7)$$

3 Détection basée sur les transformations

Notre approche est basée sur ce que nous nommons le *warp-based autocorrelation*, ainsi définit :

$$c_I(\boldsymbol{q},\boldsymbol{\gamma};\delta\boldsymbol{q}) = \int_C g(\boldsymbol{\tau};\boldsymbol{\sigma}_0)\left\|\tilde{I}(\boldsymbol{q}+\boldsymbol{\tau}+\boldsymbol{A}(\boldsymbol{\gamma})^{-1}\delta\boldsymbol{q}) - \tilde{I}(\boldsymbol{q}+\boldsymbol{\tau})\right\|^2 d\boldsymbol{\tau}. \quad (8)$$

Nous rappelons que la matrice affine $\boldsymbol{A}(\boldsymbol{\gamma})$ est définie comme la transformation locale d'une ellipse définie dans l'image de référence I_0 vers un cercle fixe C définit dans l'image transformée I et centrée en $\boldsymbol{q}$. $\tilde{I}$ est le résultat de la convolution de I avec $g(.;\boldsymbol{\sigma}_0)$:

$$\tilde{I}(\boldsymbol{q}+\boldsymbol{\tau}) = \int_C g(\boldsymbol{\alpha};\boldsymbol{\sigma}_0)I(\boldsymbol{q}+\boldsymbol{\tau}+\boldsymbol{\alpha})d\boldsymbol{\alpha}. \quad (9)$$

Dans sa définition générale proposée dans (8), la valeur de $c(\boldsymbol{q},\boldsymbol{\gamma};\delta\boldsymbol{q})$ est la variation de l'intensité de l'image au point $\boldsymbol{q}$ vis-à-vis d'une variation locale $\delta\boldsymbol{q}$ de la région circulaire isotrope définie par $g(.;\boldsymbol{\sigma}_0)$. L'image référence I_0 est reliée au point $\boldsymbol{q}$ de l'image transformée par le matrice affine $\boldsymbol{A}$ (dans le reste de l'article $\boldsymbol{A} \equiv \boldsymbol{A}(\boldsymbol{\gamma})$) :

$$I(\boldsymbol{q}+\boldsymbol{\tau}+\boldsymbol{\alpha}) = I_0(\boldsymbol{A}^{-1}\boldsymbol{q} + \boldsymbol{A}^{-1}\boldsymbol{\tau} + \boldsymbol{A}^{-1}\boldsymbol{\alpha}). \quad (10)$$

La fonction optimisée de $c_I(\boldsymbol{q},\boldsymbol{\gamma};\delta\boldsymbol{q})$ sur $\delta\boldsymbol{q}$ mesure la saillance au point $(\boldsymbol{q}\ \boldsymbol{\gamma})^\top$, notée $s_I(\boldsymbol{q},\boldsymbol{\gamma})$:

$$s_I(\boldsymbol{q},\boldsymbol{\gamma}) = \min_{\delta\boldsymbol{q}\in\mathbb{R}^2} c_I(\boldsymbol{q},\boldsymbol{\gamma};\delta\boldsymbol{q}). \quad (11)$$

L'idée sous-jacente de notre approche *warp-based autocorrelation*, est que si l'image I_0 est transformée localement au point $\tilde{\boldsymbol{q}}_0$ par une transformation affine $\tilde{\boldsymbol{A}}$, et si son image transformée I répond avec des maximas locaux au point $\tilde{\boldsymbol{q}} = \boldsymbol{A}\tilde{\boldsymbol{q}}_0$ sur la fonction de saillance (11), alors le vecteur $(\tilde{\boldsymbol{q}}\ \tilde{\boldsymbol{\gamma}})^\top$ correspond au point spatiale $\tilde{\boldsymbol{q}}_0$ et à la transformation $\tilde{\boldsymbol{A}}$ est qualifiée *saillant*. Cependant, la région circulaire C, centrée au point $\tilde{\boldsymbol{q}} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{q}}_0$ et définie dans I, est transformée en région elliptique E, centrée en $\tilde{\boldsymbol{q}}_0$ et définie dans I_0. Cette région est qualifiée de *covariante* :

$$E = \left\{\boldsymbol{q}_0 \in I_0 : (\boldsymbol{q}_0 - \tilde{\boldsymbol{q}}_0)^\top \tilde{\boldsymbol{A}}^\top \tilde{\boldsymbol{A}}\,(\boldsymbol{q}_0 - \tilde{\boldsymbol{q}}_0) \le r^2\right\}. \quad (12)$$

Ici, r est une constante correspondant au diamètre de C.
Comme l'image donnée est I_0, ceci nous amène à dériver l'expression de la fonction de saillance $s(\boldsymbol{q}_0, \boldsymbol{\gamma})$ correspondante à l'image I_0, c.-à-d., écrire $c(.,.;.)$ en fonction de I_0, $\boldsymbol{q}_0$, $\boldsymbol{\gamma}$ et $\boldsymbol{\delta q}_0$.
Commençons à partir des équations (9) et (10), on peut écrire :

$$\tilde{I}(\boldsymbol{q}+\boldsymbol{\tau}) = \int_C g(\boldsymbol{\alpha};\boldsymbol{\sigma}_0) I_0(\boldsymbol{A}^{-1}\boldsymbol{q} + \boldsymbol{A}^{-1}\boldsymbol{\tau} + \boldsymbol{A}^{-1}\boldsymbol{\alpha}) d\boldsymbol{\alpha}. \quad (13)$$

La définition de la région circulaire c'est :

$$C = \left\{ \boldsymbol{p} \in I : (\boldsymbol{p}-\boldsymbol{q})^\top (\boldsymbol{p}-\boldsymbol{q}) <= r^2 \right\}. \quad (14)$$

Le changement de variables $\boldsymbol{\alpha}_0 = \boldsymbol{A}^{-1}\boldsymbol{\alpha}$ dans (13) conduit à :

$$\tilde{I}(\boldsymbol{q}+\boldsymbol{\tau}) = \int_E g(\boldsymbol{\alpha}_0;\boldsymbol{\sigma}) I_0(\boldsymbol{q}_0 + \boldsymbol{A}^{-1}\boldsymbol{\tau} + \boldsymbol{\alpha}_0) d\boldsymbol{\alpha}_0. \quad (15)$$

Avec $\boldsymbol{q}_0 = \boldsymbol{A}^{-1}\boldsymbol{q}$, $\boldsymbol{\sigma} = \boldsymbol{A}^{-1}\boldsymbol{\sigma}_0\boldsymbol{A}^{-\top}$ est une covariance Gaussienne anisotrope et E une région elliptique définie dans I_0 :

$$E = \left\{ \boldsymbol{p}_0 \in I_0 : (\boldsymbol{p}_0-\boldsymbol{q}_0)^\top \boldsymbol{A}^\top \boldsymbol{A} (\boldsymbol{p}_0-\boldsymbol{q}_0) <= r^2 \right\}. \quad (16)$$

Le résultat obtenu dans (15) montre que la convolution d'une image transformée avec une Gaussienne isotrope et définie dans une région circulaire, est équivalente, à la convolution de l'image origine avec une Gaussienne transformée avec la même transformation, c.-à-d., une Gaussienne anisotrope définie sur une région elliptique.
Approximation de Gauss-Newton. L'approximation de Gauss-Newton, permet d'écrire :

$$\tilde{I}(\boldsymbol{q}+\boldsymbol{\tau}+\boldsymbol{A}^{-1}\boldsymbol{\delta q}) - \tilde{I}(\boldsymbol{q}+\boldsymbol{\tau}) = \left(\boldsymbol{\nabla}\tilde{I}(\boldsymbol{q}+\boldsymbol{\tau})\right)^\top \boldsymbol{A}^{-1}\boldsymbol{\delta q}. \quad (17)$$

Ici $\boldsymbol{\nabla}\tilde{I} = \frac{\partial \tilde{I}}{\partial \boldsymbol{q}}$. En accord avec la propriété de commutativité de la convolution [3], la dérivée peut être obtenue d'une façon équivalente, par la convolution de l'image originale avec la dérivée de la Gaussienne. Cela, conduit à :

$$\boldsymbol{\nabla}\tilde{I}(\boldsymbol{q}+\boldsymbol{\tau}) = \boldsymbol{A}^{-1}(\boldsymbol{\nabla} g(.;\boldsymbol{\sigma}) * I_0)(\boldsymbol{q}_0 + \boldsymbol{A}^{-1}\boldsymbol{\tau}). \quad (18)$$

Ici $*$ est l'opérateur de convolution. $\boldsymbol{\nabla} g(.;\boldsymbol{\sigma})$ est la dérivée de $g(.;\boldsymbol{\sigma})$. La théorie liée à l'espace d'échelle et la dérivée normalisée de la Gaussienne, permet d'écrire :

$$\boldsymbol{\nabla}_{\zeta-norm} g(.;\boldsymbol{\sigma}) = \boldsymbol{\sigma}^\zeta \boldsymbol{\nabla} g(.;\boldsymbol{\sigma}). \quad (19)$$

Ici $\nabla_{\zeta-norm}$ est l'opérateur de la dérivée normalisée. $\zeta \in [0,1]$ est un paramètre lié à la dimensionnalité des éléments saillants dans l'image. Ainsi, la dérivée normalisée de l'image $\tilde{I}$ est donnée par :

$$\boldsymbol{\nabla}\tilde{I}(\boldsymbol{q}+\boldsymbol{\tau}) = \boldsymbol{A}^{-1}\boldsymbol{\sigma}^\zeta(\boldsymbol{\nabla} g(.;\boldsymbol{\sigma}) * I_0)(\boldsymbol{q}_0 + \boldsymbol{A}^{-1}\boldsymbol{\tau}). \quad (20)$$

3. $\boldsymbol{\nabla}(f_a * f_b) = (\boldsymbol{\nabla} f_a) * f_b = f_a * (\boldsymbol{\nabla} f_b)$.

La substitution de l'expression de $\boldsymbol{\nabla}\tilde{I}(\boldsymbol{q}+\boldsymbol{\tau})$ obtenue à partir de (20) dans (8), conduit à écrire : $c_I(\boldsymbol{q},\boldsymbol{\gamma};\boldsymbol{\delta q}) =$

$$\int_C g(\boldsymbol{\tau};\sigma_0) \left\| \left(\boldsymbol{A}^{-1}\boldsymbol{\sigma}^\zeta(\boldsymbol{\nabla} g(.;\boldsymbol{\sigma}) * I_0)(\boldsymbol{q}_0 + \boldsymbol{A}^{-1}\boldsymbol{\tau})\right)^\top \boldsymbol{A}^{-1}\boldsymbol{\delta q} \right\|^2 d\boldsymbol{\tau}. \quad (21)$$

En substituant $\boldsymbol{\tau}$ dans (21) avec $\boldsymbol{\tau}_0 = \boldsymbol{A}^{-1}\boldsymbol{\tau}$, on obtient que : $c_I(\boldsymbol{q},\boldsymbol{\gamma};\boldsymbol{\delta q}) =$

$$\int_E g(\boldsymbol{\tau}_0;\boldsymbol{\sigma}) \left\| \left(\boldsymbol{A}^{-1}\boldsymbol{\sigma}^\zeta((\boldsymbol{\nabla} g(.;\boldsymbol{\sigma})) * I_0)(\boldsymbol{q}_0 + \boldsymbol{\tau}_0)\right)^\top \boldsymbol{A}^{-1}\boldsymbol{\delta q} \right\|^2 d\boldsymbol{\tau}_0. \quad (22)$$

Sachant que $\boldsymbol{\delta q}_0 = \boldsymbol{A}^{-1}\boldsymbol{\delta q}$, la dernière intégrale (22) peut être réarrangée comme suit :

$$c_I(\boldsymbol{q},\boldsymbol{\gamma};\boldsymbol{\delta q}) \approx \boldsymbol{\delta q}_0{}^\top \boldsymbol{A}^{-\top} \boldsymbol{M}(\boldsymbol{q}_0,\boldsymbol{\gamma}) \boldsymbol{A}^{-1} \boldsymbol{\delta q}_0, \quad (23)$$

avec $\boldsymbol{M}(\boldsymbol{q}_0,\boldsymbol{\gamma}) =$

$$\boldsymbol{\sigma}^\zeta \left(g(.;\boldsymbol{\sigma}) * \left((\boldsymbol{\nabla} g(.;\boldsymbol{\sigma}) * I_0) (\boldsymbol{\nabla} g(.;\boldsymbol{\sigma}) * I_0)^\top \right) \right) (\boldsymbol{q}_0) \boldsymbol{\sigma}^{\zeta\top}. \quad (24)$$

Par conséquent, la définition du *warp-based autocorrelation* en fonction de l'image originale I_0, est donnée par :

$$c(\boldsymbol{q}_0,\boldsymbol{\gamma};\boldsymbol{\delta q}_0) \approx \boldsymbol{\delta q}_0{}^\top \boldsymbol{A}^{-\top} \boldsymbol{M}(\boldsymbol{q}_0,\boldsymbol{\gamma}) \boldsymbol{A}^{-1} \boldsymbol{\delta q}_0. \quad (25)$$

En considérant que le principe de détection correspondant est :

$$\begin{aligned} (\tilde{\boldsymbol{q}}_0, \tilde{\boldsymbol{\gamma}}) &= \underset{(\boldsymbol{q}_0,\boldsymbol{\gamma}) \in I_0 \times \mathbb{R}^3}{\arg\operatorname{local\,max}} s(\boldsymbol{q}_0,\boldsymbol{\gamma}), \\ \text{avec :} \quad s(\boldsymbol{q}_0,\boldsymbol{\gamma}) &= \min_{\boldsymbol{\delta}_{\boldsymbol{q}0} \in \mathbb{R}^2} c(\boldsymbol{q}_0,\boldsymbol{\gamma};\boldsymbol{\delta q}_0), \end{aligned} \quad (26)$$

l'expression de $c(\boldsymbol{q}_0,\boldsymbol{\gamma};\boldsymbol{\delta q}_0)$ est la forme quadratique correspondant à la matrice : $\boldsymbol{A}^{-\top}\boldsymbol{M}(\boldsymbol{q}_0,\boldsymbol{\gamma})\boldsymbol{A}^{-1}$. Comme dans le cas de Harris et Stephen's, la fonction de saillance est :

$$s(\boldsymbol{q}_0,\boldsymbol{\gamma}) = \lambda_{min}\left(\boldsymbol{A}^{-\top}\boldsymbol{M}(\boldsymbol{q}_0,\boldsymbol{\gamma})\boldsymbol{A}^{-1}\right). \quad (27)$$

En fait, le critère approprié en terme de répétabilité, est établi ainsi :

$$s(\boldsymbol{q}_0,\boldsymbol{\gamma}) = \frac{\det(\boldsymbol{A}^{-\top}\boldsymbol{M}(\boldsymbol{q}_0,\boldsymbol{\gamma})\boldsymbol{A}^{-1})}{\operatorname{trace}(\boldsymbol{A}^{-\top}\boldsymbol{M}(\boldsymbol{q}_0,\boldsymbol{\gamma})\boldsymbol{A}^{-1})}. \quad (28)$$

Ici, $\det$ et $trace$, désignent le déterminant et la trace respectivement.

3.1 Covariance du warp-based autocorrelation

En considérant que I_0 et I_0' sont deux images liées par la transformation affine $\boldsymbol{w}(\boldsymbol{q},\boldsymbol{\gamma}) = \boldsymbol{B}\boldsymbol{q}+\boldsymbol{t}$ tel que $I_0(\boldsymbol{q}) = I_0'(\boldsymbol{B}\boldsymbol{q}+\boldsymbol{t})$. La transformation affine $\boldsymbol{w}(\boldsymbol{q},\boldsymbol{\gamma})$ transforme la région elliptique E définie dans I_0 (équation (16)) en région elliptique E' dans I_0', centrée en $\boldsymbol{q}_0' = \boldsymbol{B}\boldsymbol{q}_0 + \boldsymbol{t}$ et définie par la matrice affine $\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{B}^{-1}$:

$$E' = \left\{ \boldsymbol{p}_0 \in I_0' : \left(\boldsymbol{p}_0 - \boldsymbol{q}_0'\right)^\top \boldsymbol{A}'^\top \boldsymbol{A}' \left(\boldsymbol{p}_0 - \boldsymbol{q}_0'\right) <= r^2 \right\}. \quad (29)$$

En utilisant les notations $s(\boldsymbol{q}_0, \boldsymbol{A})$ et $c(\boldsymbol{q}_0, \boldsymbol{A}; \boldsymbol{\delta q}_0)$ à la place de $s(\boldsymbol{q}_0, \boldsymbol{\gamma})$ et $c(\boldsymbol{q}_0, \boldsymbol{\gamma}; \boldsymbol{\delta q}_0)$ respectivement, ainsi la condition de covariance est définie par :

$$\begin{aligned}(\tilde{\boldsymbol{q}}_0, \tilde{\boldsymbol{A}}) &= \underset{(\boldsymbol{q}_0, \boldsymbol{A}) \in I_0 \times \mathbb{R}^3}{\arg \text{local max}}\ s(\boldsymbol{q}_0, \boldsymbol{A}) \iff \\ (\tilde{\boldsymbol{q}}'_0, \tilde{\boldsymbol{A}}') &= \underset{(\boldsymbol{q}'_0, \boldsymbol{A}') \in I'_0 \times \mathbb{R}^3}{\arg \text{local max}}\ s'(\boldsymbol{q}'_0, \boldsymbol{A}') \\ \text{tel que : } \tilde{\boldsymbol{q}}'_0 &= \boldsymbol{B}\tilde{\boldsymbol{q}}_0 + \boldsymbol{t}\ ,\ \tilde{\boldsymbol{A}}' = \tilde{\boldsymbol{A}}\boldsymbol{B}^{-1}.\end{aligned} \tag{30}$$

$s(.,.)$ et $s'(.,.)$ sont les fonctions de saillances correspondantes à I et I' respectivement.
L'expression de $c(.,.;.)$ correspondante à la fonction $s(.,.)$ est donnée par l'équation (25) :

$$c(\boldsymbol{q}_0, \boldsymbol{A}; \boldsymbol{\delta q}_0) \approx \boldsymbol{\delta q}_0{}^\top \boldsymbol{A}^{-\top} \boldsymbol{M}(\boldsymbol{q}_0, \boldsymbol{A}) \boldsymbol{A}^{-1} \boldsymbol{\delta q}_0. \tag{31}$$

Il est facile de montrer que $M(.,.)$ peut être écrite comme :

$$\boldsymbol{M}(\boldsymbol{q}_0, \boldsymbol{A}) = \boldsymbol{B}^\top \boldsymbol{M}'(\boldsymbol{B}\boldsymbol{q}_0 + \boldsymbol{t}, \boldsymbol{A}\boldsymbol{B}^{-1}) \boldsymbol{B} \tag{32}$$

cela donne :

$$c(\boldsymbol{q}_0, \boldsymbol{A}; \boldsymbol{\delta q}_0) \approx \boldsymbol{\delta q}_0{}^\top \boldsymbol{A}^{-\top} \boldsymbol{B}^\top \boldsymbol{M}'(\boldsymbol{B}\boldsymbol{q}_0 + \boldsymbol{t}, \boldsymbol{A}\boldsymbol{B}^{-1}) \boldsymbol{B}\boldsymbol{A}^{-1} \boldsymbol{\delta q}_0. \tag{33}$$

Après le réarrangement, on obtient :

$$c(\boldsymbol{q}_0, \boldsymbol{A}; \boldsymbol{\delta q}_0) \approx \boldsymbol{\delta q}_0{}^\top \left(\boldsymbol{A}\boldsymbol{B}^{-1}\right)^{-\top} \boldsymbol{M}'(\boldsymbol{B}\boldsymbol{q}_0 + \boldsymbol{t}, \boldsymbol{A}\boldsymbol{B}^{-1}) \left(\boldsymbol{A}\boldsymbol{B}^{-1}\right)^{-1} \boldsymbol{\delta q}_0. \tag{34}$$

Les equations $\boldsymbol{q}'_0 = \boldsymbol{B}\boldsymbol{q}_0 + \boldsymbol{t}$ et $\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{B}^{-1}$ donnent :

$$c(\boldsymbol{q}_0, \boldsymbol{A}; \boldsymbol{\delta q}_0) \approx \boldsymbol{\delta q}_0{}^\top \boldsymbol{A}'^{-\top} \boldsymbol{M}'(\boldsymbol{q}'_0, \boldsymbol{A}') \boldsymbol{A}'^{-1} \boldsymbol{\delta q}_0. \tag{35}$$

En fin on obtient :

$$c(\boldsymbol{q}_0, \boldsymbol{A}; \boldsymbol{\delta q}_0) \approx c'(\boldsymbol{q}'_0, \boldsymbol{A}'; \boldsymbol{\delta q}_0), \tag{36}$$

ce qui implique

$$s(\boldsymbol{q}_0, \boldsymbol{A}) \approx s'(\boldsymbol{q}'_0, \boldsymbol{A}') \tag{37}$$

Ceci mène à l'approximation :

$$\begin{aligned}(\tilde{\boldsymbol{q}}_0, \tilde{\boldsymbol{A}}) &= \underset{(\boldsymbol{q}_0, \boldsymbol{A}) \in I_0 \times \mathbb{R}^3}{\arg \text{local max}}\ s(\boldsymbol{q}_0, \boldsymbol{A}) \iff \\ (\tilde{\boldsymbol{q}}'_0, \tilde{\boldsymbol{A}}') &= \underset{(\boldsymbol{q}'_0, \boldsymbol{A}') \in I'_0 \times \mathbb{R}^3}{\arg \text{local max}}\ s'(\boldsymbol{q}'_0, \boldsymbol{A}') \\ \text{avec : } \tilde{\boldsymbol{q}}'_0 &\approx \boldsymbol{B}\tilde{\boldsymbol{q}}_0 + \boldsymbol{t}\ ,\ \tilde{\boldsymbol{A}}' \approx \tilde{\boldsymbol{A}}\boldsymbol{B}^{-1}.\end{aligned} \tag{38}$$

Nous concluons que la fonction *warp-based autocorrelation* est approximativement covariante vis à vis des transformations affines.

3.2 Paramétrisation par points de contrôle

Des paramètres avec différentes unités vont générer des matrices de transformation $\boldsymbol{A}(\gamma)$ non normalisées, rendant à leur tour les matrices $\boldsymbol{M}$, liées la fonction de *warp-based autocorrelation*, non normalisées. Notre approche pour résoudre ce problème utilise une paramétrisation à base de points de contrôle comme suit :

1. Les points d'entrés sont paramétrées comme suit :

$$\boldsymbol{m}_0^i = \begin{pmatrix} x \\ y \end{pmatrix},\ \boldsymbol{m}_1^i = \begin{pmatrix} x + s_C + h_1 \\ y + h_2 \end{pmatrix},\ \boldsymbol{m}_2^i = \begin{pmatrix} x + h_2 \\ y + s_C - h_1 \end{pmatrix}, \tag{39}$$

2. Tandis que, les points de sorties sont fixes :

$$\boldsymbol{m}_0^o = \begin{pmatrix} x \\ y \end{pmatrix},\ \boldsymbol{m}_1^o = \begin{pmatrix} x + r \\ y \end{pmatrix},\ \boldsymbol{m}_2^o = \begin{pmatrix} x \\ y + r \end{pmatrix}, \tag{40}$$

3. Notant $\boldsymbol{M}^i$ et $\boldsymbol{M}^o$ deux ensembles d'entrée et de sortie respectivement :

$$\boldsymbol{M}^i = \begin{pmatrix} \boldsymbol{m}_0^i & \boldsymbol{m}_1^i & \boldsymbol{m}_2^i \end{pmatrix},\ \boldsymbol{M}^o = \begin{pmatrix} \boldsymbol{m}_0^o & \boldsymbol{m}_1^o & \boldsymbol{m}_2^o \end{pmatrix}. \tag{41}$$

La matrice de transformation $\boldsymbol{A}$ est calculée pour satisfaire la relation suivante : $\boldsymbol{M}^o = \boldsymbol{A}\boldsymbol{M}^i(\boldsymbol{\gamma})$.
Ici, $\boldsymbol{q}_0 = (x\ y)^\top$ sont les coordonnées spatiales d'un point dans l'image originale. r est une valeur fixée. En fait, la transformation affine $\boldsymbol{A}(\gamma)$ est définie comme la transformation d'une ellipse vers un cercle fixe. L'avantage de cette approche est d'utiliser les mêmes unités (pixels) pour tous les paramètres affines $\boldsymbol{q}_0 = (x\ y)^\top$ et $\boldsymbol{\gamma} = (s_C\ h_1\ h_2)^\top$. Cela conduit à une structure de tenseur $\boldsymbol{M}$ normalisée.

4 Résultats expérimentaux

Les expériences sont effectuées sur des séquences d'images réelles [4], obtenues à partir de la base de données de [5, 2]. Trois différents types de détecteurs sont inclus dans notre comparaison : Harris-Affine [1, 8], Hessian-Affine [1] et MSER [3]. Nous avons utilisé les implémentations originales des auteurs en ajustant les valeurs des paramètres par défaut recommandés par les auteurs. Les paramètres liés à notre détecteur comprennent : r le rayon du cercle prédéfini, σ_0 la variance de la Gaussienne isotrope $g(.,\sigma_0)$, k la taille de l'espace affine échantillonné

Figure 1 – *Quelques échantillons d'images de test. Colonne* (1) *:* 2 *images d'une séquence de changement d'échelle + rotation (*1*ère et* 6*ème séquences). Colonne* (2) *:* 2 *images d'une séquence de changement d'angle de vue (*1*ère et* 6*ème séquences). Colonne* (3) *:* 2 *images d'une séquence de changement d'illumination (*1*ère et* 6*ème séquences).*

4. http ://www.robots.ox.ac.uk/˜vgg/research/affine/
https ://lear.inrialpes.fr/people/Mikolajczyk/Database

(dans lequel, la recherche exhaustive est effectuée, *e.g.*, nombre d'échelles à tester), th le seuil de saillance (comme pour le détecteur de coins de Harris) pour éliminer les régions instables. Les valeurs expérimentales sont :

$$r = 3,\ k = 100,\ \zeta = 0.25,\ \sigma_0 = 2,\ th = 0.1$$

4.1 Évaluation par l'erreur de chevauchement

Nous comparons les performances de notre détecteur à celles obtenues par les détecteurs ci-dessus. En fait, ces détecteurs (*i.e.*, Harris-Affine [1, 8], Hessian-Affine [1] et MSER [3]) outrepassent les autres approches existantes. La comparaison est basée sur la mesure de répétabilité [2]. La répétabilité pour une paire d'images est calculée comme le rapport entre le nombre de correspondances point-à-points et le nombre minimum de points détectés dans les deux images [5]. Deux régions sont considérées en correspondance si l'erreur de chevauchement liée à la surface commune couverte par les deux régions est $\epsilon < 0.4$:

$$\epsilon = 1 - \frac{\boldsymbol{R}(\boldsymbol{E}_1) \cap \boldsymbol{R}(\boldsymbol{B}^\top \boldsymbol{E}_2 \boldsymbol{B})}{\boldsymbol{R}(\boldsymbol{E}_1) \cup \boldsymbol{R}(\boldsymbol{B}^\top \boldsymbol{E}_2 \boldsymbol{B})}. \tag{42}$$

Ici $\boldsymbol{R}(\boldsymbol{E}_1)$ et $\boldsymbol{R}(\boldsymbol{E}_2)$ sont deux régions elliptiques définies par : $\boldsymbol{q}^\top \boldsymbol{E}_1 \boldsymbol{q} = 1$ et $\boldsymbol{q}^\top \boldsymbol{E}_2 \boldsymbol{q} = 1$ respectivement, $\cap$ et $\cup$ correspondent aux opérateurs d'intersection et d'union respectivement. $\boldsymbol{B}$ est la transformation affine entre les deux images.

4.2 Répétabilité

Pour tous les détecteurs inclus (sauf indication contraire), nous avons fixé le seuil de l'erreur de chevauchement à 40% La taille normalisée des régions est ajustée à 30 pixels [2]. Nous évaluons la répétabilité de chaque détecteur vis à vis des changements d'échelle, changement d'angle de vue, changement d'éclairage et changement de flou. Des échantillons de séquences d'images utilisées sont affichées dans la figure 1. Le test de répétabilité permet d'évaluer comment le nombre de correspondances change entre l'image de référence et la séquence, en fonction du changement des transformations. Pour chaque image de la séquence, le nombre relatif et total des régions détectées sont enregistrées. Le détecteur idéal a un score de répétabilité élevé et un grand nombre de correspondances.

Changement d'échelle. Les résultats de ces évaluations sont reportés dans les figures 2(a) et 2(b). Bien que simple (notre algorithme est effectué en une seule étape), notre détecteur dispose d'un score de répétabilité meilleurs que ceux obtenus par les meilleures approches existantes. La figure 2(b) montre que notre détecteur a un nombre de correspondances supérieur à celui obtenu avec MSER, et inférieur à ceux obtenus avec Harris-Affine et Hessian-Affine. Cela signifie que nos régions détectées sont plus stables que celles obtenues par les autres détecteurs. Généralement, cela est lié à la propriété de covariance de la fonction *warp-based autocorrelation* par rapport aux transformations affines.

Changement d'angle de vue. Les figures 2(c) et 2(d) montrent les résultats de répétabilité et du nombre de correspondances en fonction des transformations liées au changement d'angle de vue. Le meilleur taux de répétabilité est obtenu avec le détecteur MSER suivi de notre détecteur (figure 2(c)). Cela est dû à la précision élevée de MSER, en particulier pour les régions homogènes avec des limites distinctives [2]. Le plus grand nombre de régions correspondantes est donné par Hessian-Affine (≈ 3000) suivi par Harris-Affine (≈ 2700) (voir Fig. 2(d)). Notre détecteur donne ≈ 500 régions et MSER est dernier avec ≈ 250 régions.

Changement d'éclairage. Les figures 2(e) et 2(f) affichent les résultats liés au changement d'éclairage. Il est facile de voir que notre détecteur et MSER obtiennent les taux les plus élevés de répétabilité. Les courbes de répétabilité liées à notre détecteur et MSER sont approximativement horizontales, soit un taux de répétabilité presque constante. Ceci montre qu'ils ont une bonne robustesse au changement d'éclairage. Le nombre de régions détectées avec notre détecteur est beaucoup plus grand que ceux obtenus avec les autres détecteurs (voir Fig. 2(f)). Cela montre clairement que notre détecteur répond à différents types d'éléments saillants, à savoir coins, tâches, etc.

Changement de flou. Les figures 2(g) et 2(h) montrent les résultats liés au changement croissant de flou. Notre détecteur est le plus performant, suivi par Hessian-Affine et Harris-Affine. La courbe de répétabilité pour notre détecteur est presque horizontale (figure 2(g)). Cela preuve que notre détecteur à un très haut niveau robustesse par rapport à la variation de flou. Le détecteur MSER est nettement le plus sensible au flou parce que les frontières des régions deviennent lisses, et le processus de segmentation est moins précis [2]. Le Hessian-Affine donne le plus grand nombre de régions suivies par notre détecteur. Le nombre de régions détectées avec MSER est très faible, aussi parce que les frontières de régions deviennent lisses.

La figure 3 montre des régions générées par, notre détecteur WBA, Hessian-Affine et MSER sur des sous-parties en correspondances de la première et la quatrième (angle de vue 40°) images de la séquence graffiti (mur), figure 1(c.-à-d., changement d'angle de vue). Les régions détectées sont en jaune et les régions correspondantes projetées à partir de la référence sont en bleu. Les ellipses sont à un facteur de 0.5 de la taille originale détectée.

5 Conclusion

Dans cet article, nous avons introduit une approche simple pour détecter les régions affines covariantes dans les images. Elle est basée sur ce que nous appelons la fonction *warp-based autocorrelation*. La fonction *warp-based autocorrelation* est covariante par rapport aux transformations affines des régions. En combinant la fonction *warp-based autocorrelation* avec un paramétrage à base de points de contrôle, on obtient un nouveau détecteur, répétable et précis vis à vis des transformations affines. Notre méthode est simple et efficace. Elle surpasse nettement les méthodes

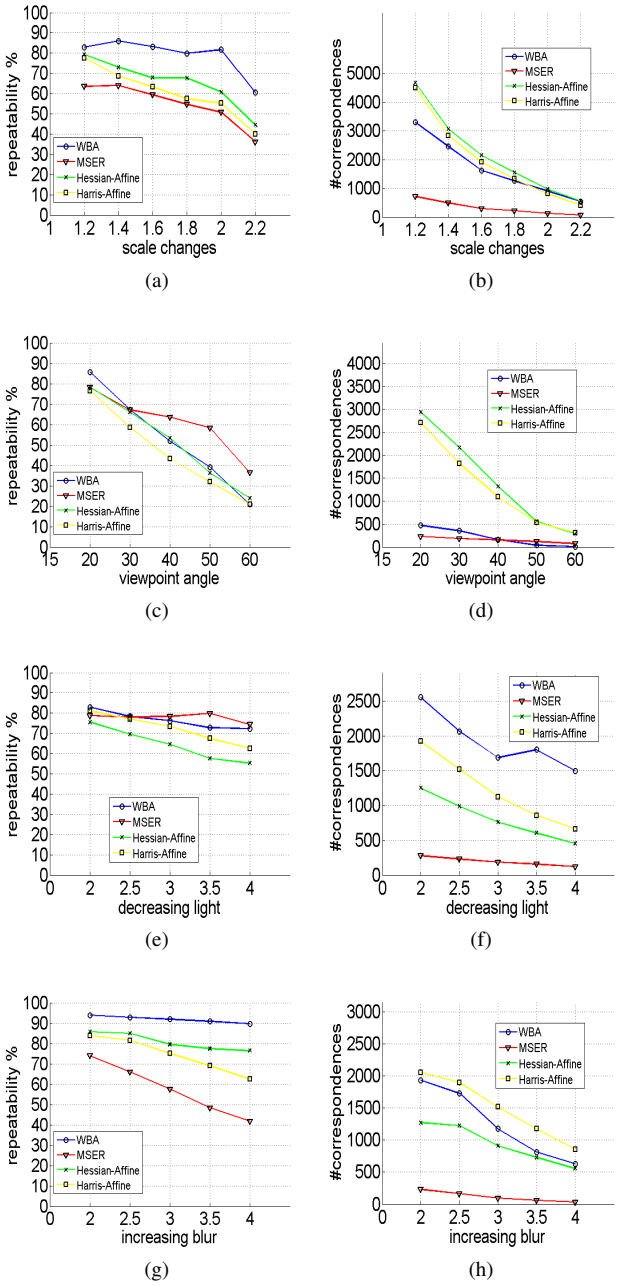


Figure 2 – *Répétabilitées et nombres de régions détectées (erreur de chevauchement = 40%, taille normalisée = 30 pixels) par rapport au changement d'échelle+rotation (la séquence Bateau, Fig. 1), changement d'angle de vue (la séquence Mur, Fig. 1), changement d'éclairage (la séquence Leuven, Fig. 1) et changement de flou (la séquence Motos). (a) Taux de répétabilité par rapport au changement d'échelle. (b) Nombre de régions détectées par rapport au changement d'échelle. (c) Taux de répétabilité par rapport au changement d'angle de vue. (d) Nombre de régions détectées par rapport au changement d'angle de vue. (e) Taux de répétabilité par rapport au changement d'éclairage. (f) Nombre de régions détectées par rapport au changement d'éclairage. (g) Taux de répétabilité par rapport au changement de flou. (h) Nombre de régions détectées par rapport au changement de flou.*

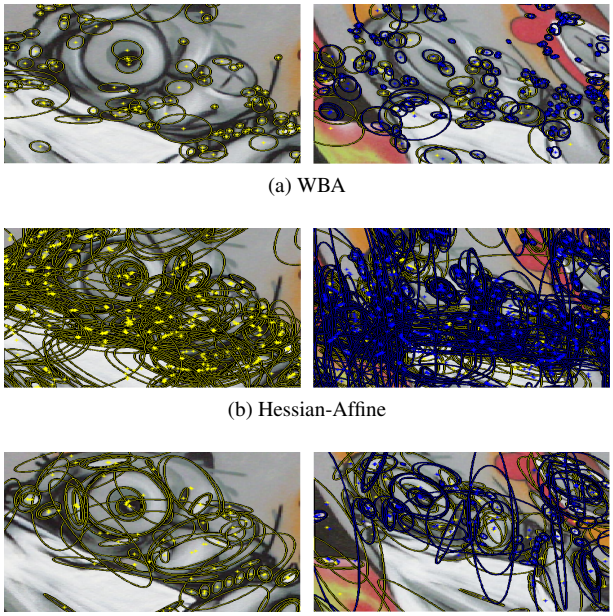


Figure 3 – *Exemple de régions détectées par : notre détecteur WBA, Hessian-Affine et MSER sur des sous-parties en correspondances de la 1ère (la référence) et la 4ème (angle de vue 40°) images de la séquence graffiti (mur), figure 1 (c.-à-d., changement d'angle de vue). Les régions détectées sont en jaune et les régions correspondantes projetées à partir de la référence sont en bleu. Les ellipses sont à un facteur de 0.5 de la taille originale détectée.*

existantes. Nos expériences ont montré l'efficacité de notre approche. Dans l'avenir, nous allons étendre notre approche vers les images 3D.

Références

[1] K. Mikolajczyk et C. Schmid. An affine invariant interest point detector. *ECCV*, 2002.

[2] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, et L.V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1) :43–72, 2005.

[3] J. Matas, O. Chum, M. Urban, et T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 22(10) :761–767, 2004.

[4] C. Harris et M. Stephens. A combined corner and edge detector. *Alvey Vision Conference*, 1988.

[5] K. Mikolajczyk et C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1) :63–86, 2004.

[6] T. Tuytelaars et L. Van Gool. Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision*, 59(1) :61–85, 2004.

[7] T. Tuytelaars et L. Van Gool. Wide baseline stereo matching based on local, affinely invariant regions. *BMVC*, 2002.

[8] F. Schaffalitzky et A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps ?". *ECCV*, 2002.

Une approche pour la catégorisation des objets 3D basée sur la théorie des fonctions de croyance

Hedi Tabia[1] Mohamed Daoudi[2] Jean-Philippe Vandeborre[2] Olivier Colot[1]

[1] Université Lille Nord de France / LAGIS - FRE CNRS 3303 / Université Lille 1, France.
e-mail : hedi.tabia@telecom-lille1.eu ; olivier.colot@univ-lille1.fr
[2] TELECOM Lille 1 / Institut TELECOM / LIFL UMR CNRS 8022 / Université de Lille 1, France.
e-mail : mohamed.daoudi@lifl.fr ; jean-philippe.vandeborre@telecom-lille1.eu

Résumé

Le groupement des objets 3D en catégories significatives est un problème très important dans le traitement des formes 3D. En introduisant une nouvelle technique de classification basée sur les fonctions de croyance, nous réussissons à catégoriser les objets 3D. Cette technique comporte deux étapes. Une première étape d'apprentissage, où les objets 3D d'une même catégorie sont traités et où un ensemble des parties représentatives de ces objets est construit, et une deuxième étape d'étiquetage, où des objets inconnus sont classifiés par catégorie. Le classifieur a été conçu et évalué sur une base de données de 400 objets 3D. Notre système atteint un taux de bonne reconnaissance de l'ordre de 85%.

Mots clefs

Catégorisation, classification, objets 3D, fonctions de croyance.

1 Introduction

L'indexation d'objet 3D est un sous-domaine très important de la vision par ordinateur et du multimédia. De nombreuses organisations ont d'importantes collections d'objets 3D sous format numérique disponibles pour un accès en ligne. L'organisation de ces collections en catégories dans le but de l'indexation est indispensable. Ces dernières années, de nombreux systèmes ont été proposés pour la recherche d'informations 3D.

Par exemple, Kazhdan et al. [1] décrivent une approche générale basée sur les harmoniques sphériques. Antini et al. [2] présentent une approche basée sur les corrélogrammes de courbures. Filali et al. [3] proposent une adaptation du k-plus proche voisin pour choisir les points de vue caractéristiques d'un modèle 3D.

Contrairement à l'indexation 3D où les objets sont comparés deux à deux, la classification 3D, qui consiste à affecter un objet requête à une catégorie, reste encore un problème ouvert. Peu de contributions telles que [4, 5] ont abordé ce problème. Leurs méthodes sont basées sur un classifieur Bayesien et sont limitées à des données très spécifiques.

Notre travail aborde le problème de catégorisation avec une approche basée sur les parties. Elle consiste à capturer un modèle compact d'une catégorie en construisant un ensemble de parties représentatives des différents objets dans cette catégorie. Dans ce but, les objets d'une même catégorie sont segmentés en plusieurs parties. Nous insistons ici sur le fait que nos parties sont des caractéristiques locales des objets et sont représentées par des descripteurs invariants.

Une fois que les parties provenant des objets de la même catégorie sont extraites, nous construisons un ensemble des parties représentatives. Cet ensemble nous permet de représenter tous les objets dans cette catégorie. Un moyen simple de construire cet ensemble est l'utilisation de techniques de quantification. Dans ce papier, nous utilisons une variante crédibiliste de k-plus proches voisins. Les centroïdes des clusters résultants servent comme des parties représentatives de notre catégorie. Ce processus est réitéré pour toutes les catégories dans l'ensemble d'apprentissage.

L'étiquetage des objets inconnus est réalisé par l'étiquetage de leurs parties associées. Ici, nous supposons qu'un objet appartient à une catégorie donnée lorsque la plupart de ses parties appartiennent à la même catégorie. Afin d'atteindre cet objectif, nous utilisons la théorie des fonctions de croyance. Plus précisément, chaque partie de l'objet à étiqueter est considérée comme une source d'informations fournissant certaines hypothèses concernant la catégorie d'appartenance de cet objet. En se basant sur ce raisonnement, les parties de l'objet sont comparées avec les parties représentatives d'une catégorie donnée et un ensemble de masses de croyance est calculé. À la suite de l'examen de chaque partie d'objet, on obtient un ensemble de *Basic Belief Assignments* (BBAs) qui peuvent être combinées en utilisant la règle de combinaison de Dempster pour former une BBA synthétisant une croyance finale concernant la catégorie de l'objet entier.

La suite du papier est organisée comme suit : dans la section 2, la phase d'apprentissage est présentée. Puis, dans la section 3, la phase de l'étiquetage est détaillée. La section 4 présente les résultats expérimentaux. La conclusion est présentée à la section 5.

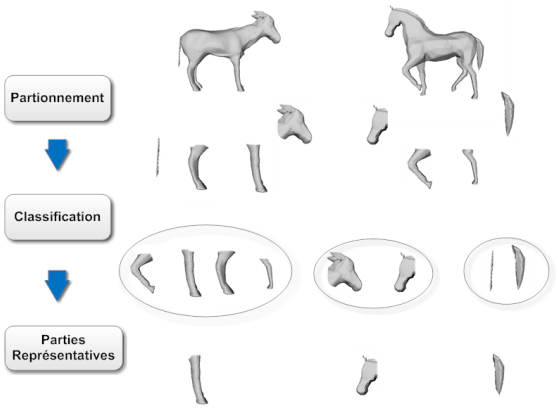


Figure 1 – *L'architecture du système d'apprentissage : Etant donnée une catégorie d'objets 3D. le système permet de construire un ensemble de parties représentatives. Les objets 3D dans cette figure sont issus de la catégorie "animal" de la base SHREC07.*

2 Modèle d'une catégorie

A partir d'un ensemble d'objets 3D d'apprentissage, la méthode que nous proposons pour construire le modèle de chaque catégorie est basée sur l'hypothèse suivante : "les objets 3D appartenant à la même catégorie possèdent les mêmes parties". Plus précisément, nous supposons que chaque catégorie peut être représentée par un ensemble de parties représentatives, avec lesquelles les objets dans cette catégorie peuvent être décrits. Dans cette section, nous mettons l'accent sur la construction de cet ensemble pour chaque catégorie. La figure 1 présente les différentes étapes du processus d'apprentissage. Elle montre trois étapes principales : (1) partionnement des objets 3D ; (2) classification des parties ; (3) calcul des parties représentatives.

2.1 Partitionnement d'objet 3D

Notre but est d'extraire les parties d'objets d'une même catégorie. Ces parties seront ensuite utilisées pour trouver les parties représentatives ainsi que le modèle de la catégorie. Il faut bien noter que le partionnement peut être sémantique ou non. Il y a plusieurs manières de partitionner un objet 3D. La façon la plus simple se base sur l'échantillonnage des points. La technique consiste à sélectionner un ensemble de sommets du maillage 3D et à associer à chaque point de l'échantillon une sous-partie de l'objet. Pour notre analyse, nous utilisons la technique des points les plus éloignés [6]. Pour calculer cet échantillonnage, Peyré et Cohen ont proposé un algorithme basé sur le Fast Marching [7].

Après le partitionnement des objets, les parties sont représentées par des descripteurs locaux qui décrivent leur géométrie. La performance de tout classifieur dépend fortement de la capacité discriminante des descripteurs. Choisir le bon descripteur est crucial. Les descripteurs doivent être suffisamment riches pour discriminer entre les différentes parties, et en même temps ils doivent être invariants sous les différentes transformations que peut subir une forme. Nous présentons ici quelques descripteurs qui peuvent être utilisés dans notre analyse.

- **GD2 :** Le GD2 est une distribution locale des distances géodésiques. Il a été utilisé par Osada et al. [8] pour la reconnaissance des formes 3D en mesurant les distances euclidiennes entre des points de surface tirés aléatoirement.
- **Gcords :** Le descripteur des cordes géodésiques est défini comme une distribution de distances géodésiques d'un point source à un ensemble de points. Ce descripteur est une extension de la distribution de cordes euclidiennes introduite par Paquet et Rioux. [9].
- **HSI :** Histogramme d'indices de forme. L'indice de forme, d'abord introduit par Koendrink et al. [10], est défini comme une fonction de deux courbures principales sur la surface 3D.

2.2 Parties représentatives

Dans cette section, nous proposons une méthode pour trouver des parties représentatives de chaque catégorie. Le but de la méthode consiste à rechercher une partition stable dans l'ensemble $P^C = \{P_1, ..., P_N\}$ (l'ensemble de toutes les parties dans une catégorie donnée C) vis-à-vis d'une règle de décision $\mathbb{R}$. La règle de décision utilisée est une extension aux fonctions de croyance de l'algorithme des plus proches voisins qui a été introduit par Zouhal et Denoeux [11]. Cet algorithme consiste à construire à partir d'un ensemble d'apprentissage $L = \{(P_i, \omega_i), i = 1, ..., n\}$ une fonction $\mathbb{R}$ par la méthode classique des k plus proches voisins. Dans sa version initiale, chaque voisin de P dans L est considéré comme une source d'informations quantifiant le degré de croyance sur l'appartenance du vecteur P à une classe dans $\Omega = \{\omega_i\}_{i=1,...,n}$ (dans notre cas P désigne une partie). Ainsi, une fonction de croyance m_k est directement construite en utilisant les informations apportées par les vecteurs P_k situés dans le voisinage du vecteur inconnu P par :

$$\begin{aligned} m_k(\omega_k) &= \chi e^{(-\gamma_k D^2)} \\ m_k(\Omega) &= 1 - m_k(\omega_k) \\ m_k(A) &= 0 \; \forall A \in 2^\Omega \backslash \{\Omega, \{\omega_k\}\} \end{aligned} \quad (1)$$

où D est la distance au vecteur P_k, χ un paramètre d'affaiblissement et γ_k est un paramètre de la classe ω_k déterminé par un processus d'optimisation proposé par Zouhal et Denoeux [11]. La masse de croyance allouée au singleton $\{\omega_k\}$ est donc un affaiblissement de la distance entre P et P_k. Les fonctions de croyance obtenues pour chaque voisin sont ensuite fusionnées sous la forme d'une fonction de croyance m définie sur Ω par la règle de combinaison de Dempster. Enfin, la décision quant à l'étiquetage de la partie P est prise en analysant la fonction de probabilité pignistique déduite à partir de m.

La règle de discrimination $\mathbb{R}$ ainsi construite permet d'introduire un algorithme itératif qui consiste à construire une suite de partitions crédales de P^C étant donné un nombre de classes. A partir d'une partition aléatoire initiale, chaque vecteur à classifier est tiré aléatoirement dans P^C et est ensuite classé par la règle de décision $\mathbb{R}$. L'itération se termine lorsqu'une partition stable est obtenue c'est-à-dire lorsque l'on n'observe aucun changement dans l'attribution des étiquettes. Le nombre de classes à l'itération initiale est fixé par le nombre d'individus dans P^C, ce qui permet d'obtenir un algorithme non paramétrique. Dans les itérations suivantes, le nombre de classes est réduit naturellement par l'attribution des vecteurs à une classe dans Ω. Ainsi l'algorithme obtenu peut être déroulé de la manière décrite dans l'Algorithme 1.
La partition finale obtenue est une partition crédale, c'est-à-dire que chaque partie est caractérisée par une fonction de croyance qui quantifie le degré d'appartenance. Il est à noter que le problème du conflit qui peut apparaître lors de la combinaison des fonctions de croyance devient crucial et qu'il doit être résolu avec un opérateur adéquat . Dans notre cas, nous avons opté pour l'utilisation d'un opérateur purement conjonctif.

Algorithm 1 : Classification des parties d'objets 3D d'une même catégorie

1: *Chaque partie dans P^C est considérée comme une classe. Ω est initialisé comme : $\Omega = \{\omega_1 ..., \omega_N\}$. N est le nombre des parties dans P^C.*
2: **repeat**
3: *Ordonner aléatoirement les parties dans P^C.*
4: **for all** P_i dans P^C en respectant l'ordre **do**
5: *Calculer m_k (pour chaque k-plus proche voisin de P_i)*
6: *Calculer m_i par combinaison des m_k utilisant la règle de Dempster*
7: *Calculer $BetP$ la probabilité pignistique induite par m_i*
8: *Changer l'étiquette de P_i dans Ω*
9: **end for**
10: *Analyser et réduire le nombre de classes dans Ω*
11: **until** Obtention d'une partition stable de Ω

Une fois le processus de regroupement des parties terminé, nous calculons le centroïde de chaque cluster. Les centroïdes ainsi déterminés présentent les parties représentatives de la catégorie C. Dans ce papier, une partie représentative est notée par R.

3 Etiquetage d'un nouvel objet 3D

Pendant le processus d'étiquetage, chaque catégorie d'objets 3D est considérée comme un ensemble de parties représentatives. L'étiquetage des objets 3D inconnus est réalisé par l'étiquetage de leurs parties associées. Un objet appartient à une catégorie donnée lorsque la plupart de ses parties appartiennent à la même catégorie. Dans cette section, nous mettons l'accent sur la mise en correspondance entre les parties d'un objet inconnu et les parties représentatives de chaque catégorie ainsi que l'étiquetage de l'objet 3D en entier. Soit O un objet 3D à classifier. On suppose que O est composé de N parties $\{P_i\}_{1\leq i\leq N}$. Disposant de J catégories $\Omega_C = \{C_j\}_{1\leq j\leq J}$. Une catégorie C_j est représentée par n_j parties représentatives $\{R_j r\}_{1\leq r\leq n_j}$

Similarity	C_1	C_2	...	C_J
P_1	$S(P_1, C_1)$	$S(P_1, C_2)$	...	$S(P_1, C_J)$
P_2	$S(P_2, C_1)$	$S(P_2, C_2)$	...	$S(P_2, C_J)$
...	...	...	...	...
P_I	$S(P_I, C_1)$	$S(P_I, C_2)$	...	$S(P_I, C_J)$

Tableau 1 – *Similarités entre parties 3D et catégories d'objets 3D.*

3.1 Mise en correspondance des parties 3D

La relation entre une partie 3D P_i et une catégorie donnée C_j est calculée en se basant sur la théorie des fonctions de croyance. Notre idée consiste à considérer les parties représentatives $\{R_{jr}\}_{1\leq r\leq n_j}$ comme des sources indépendantes donnant des informations concernant cette relation. Plus formellement, nous définissons une fonction de masse m_{ij}, qui quantifie le degré de croyance donné à l'hypothèse " P_i correspond à C_j".
P_i correspond à C_j si P_i tend vers une partie représentative de C_j. Dans le cas contraire, nous considérons que P_i correspond à $\overline{C_j}$ (le complément de C_j dans Ω_c). Ainsi la masse m_{ij} peut s'écrire sous la forme :

$$\begin{aligned} m_{ij}(C_j) &= \mu \cdot S(P_i, C_j) \\ m_{ij}(\overline{C_j}) &= \mu \cdot (1 - S(P_i, C_j)) \\ m_{ij}(\Omega_c) &= 1 - \mu \end{aligned} \tag{2}$$

$S(P_i, C_j) = e^{(-D(P_i, R_{P_i}))}$ est une fonction de la distance entre la partie P_i et sa plus proche partie représentative R_{P_i} dans la catégorie C_j. μ est un coefficient d'affaiblissement associé à la catégorie C_j. La table 1 montre les similarités $S(P_i, C_j)$ sous forme d'une matrice où les lignes représentent les parties de l'objet inconnu et les colonnes représentent les J catégories.
À la suite de l'examen de chaque catégorie, on obtient J BBAs comme le montre la table 2. Ces masses peuvent être combinées en utilisant la règle de combinaison de Dempster pour former une BBA m_i synthétisant une croyance définitive sur l'attachement de P_i à chaque catégorie. La table 3 montre la BBA résultante m_i.

3.2 Etiquetage de l'objet 3D

Afin d'obtenir une décision finale sur la catégorie de l'objet entier, toutes les masses m_i $i \in [1..I]$ sont combinées en utilisant la règle de Dempster de combinaison. Ainsi le processus d'étiquetage d'un objet inconnu peut se résumer dans l'algorithme 2.

BBA	C_1			...			C_J		
P_i	$m_{i1}(C_1)$	$m_{i1}(\overline{C_1})$	$m_{i1}(\Omega_c)$	...	...	...	$m_{iJ}(C_J)$	$m_{iJ}(\overline{C_J})$	$m_{iJ}(\Omega_c)$

Tableau 2 – *Modélisation d'une partie avec une masse de croyance.*

BBA	C_1	...	C_J	$\{C_1, C_2\}$	...	Ω_c	Φ
P_i	$m_i(C_1)$	...	$m_i(C_J)$	$m_i(\{C_1, C_2\})$	...	$m_i(\Omega_c)$	$m_i(\Phi)$

Tableau 3 – *La distribution finale de la masse de croyance pour une partie P_i.*

Algorithm 2 : Etiquetage d'un objet requête

1: Données : *Etant donné J catégories $\Omega_c = \{C_j\}_{1\leq j\leq J}$. Chaque catégorie est représentée par n_j parties représentatives $\{R_{jr}\}_{1\leq r\leq n_j}$. Etant donné un objet inconnu O.*
2: *Partionner O en N parties $\{P_i\}_{1\leq i\leq N}$.*
3: **for all** P_i dans $\{P_i\}_{1\leq i\leq N}$ **do**
4: **for all** C_j dans Ω_c **do**
5: *Trouver la partie representative R_{P_i} la plus proche de P_i.*
6: *Calculer m_{ij} (g to eq.2).*
7: **end for**
8: *Calculer m_i par combinaison des masses m_{ij} en utilisant la règle de combinaison de Dempster*
9: **end for**
10: *Calculer m par combinaison des m_i en utilisant la règle de combinaison de Dempster*
11: *Déduire les probabilités pignistiques induites par m*
12: *Etiqueter O suivant les probabilités pignistiques.*

4 Résultats expérimentaux

Dans cette section, nous présentons les résultats expérimentaux, la base de données et l'évaluation. Les algorithmes que nous avons décrits dans les sections précédentes ont été developpés avec MATLAB. Le système se compose d'une phase off-ligne pour l'apprentissage, où les parties représentatives de chaque catégorie sont extraites et une phase on-ligne pour l'étiquetage. Au cours de nos expériences, nous évaluons notre méthode selon deux points de vue. Qualitativement, nous présentons quelques résultats sous forme d'une matrice de confusion. Quantitativement, nous comparons le taux de bonne reconnaissance de notre méthode avec le classifieur bayésien Huber et al.[4]. En outre, nous évaluons le taux de bonne reconnaissance de notre classifieur en fonction des descripteurs utilisés.

4.1 Description de la base SHREC07

La base a été proposée par Marini et al. [12] dans le cadre de la compétition SHREC07 pour l'indexation des objets 3D. Les objets de cette base présentent des variations diverses allant du changement de pose, jusqu'à des transformations non-rigides au sein d'une même classe en passant par des modifications topologiques. La figure 2 montre la base de données SHREC07.

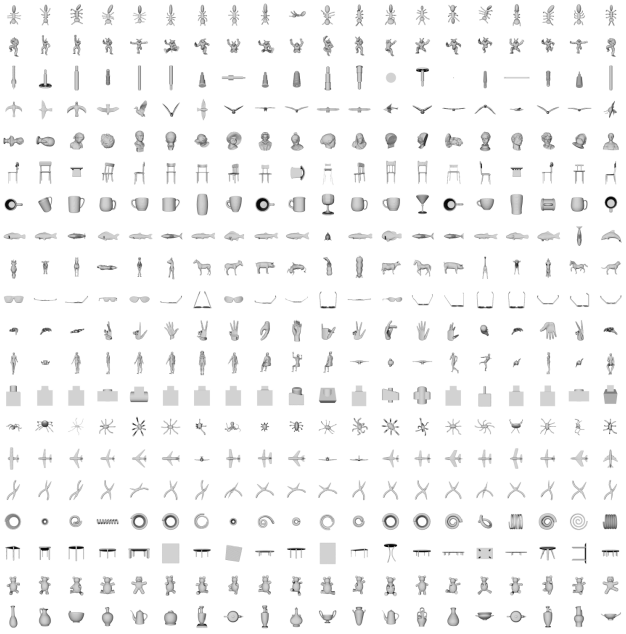


Figure 2 – *Aperçu sur la base d'objets 3D SHREC07.*

4.2 Résultats de la catégorisation

D'un point de vue qualitatif, la figure 3 montre une matrice de confusion. Les lignes de cette matrice correspondent à des parties extraites d'un objet requête correspondant à un modèle humain 3D. Les colonnes de cette matrice correspondent aux différentes catégories indiquées dans les lignes de la figure 2 (dans le même ordre). La brillance de chaque élément (i, j), dans cette matrice, est proportionnelle à la similarité entre les parties et les catégories. Les éléments d'une couleur froide représentent les meilleures correspondances, tandis que les éléments de couleur plus chaude indiquent une mauvaise correspondance. On remarque, dans cette visualisation, que les parties de l'objet humain ont tendance à converger vers la catégorie 12 qui correspond à celle des humains dans la figure 2. Ces résultats confirment notre hypothèse que les objets 3D dans la même catégorie partagent les mêmes parties.

La figure 4 montre une deuxième matrice de confusion. Les lignes de cette matrice représentent un ensemble d'objets 3D requêtes (les objets correspondent à la dernière colonne de la figure 2) et les colonnes représentent les catégories présentées par les lignes de la figure 2. La brillance des

éléments sur la diagonale de la matrice montre l'efficacité de notre classifieur.
Plus quantitativement, le tableau 4 montre les résultats de comparaison entre notre classifieur crédibiliste et le classifieur bayésien [4] (GD2 a été utilisé comme descripteur de parties dans cette comparaison). Notre classifieur a montré respectivement un taux de bonne reconnaissance de 89,6 % et 81,25% sur les données d'apprentissage et les données de test. Ce qui veut dire 84% de taux de bonne reconnaissance sur l'ensemble de la base entière. L'utilisation d'un classificateur bayésien [4] ne rapporte que 66,63% de taux de bonne reconnaissance.
En plus de l'utilisation du descripteur GD2, nous avons également testé les descripteurs HSI et Gcords. Le taux de bonne reconnaissance du classifieur basé sur HSI et Gcords n'a pas dépassé 69%, tandis qu'avec GD2 nous avons réussi à atteindre 84%. Ces résultats montrent l'efficacité de GD2 et son pouvoir discriminant dans la description des formes 3D. Une combinaison des GD2 et HSI donne un meilleur taux de bonne reconnaissance que GD2. La table 5 présente les résultats de la classification en fonction de ces descripteurs à la fois indépendants et d'autres combinés. La combinaison est basée sur la distance moyenne.

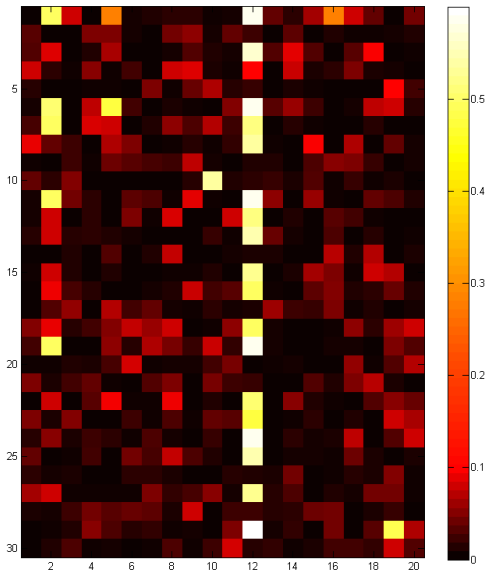


Figure 3 – *Matrice de confusion des parties d'un objet humain en fonction de l'ensemble des catégorie d'apprentissage. Les lignes présentent les parties requêtes. Les colonnes présentent les catégories d'objets. (GD2 a été utilisé comme descripteur de parties dans cette comparaison)*

4.3 Base d'apprentissage

Les performances de la classification dépendent de la taille de l'ensemble d'apprentissage : Plus cet ensemble est important, plus la performance de classification augmente. La table 6 compare le taux de bonne reconnaissance de la classification (basée sur le descripteur GD2) en fonction de

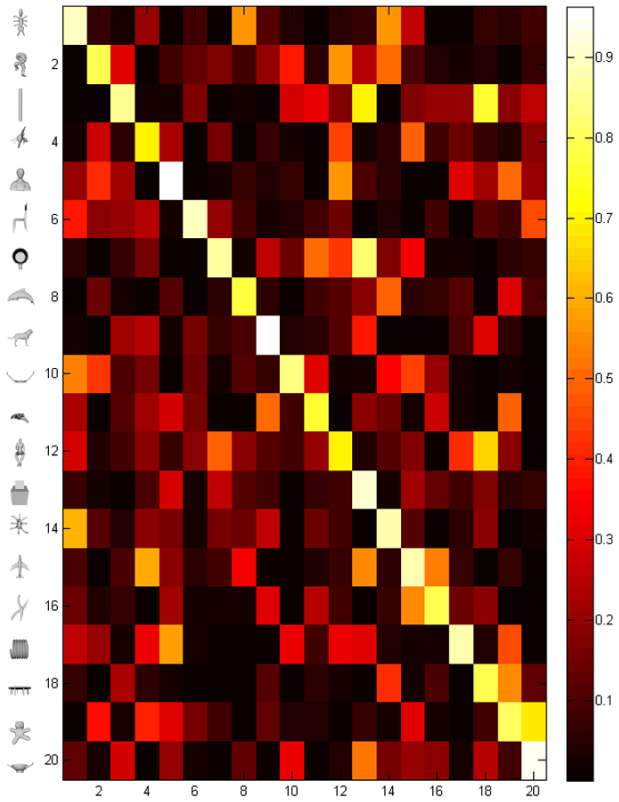


Figure 4 – *Matrice de confusion d'un exemple de catégorisation d'objets 3D. Les lignes correspondent à des objets requêtes. Les colonnes présentent les catégories d'objets. (GD2 a été utilisé comme descripteur de parties dans cette comparaison)*

la taille de l'ensemble d'apprentissage. Nous pouvons remarquer que l'augmentation de l'ensemble d'apprentissage améliore le taux de bonne reconnaissance de la classification.

5 Conclusion

Dans ce papier, nous avons présenté une nouvelle méthode pour catégoriser les objets 3D. Cette méthode se base sur la théorie des fonctions de croyance. Le processus de catégorisation est totalement automatisé. Il comporte deux étapes différentes. Une étape d'apprentissage qui consiste à trouver le modèle d'une catégorie, comporte de même deux étapes : 1) partionnement d'objet et 2) la construction des parties représentatives. La deuxième étape est l'étiquetage d'un objet inconnu dans lequel les fonctions de croyance ont été également utilisées. Le classifieur a été évalué sur une base de données de 400 modèles 3D. Notre système réussit à atteindre un taux de bonne reconnaissance de classification dépassant le 84%.

Références

[1] M. Kazhdan, T. Funkhouser, et S Rusinkiewicz. Rotation invariant spherical harmonic representation of

Base de test	Taille de la base	Notre méthode Taux de bonne reconnaissance (%)	classifieur bayesien Taux de bonne reconnaissance (%)
Base d'apprentissage	100	89.6	68.9
Base de test	300	81.25	65.5
Base entière	400	84	66.63

Tableau 4 – *Resultats de classification en fonction des classifieurs.*

Test	GD2	HSI	Gcords	GD2-HSI	GD2-Gcords	HSI-Gcords
Base d'apprentissage	89.6	68.3	71.6	91.2	90.2	80.4
Base de test	81.25	67.25	67.6	83.3	81.5	74.3
Base entière	84	67.6	68.9	85.9	84.4	76.33

Tableau 5 – *Le taux de bonne reconnaissance de classification en fonction des descripteurs utilisés.*

Apprentissage Taille	Test Taille	Taux de bonne reconnaissance %
100	400	84%
150	400	84.2%
200	400	85.6%
300	400	87.5%
400	400	88.1%

Tableau 6 – *Résultats de classification en fonction de la taille de la base d'apprentissage.*

3d shape descriptors. *Geometry Processing, Aachen, Germany*, 2003.

[2] G. Antini, S. Berretti, A. Del Bimbo, et P. Pala. Retrieval of 3d objects using curvature correlograms. Dans *IEEE International Conference on Multimedia & Expo*, July 2005.

[3] T. Filali Ansary, M. Daoudi, et J-P. Vandeborre. A bayesian 3D search engine using adaptive views clustering. *IEEE Transactions on Multimedia*, 9(1) :78–88, January 2007.

[4] D. Huber, A. Kapuria, R. Donamukkala, et M. Hebert. Parts-based 3d object classification. Dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[5] R. Donamukkala, D. Huber, A. Kapuria, et M. Hebert. Automatic class selection and prototyping for 3-d object classification. *3-D Digital Imaging and Modeling (3DIM)*, 2005.

[6] T-F. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theoretical Computer Science*, 1985.

[7] G. Peyre et L-D. Cohen. Geodesic remeshing using front propagation. *International Journal of Computer Vision*, 69 :145–156, 2006.

[8] R. Osada, T. Funkhouser, B. Chazelle, et D. Dobkin. Shape distributions. *ACM Transactions on Graphics*, 21(4) :807–832, 2002.

[9] E. Paquet et M. Rioux. Nefertiti : a query by content system for three-dimensional model and image databases management. *Image and Vision Computing*, 17 :157–166, 1999.

[10] J. Koenderink et A. van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 10 :557–565, 1992.

[11] L-M. Zouhal et T. Denoeux. An evidence-theoretic k-nn rule with parameter optimization. *IEEE Transactions on Systems*, 28 :263–271, May 1998.

[12] S. Marini, L. Paraboschi, et S. Biasotti. Shape retrieval contest 2007 : Partial matching track. *SHREC (in conjunction with IEEE Shape Modeling International)*, pages 13–16, 2007.

Reconnaissance de visages en 3D orientée région

P. Lemaire[1] P. Szeptycki[1] M. Ardabilian[1] L. Chen[1]

[1] LIRIS (Laboratoire d'InfoRmatique en Image et Systèmes d'information) -

Ecole Centrale de Lyon, Département Math-Info
Bât. E6, 36 avenue Guy de Collongue, 69134 Ecully Cedex – France

{pierre.lemaire, przemyslaw.szeptycki, mohsen.ardabilian, liming.chen}@ec-lyon.fr

Résumé

En reconnaissance faciale 3D, la robustesse des algorithmes aux expressions faciales demeure un défi. Dans cet article, nous abordons ce problème par une approche orientée régions. Le framework présenté est basé sur une nouvelle paramétrisation 3D du visage, invariante aux transformations non élastiques, utilisant les distances géodésiques. Il inclut un algorithme de segmentation automatique du visage en 3D inspirée par l'anatomie faciale, et introduit le concept de potentiel de déformation. Il comprend également une nouvelle méthode de fusion par pondération des scores de similarité. Enfin, nous présentons une évaluation des bénéfices de cette segmentation utilisant la base Bosphorus.

Mots clefs

Reconnaissance faciale, 3D, expressions faciales, région, segmentation, fusion de scores de similarité, distances géodésiques, ICP.

1 Introduction

Dans le domaine de la biométrie, la reconnaissance faciale est une modalités séduisante, car peu contraignante vis-à-vis des conditions d'acquisition et sans contact. Elle est néanmoins confrontée à de nombreux problèmes, parmi lesquels on peut notamment compter les variations d'éclairage, de pose, d'expression, les occultations [1]. La reconnaissance faciale en 3D permet d'en résoudre certains partiellement, notamment en ce qui concerne la pose et l'éclairage ; en revanche, elle accroit le temps de calcul sans pour autant résoudre le problème des expressions faciales [2].

De manière générale, un scénario de reconnaissance faciale classique compare un élément candidat a priori inconnu (probe) à un ensemble d'éléments connus (gallery). Le calcul d'un score de similarité entre le modèle probe et l'ensemble des modèles de la gallery permet d'identifier, ou de rejeter le candidat.

Plusieurs méthodes ont été proposées afin de permettre la reconnaissance automatique de visages en 3D.

La méthode considérée comme référence est une méthode de recalage de surfaces rigides, dite Iterative Closest Point (ICP) [3]. Le principe est de recaler par itérations successives 2 surfaces rigides en calculant la transformation (translation et rotation) optimale permettant de minimiser la distance point à point entre les 2 surfaces. L'écart restant est considéré comme une mesure de similarité.

Ben Amor et al. [9] proposent une méthode appelée R-ICP inspirée d'ICP sur l'ensemble du visage, mais ne prenant en compte que la partie statique du visage dans l'étape de recalage. Le score de similarité est obtenu en effectuant une somme pondérée sur les distances point à point entre la partie statique et la partie rigide du visage.

Faltemier et al. [8] proposent d'extraire 28 régions sphériques du visage, et d'effectuer un recalage rigide (par méthode ICP) entre probe et gallery sur chacune d'entre elles. La fusion des scores est inspirée de la méthode Borda Count. Pris individuellement, les scores par région indiquent l'importance des régions statiques du visage relativement aux régions mimiques dans la reconnaissance faciale en 3D.

Drira et al. [6] proposent de représenter le visage 3D sous forme de courbes géodésiques, et d'étudier le coût des déformations nécessaires à appliquer à ces courbes pour passer d'un visage probe à un visage gallery, ce qu'ils appellent un chemin géodésique, en tant que score de similarité. Cette méthode les amène à sélectionner et pondérer les meilleures courbes, correspondant globalement à celles subissant le moins de déformations lors des expressions.

Kakadiaris et al. [7] propose de modéliser le visage à l'aide d'un modèle annoté et déformable. A l'aide de ce dernier, on projette des caractéristiques (orientation des normales à la surface notamment) du visage 3D étudié sur un plan 2D, normalisé d'où sont extraits des descripteurs à l'aide de transformées en ondelettes. La fusion des descripteurs est pondérée selon la région du visage d'où sont extraites les caractéristiques, afin d'être moins sensible aux expressions.

Mian et al. [4] et Huang et al. [5] proposent une méthode holistique. Ils extraient un nombre réduit de points caractéristiques de la surface 3D, formant ainsi un graphe,

et en étudient la similarité en tant que mesure de similarité entre visages 3D. Les auteurs reportent des résultats relativement sensibles aux expressions.
La plupart des méthodes proposées dans l'état de l'art emploie ainsi diverses techniques dont l'objectif est souvent de réduire la sensibilité aux expressions.
Dans cet article, nous nous proposons d'adresser le problème de la reconnaissance faciale en 3D, et plus précisément celui de la robustesse aux expressions, à l'aide du système (*framework*) suivant. Dans un premier temps, le visage en 3D est prétraité et on en extrait des points caractéristiques. Puis on effectue une paramétrisation de la surface 3D à l'aide des distances géodésiques. Nous apparions ensuite le visage 'probe' au visage 'gallery', préalablement et automatiquement segmenté en régions, à l'aide de notre paramétrisation. Enfin, le score de similarité entre 2 visages est obtenu comme une somme pondérée des scores sur chacune des régions. Le score associé à une région est calculé sur la base d'une méthode de recalage rigide.
La suite de cet article est organisée de manière suivante.
La section 2 est consacrée à la présentation des bases de notre théorie et la présentation de notre approche. Dans la troisième partie, nous évaluons les bénéfices de notre approche à l'aide d'une expérimentation sur la base Bosphorus [11]. Nous concluons finalement par la section quatre tout en présentant quelques perspectives.

2 Approche proposée

La notion d'expression faciale étant très subjective, nous privilégions l'étude des déformations causées par les expressions. Ces dernières peuvent être décomposées sous forme de l'activation d'unités d'action (Action Units, AU), telles que décrites par Ekman et al. [10] au sein du Facial Action Coding System (FACS). Chaque AU peut se décrire comme étant la contraction ou la détente d'un ou plusieurs muscles du visage. D'autre part, Ben Amor et al. [9] ont mené une étude quant à l'influence de l'activation des principaux muscles et groupes musculaires faciaux (figure 1) sur les déformations de la surface faciale. Cette étude révèle que l'activation de certains muscles et groupes musculaires engendre des déformations avec une amplitude variant selon la région où ils sont situés sur le visage (figure 2). Plus précisément, l'activation de groupes musculaires sur le haut du visage engendre moins de déformations que l'activation de groupes musculaires sur le bas du visage.
Cette expérimentation est en adéquation avec le point de vue anatomique selon lequel certaines régions du visage sont statiques tandis que d'autres sont mimiques. Cela corrobore également l'idée selon laquelle l'activation d'une AU engendre des déformations locales, tandis que d'autres parties du visage restent statiques ou quasiment statiques. Cette étude anatomique nous confirme en outre l'existence de points de repère anthropométriques tels que les coins intérieurs des yeux et le bout du nez, dont la présence et la localisation sur la surface du visage est indépendante des expressions et des morphologies.

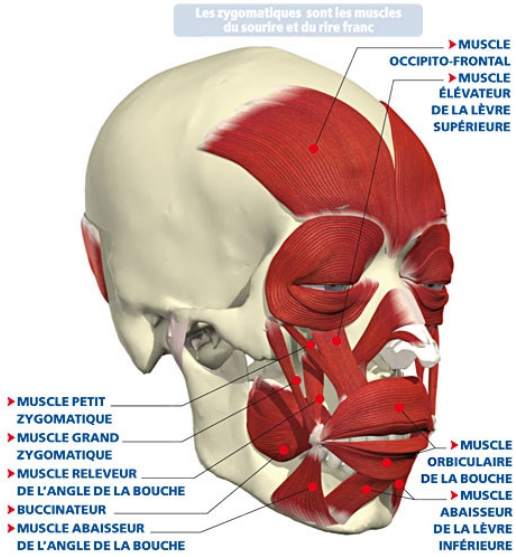


Figure 1 - *Anatomie du visage humain*

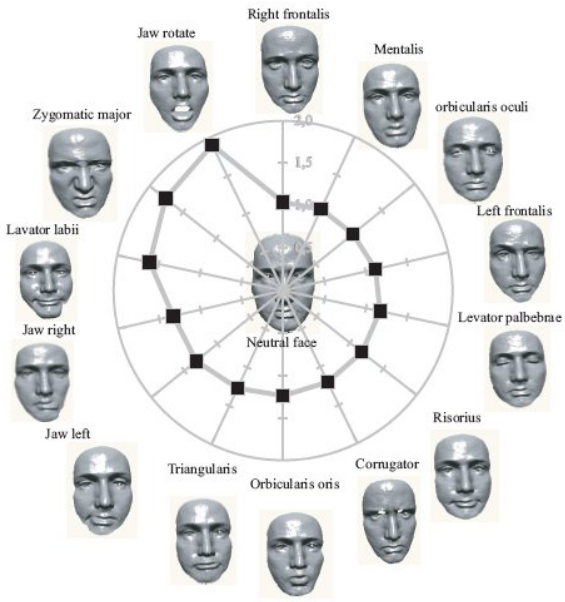


Figure 2 - *Classement de l'amplitude des déformations en fonction du groupe musculaire activé*

Malheureusement, la modélisation complète du système musculaire facial chez l'humain est difficile à mettre en œuvre. D'autre part, l'étude présentée dans [16] montre que les expressions engendrent des déformations élastiques de la surface du visage.
Nous cherchons donc dans ce travail à segmenter le visage en plusieurs régions, correspondant aux régions déformées ou non par les principales AU. Cette segmentation se doit d'être indépendante autant que possible des déformations élastiques et non élastiques du visage, pour un individu

donné. Elle doit permettre une meilleure robustesse aux expressions dans un scénario de reconnaissance ou d'identification.

2.1 Schéma global

Globalement, l'ensemble de l'approche de reconnaissance basée région peut être scindée en deux phases, l'une hors-ligne et l'autre en-ligne (figure 3). La première phase est consacrée au traitement et à l'analyse des modèles de la gallery. Pendant la phase en-ligne, un modèle probe est apparié avec les modèles de la gallery, et les mesures de similarité sont calculées. Une mesure de similarité entre deux visages est la pondération des mesures de similarité par région, dont les coefficients ont été appris suite à un apprentissage qui sera abordé dans la section 2.5.
Dans la suite nous détaillons les étapes de l'approche proposée.

2.2 Prétraitement

2.2.1 Qualité des modèles

Les objectifs du prétraitement des modèles 3D sont de minimiser l'influence de la qualité de l'acquisition à l'étape de la reconnaissance. Les données sont en effet généralement des images de profondeur et non pas des modèles 3D complets, ce qui implique que certaines parties du visage peuvent être manquantes. Ils comportent souvent également des pics et des trous, et font presque systématiquement l'objet d'un bruit d'acquisition. Pour assurer l'invariance aux poses, un modèle 3D complet est généré. Afin de supprimer les pics, on applique un filtre médian aux points dont les coordonnées en Z sont détectées comme aberrantes. Afin de corriger les trous, c'est-à-dire de déterminer les coordonnées en Z des points manquant sur l'image de profondeur, on opère une régression linéaire sur l'estimation des coefficients polynomiaux biquadratiques. Le détail de cette partie est disponible dans [12].

2.2.2 Localisation des points de repère

La segmentation du visage, c'est-à-dire la localisation des zones de déformations liées aux AU, repose en grande partie sur une bonne localisation, c'est-à-dire robuste et précise, des points de repère. Cette localisation doit être invariante aux poses et aux conditions d'éclairage. Elle ne doit donc de préférence pas utiliser l'information texturale. Notre approche de localisation de points de repère anthropométriques est basée sur l'utilisation des

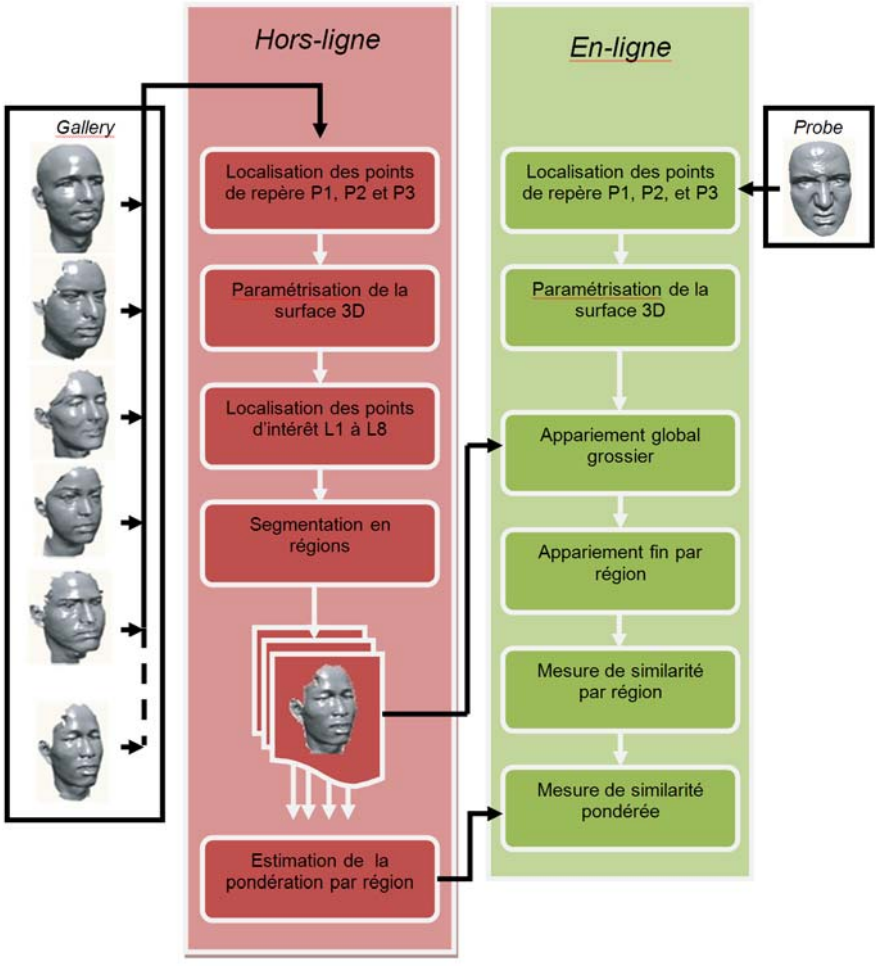


Figure 3 - *Schéma global de notre framework*

courbures moyenne et gaussienne, et un modèle générique. Par seuillages successifs, et en étudiant les courbures moyennes et gaussiennes à différents rayons, nous localisons les coins intérieurs des yeux (notés $\mathbf{P_1}$ et $\mathbf{P_2}$) et le bout du nez (noté $\mathbf{P_3}$), en affinant progressivement la précision de la localisation. Le détail de cet algorithme est disponible dans [12].

2.3 Paramétrisation de la surface 3D faciale

L'étape de paramétrisation vise à accorder les mêmes coordonnées à un même point physique, dans deux modèles 3D faciaux différents d'un même individu. On suppose que les points de repère anthropométriques $\mathbf{P_1}$ à $\mathbf{P_3}$ (coins intérieurs des yeux et bout du nez) sont localisés de manière précise. Ils sont naturellement distincts. Chaque point **p** d'un modèle facial 3D est alors décrit de manière unique par ses distances géodésiques ($\mathbf{d_1}$, $\mathbf{d_2}$, $\mathbf{d_3}$) à ces 3 points (figure 4). Cette paramétrisation comporte l'avantage d'être invariante aux déformations non élastiques, c'est-à-dire aux translations et aux rotations. Contrairement à la représentation euclidienne donc, cette paramétrisation n'exige pas de recalage relatif à la pose. Elle est appliquée à chaque visage de la gallery, mais aussi à chaque visage probe. A un point **p** de coordonnées ($\mathbf{d_1}$, $\mathbf{d_2}$, $\mathbf{d_3}$) d'un visage gallery, on apparie **p'** de coordonnées ($\mathbf{d_1'}$, $\mathbf{d_2'}$, $\mathbf{d_3'}$) du visage probe tel que la distance **D** entre ($\mathbf{d_1}$, $\mathbf{d_2}$, $\mathbf{d_3}$) et ($\mathbf{d_1'}$, $\mathbf{d_2'}$, $\mathbf{d_3'}$) soit minimale. Dans nos travaux, **D** est une distance euclidienne.

2.4 Segmentation en régions

La segmentation du visage constitue l'une des étapes centrales de l'approche de reconnaissance par région. C'est à cette étape que les régions statiques et mimiques du visage sont localisées automatiquement, pour chaque visage de la gallery.

La segmentation d'un visage selon des critères purement anatomiques est délicate, étant donné qu'aucun champ d'action d'un groupe musculaire ou d'une AU ne peut semble-t-il être localisé simplement selon des critères liés à la forme ou à la texture du visage. Nous avons donc choisi de déterminer nos régions de manière empirique, à l'aide de cartes dites de potentiel de déformation. Le potentiel de déformation est ici la mise en correspondance, via la paramétrisation précédemment exposée, d'une mesure de courbure entre 2 visages différents. En l'occurrence, la mesure de courbure dont nous nous sommes servis est le shape-index [13] avec un rayon de 25mm, indiquant la topologie de la surface au voisinage d'un point du modèle 3D. En appariant de la sorte plusieurs visages d'une même personne soumis à l'activation d'une ou plusieurs Action Units (AU), nous pouvons mettre en évidence à l'aide du potentiel de déformation les régions statiques (le shape-index varie peu d'un visage à l'autre, c'est-à-dire que sa variance est faible) et les régions mimiques (le shape-index varie largement d'un visage à l'autre, c'est-à-dire que sa variance est grande).

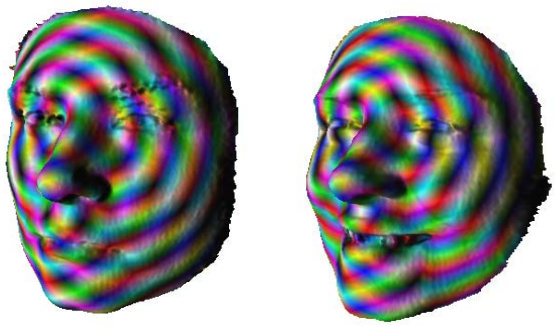


Figure 4 - *La paramétrisation du visage à partir des points d'intérêt $\boldsymbol{P_1}$ à $\boldsymbol{P_3}$, utilisant les distances géodésiques. Ici, les composantes RVB correspondent respectivement aux distances géodésiques à $\boldsymbol{P_1}$, $\boldsymbol{P_2}$ et $\boldsymbol{P_3}$.*

A la suite de nos observations sur le potentiel de déformation à l'aide de la base Bosphorus [11] (détaillée dans la partie 3), nous avons décidé de découper le visage en 8 régions, correspondant au nez, aux yeux gauche et droit, aux pommettes gauche et droite, au front ainsi qu'à la mâchoire, répartie en 2 régions séparées (figure 5). Ces régions correspondent à des régions globalement stables, ou globalement mimiques, selon nos observations sur le potentiel de déformation, sur plusieurs visages. La séparation de la bouche en 2 régions distinctes est supposée permettre une meilleure robustesse aux problèmes d'expressions partielles, de modèles partiels ou aux occultations. Afin de segmenter la surface faciale 3D de la sorte, nous localisons d'abord les points de repère $\mathbf{L_1}$ à $\mathbf{L_8}$ correspondant aux représentants des 8 régions. Ceux-ci sont localisés grâce à leurs distances géodésiques à $\mathbf{P_1}$, $\mathbf{P_2}$ et $\mathbf{P_3}$, à l'aide d'un modèle générique. Ensuite, chaque point du visage est associé respectivement à la région $\mathbf{S_i}$ (centrée en $\mathbf{L_i}$) à l'aide de deux paramètres, $\mathbf{w_i}$ et $\mathbf{R_i}$, $\mathbf{i}\in[1,8]$. Plus précisément, tout point **p** du visage appartient à la région $\mathbf{S_i}$, représentée par le point de repère $\mathbf{L_i}$, lorsque

$$s_i = \min(s_j) \text{ avec } j \in [1,8]$$

où $\mathbf{s_i}$ est calculé de la manière suivante :

$$\text{Si } w_i \times d_i > R_i \text{ alors } s_i = \infty$$

$$\text{sinon } s_i = w_i \times d_i$$

où $\mathbf{d_i}$ est la distance géodésique du point **p** au point de repère $\mathbf{L_i}$.

2.5 Calcul de similarité et pondération par région

Une fois que la segmentation en régions est effectuée, on calcule un score de similarité pour chaque région. Cette étape est relativement indépendante de la segmentation en

régions du visage, et si dans cet article nous présentons des résultats relatifs à l'algorithme de référence ICP, afin d'évaluer les bénéfices de la segmentation du visage, il est envisageable d'utiliser toute autre mesure de similarité dans cette partie.

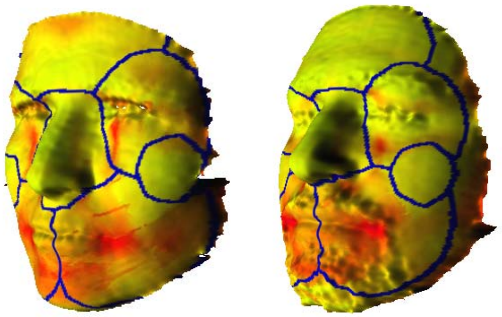


Figure 5 - *Exemples de segmentation du visage en régions, sur des cartes de potentiel de déformation*

Pour chaque point **p** du visage gallery dans une région donnée, nous apparions le point **p'** correspondant du visage probe suivant notre paramétrisation. Une itération d'ICP permet de minimiser la distance euclidienne au sens des moindres carrés entre les deux régions. C'est-à-dire qu'on calcule la rotation et la translation qui minimise cette distance. Le score de similarité par région est la moyenne des distances euclidiennes point à point entre le visage gallery et le visage probe au sein de cette région.
Finalement, le score de similarité entre deux visages est une mesure de similarité pondérée de la forme suivante :

$$M = \sum_{i=1}^{8} W_i \times d(S_i, S'_i)$$

$$\text{avec } d(S_i, S'_i) = \frac{1}{\text{card}(S_i)} \sum_{p_j \in S_i,\, p'_j \in S'_i} d_{euclid}(p_j, p'_j)$$

où card($\mathbf{S_i}$) correspond au nombre de points **p** compris dans la région $\mathbf{S_i}$. La détermination des poids lors de cette fusion de scores est déterminée à l'aide d'une base d'apprentissage. Les poids sont déterminés de sorte à minimiser l'Equal Error Rate (EER). Les détails concernant cet algorithme de fusion sont disponibles dans [14].

3 Expérimentation

L'expérimentation a été menée sur la base de visages Bosphorus [11]. Cette base est constituée des images de profondeur de 105 individus, pour la plupart des acteurs professionnels, avec 1 à 3 modèles neutres par individu. Pour chaque individu, la base comporte 34 modèles avec déformations, dont 28 obtenus par l'activation d'AU isolées et 6 par expression. La base présente également des modèles avec occultations (lunettes ou mains) ou rotations. Nous n'avons pas utilisé ces derniers dans l'expérimentation présentée dans cet article. Cette base est réputée difficile, étant centrée sur les expressions faciales.
Concernant notre expérimentation, 29 individus ont été sélectionnés aléatoirement, avec au moins un modèle neutre par individu. Ces modèles neutres constituent notre gallery. Par la suite, 80 modèles probe dont 74 avec expression et 6 neutres ont été sélectionnés pour évaluer les performances de notre algorithme et les bénéfices de la segmentation en région présentée dans cet article. La moitié de ces modèles est utilisée pour estimer la pondération optimale par région, l'autre constitue les modèles de test.
Nous avons comparé les améliorations incrémentales de notre approche (approche D) avec l'approche baseline ICP. Les améliorations successives de l'approche R-ICP [9] sont obtenues en additionnant simplement les scores par région mais sans apprentissage (approche B), en fusionnant les scores par la méthode Borda Count modifiée présentée dans [8] (approche C), et finalement en utilisant la pondération optimale des scores basée sur l'EER, telle que présentée dans [14].
Les résultats sont présentés dans le tableau 1. On remarque une amélioration progressive du taux de reconnaissance avec le raffinement des méthodes de fusion. Dans tous les cas, notre approche de segmentation du visage en régions ainsi que notre paramétrisation permettent une amélioration du taux de reconnaissance.
L'expérimentation menée jusqu'à présent, même si elle reste à affiner et à approfondir, présente une première tendance. Dans la suite, il est prévu de procéder à des tests plus significatifs en augmentant le nombre d'individus, en utilisant les bases FRGC 1.0, FRGC 2.0 [15] et finalement en procédant à une validation croisée.

Approche	**Taux d'identification**
A - baseline ICP	78.5%
B - régions sans pondération	82.5%
C - régions Borda Count	87.5%
D - régions avec apprentissage	92.5%

Tableau 1 - *Résultats de l'expérimentation : améliorations successives de notre approche comparées à l'approche baseline ICP*

4 Conclusion et perspectives

Dans cet article, nous avons présenté nos travaux sur la reconnaissance faciale en 3D orientée régions. Les expérimentations menées montrent des résultats encourageants concernant l'approche région, avec une augmentation de 14% du taux de reconnaissance par rapport à la méthode de référence, en présence d'expressions faciales. Cependant, il reste de nombreux points à améliorer. Notons parmi ceux-ci le score de similarité entre régions, pouvant être amélioré par des mesures d'interpénétration de surface par exemple. Notons également qu'une stratégie de cascade pour la

fusion des scores de similarité peut permettre de rejeter prématurément un visage, réduisant ainsi le temps de calcul. Le cas des occultations est également un travail qu'il nous reste à mener ; une approche orientée région semblerait adaptée pour mieux le traiter.

Références

[1] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face Recognition: A Literature Survey, *ACM Computing Surveys*, vol. 35, no. 4, pp. 399-458, 2003

[2] K.W. Bowyer, K. Chang, P.J. Flynn, A Survey of Approaches and Challenges in 3D and Multi-Modal 3D+2D Face Recognition, *Computer Vision and Image Understanding*, vol. 101, pp. 1-15, 2006

[3] P.J. Besl and N.D. McKay, A Method for Registration of 3-D Shapes, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239-256, Feb. 1992

[4] A. Mian, M. Bennamoun, R. Owens, An Efficient Multimodal 2D-3D Hybrid Approach to Automatic Face Recognition, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 11,pp. 1927-1943, June 2007

[5] D. Huang, M. Ardabilian, Y. Wang, L. Chen. Asymmetric 3D/2D Face Recognition Based on LBP Facial Representation and Canonical Correlation Analysis, *IEEE International Conference on Image Processing (ICIP 2009)*, Cairo, Egypt, 2009

[6] H. Drira, B. Ben Amor, A. Srivastava, M. Daoudi, A Riemannian Analysis of 3D Nose Shapes For Partial Human Biometrics, *ICCV 2009*, Kyoto, Japan, 2009

[7] I. Kakadiaris, G. Passalis, G. Toderici, N. Murtuza, T. Theoharis, Three-Dimensional Face Recognition in the Presence of Facial Expression: An Annotated Deformable Model Approach, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 640-649, avril 2007

[8] T. Faltemier, K.W. Bowyer, P.J. Flynn, A Region Ensemble for 3D Face Recognition, *IEEE Trans. Information Forensics and Security*, vol. 3, no. 1, pp. 62-73, Mar. 2008

[9] B. Ben Amor, Contributions à la reconstruction, la reconnaissance et l'authentification faciale 3D, thèse préparée au sein du laboratoire LIRIS, sous la direction de L. Chen, M. Ardabilian, soutenue le 8 décembre 2006

[10] P. Ekman, W.V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement, *Consulting Psychologists Press*, Palo Alto, 1978

[11] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, L. Akarun, Bosphorus Database for 3D Face Analysis, *Workshop on Biometrics and Identity Management (BIOID)*, Denmark, May 2008

[12] P. Szeptycki, M. Ardabilian, L. Chen. A coarse-to-fine curvature analysis-based rotation invariant 3D face landmarking, *International Conference on Biometrics: Theory, Applications and Systems*, Washington, 2009

[13] JJ. Koenderink, AJ. van Doorn, Surface shape and curvature scales, *Image and vision computing*, 1992

[14] W. Ben Soltana, M. Ardabilian, L. Chen, Comparison of 2D/3D Features and their adaptive Score Level Fusion for 3D Face Recognition, *3DPVT 2010* (soumis, accepté)

[15] P.J. Phillips, P.J. Flynn, T. Scruggs, K.W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the Face Recognition Grand Challenge, *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 947-954, 2005

[16] I. Mpiperis, S. Malassiotis, MG. Strintzis, Expression Compensation for Face Recognition Using a Polar Geodesic Representation, *3D Data Processing, Visualization, and Transmission (3DPVT'06)*, 2006

Bornes de Cramér-Rao en Estimation Fréquentielle 3-D

B. Akssase[1] M. Ouanan[1]

[1] Equipe ASIA (Analyse des Systèmes et Informatique Appliquée)

Université Moulay Ismail, Faculté des Sciences et Techniques, Errachidia
BP 509, Boutalamine, 52000 Errachidia, Maroc

{ouanan_mohammed@yahoo.fr, baksasse@yahoo.com}

Résumé

Dans ce travail nous considérons le problème de l'estimation des fréquences d'un mélange de sinusoïdes noyées dans un bruit additif dans le cas tridimensionnel (3-D). La méthode d'estimation des fréquences est de type ESPRIT en utilisant les statistiques de second ordre (SSO) en l'occurrence la fonction d'autocorrélation lorsque le bruit additif est blanc gaussien (BABG). Lorsque le bruit additif est coloré gaussien (BACG), une version de la méthode ESPRIT 3-D qui exploite les statistiques d'ordre supérieures (SOS) est utilisée. Pour le souci d'efficacité des estimateurs des fréquences 3-D, la borne de Cramér-Rao (BCR) asymptotique est calculée, ainsi l'erreur quadratique moyenne des fréquences s'approche asymptotiquement de la BCR ce qui prouve l'efficacité de la méthode d'estimation.

Mots clefs

Borne de Cramér-Rao (BCR), méthode ESPRIT, bruit additif blanc gaussien (BABG), bruit additif coloré gaussien (BACG), rapport signal sur bruit (RSB).

1

1 Introduction

La modélisation des signaux par une somme d'exponentielles perturbées par un bruit additif est utilisée dans plusieurs applications telles les télécommunications, le traitement d'antennes, l'analyse d'images par résonance magnétiques, ou encore le traitement d'images sismiques. Ce type de modélisation implique le problème fondamental de l'estimation des paramètres : fréquences, amplitudes et ordre du modèle. L'estimation des fréquences est souvent effectuée par des méthodes dites à haute résolution (HR). Ces méthodes exploitent le modèle du signal et la décomposition de l'espace des observations en deux sous-espaces orthogonaux à savoir le sous-espace signal et le sous-espace bruit. L'idée de base des méthodes HR analytiques d'estimation fréquentielle 1-D est la méthode ESPRIT (Estimation of Signal Parameters via Rotation Invariante Technique) qui s'appuie sur la propriété de l'invariance par rotation du sous-espace signal [1]. Différentes extensions de cette méthode aux cas des signaux 2-D et 3-D ont été développées récemment. Nous pouvons citer par exemple la méthode MEMP 2-D (Matrix Enhancement and Matrix Pencil) [2], la méthode ESPRIT 2-D [3], et leur extension au cas 3-D développées dans [4], [5]. Ces méthodes exploitent la structure bloc Toeplitz de la matrice d'autocorrélation. Bien qu'elles fournissent de bons résultats en présence d'un bruit additif de type blanc gaussien (BABG), les performances de ces méthodes utilisant les statistiques d'ordre deux se dégradent de manière significative en présence d'un bruit additif coloré gaussien (BACG). Ceci est dû principalement au fait que les statistiques d'ordre deux sont sensibles aux bruits additifs gaussiens (BAG) [6], [7], [8].

L'objectif de cet article est double : d'une part, il s'agit d'améliorer les performances des méthodes HR 3-D en présence d'un bruit additif coloré gaussien (BACG) en utilisant les statistiques d'ordres supérieurs. La méthode proposée est une extension de la méthode ESPRIT présenté dans [4] qui exploite les caractéristiques fréquentielles des cumulants d'ordre 4. D'autre part et pour évaluer les performances de la méthode, nous développons aussi les expressions analytiques des bornes de Cramér-Rao (BCR) asymptotiques associées aux paramètres du processus ou signal utile non bruité. Ce développement étend au cas 3-D, pour le problème considéré, les expressions des BCR 1-D et 2-D introduites dans [9], [10] et [11].

Ce travail est réalisé dans le cadre du projet soutenu par l'université UMI.

2 Modèle du signal

Soit Y un volume de données 3-D, de taille $M_1 \text{x} M_2 \text{x} M_3$, considéré comme une réalisation d'un processus stationnaire noté $\{y(m_1, m_2, m_3)\}$ pour $0 \le m_1 \le M_1 - 1,\ 0 \le m_2 \le M_2 - 1, 0 \le m_3 \le M_3 - 1$

Chaque point nommé voxel $\mathbf{m} = (\boldsymbol{m}_1, \boldsymbol{m}_2, \boldsymbol{m}_3)$ est décrit par :

$$y(\mathbf{m}) = x(\mathbf{m}, \boldsymbol{\theta}_x) + b(\mathbf{m}, \boldsymbol{\theta}_b) \quad (1)$$

Le bruit additif $b(\mathbf{m}, \boldsymbol{\theta}_b)$ est supposé coloré gaussien, indépendant statistiquement de $x(\mathbf{m}, \boldsymbol{\theta}_x)$, et définit par un vecteur de paramètre noté $\boldsymbol{\theta}_b$. Le signal utile non bruité est un processus décrit par la somme d'exponentielles complexes tridimensionnelles (SEC 3-D) suivant:

$$x(\mathbf{m}, \boldsymbol{\theta}_x) = \sum_{k=1}^{K} c_k \exp\left[j(2\pi \mathbf{m} f_k^T + \varphi_k)\right] \quad (2)$$

où les triplets $f = (f_{1k}, f_{2k}, f_{3k})$ sont les fréquences 3-D, les paramètres c_k et φ_k représentent respectivement l'amplitude à valeur réelle et la phase de la $k^{ième}$ composante harmonique. Le vecteur $\boldsymbol{\theta}_x$ contient tous les paramètres inconnus du signal $x(\mathbf{m}, \theta_x)$

$$\boldsymbol{\theta}_x = [c_1, \varphi_1, f_{11}, f_{21}, f_{31}, \cdots, c_K, \varphi_K, f_{1K}, f_{2K}, f_{3K}]^T \quad (3)$$

3 Estimation des fréquences

La méthode d'estimation HR développée dans cet article utilise les cumulants à l'ordre 4 définis pour tout $\mathbf{k}, \mathbf{h}, \mathbf{l} \in \mathbf{Z}^3$ par :

$$c_{4y}(\boldsymbol{h}, \boldsymbol{k}, \boldsymbol{l}) = cum[y(\boldsymbol{m}), y^*(\boldsymbol{m}+\boldsymbol{h}), y(\boldsymbol{m}+\boldsymbol{k}), y^*(\boldsymbol{m}+\boldsymbol{l})] \quad (4)$$

Où la fonction $(.)^*$ représente l'opérateur adjoint pour les quantités à valeurs complexes. En effet, puisque les cumulants d'ordre quatre des processus gaussiens sont nuls, les cumulants d'ordre quatre des observations coïncident théoriquement avec ceux des observations non bruitées

$$c_{4y}(\mathbf{h}) = c_{4x}(\mathbf{h}) \quad (5)$$

De plus, les cumulants d'ordre quatre du processus 3-D considéré sont décrits par une somme d'exponentielles complexes avec les mêmes fréquences recherchées [8]:

$$c_{4x}(\mathbf{h}, \mathbf{k}, \mathbf{l}) = -\sum_{k=1}^{K} c_k^4 \exp\left[j(-\mathbf{h} + \mathbf{k} + \mathbf{l}) 2\pi f_k^T\right] \quad (6)$$

L'expression (6) n'est pas directement exploitable car les cumulants d'ordre quatre sont fonction de trois variables à trois composantes donc neuf variables au total $(\mathbf{h}, \mathbf{k}, \mathbf{l}) = (h_1, h_2, h_3, k_1, k_2, k_3, l_1, l_2, l_3)$. Pour cette raison, nous considérons uniquement la "diagonale des cumulants d'ordre quatre", ne dépendant que d'une seule variable 3-D, définie de la façon suivante:

$$c_{4x}(\mathbf{h}) = c_{4x}(\mathbf{h}, \mathbf{h}, \mathbf{h}), \text{ avec } \mathbf{h} = (h_1, h_2, h_3) \quad (7)$$

Les équations (5), (6) et (7) nous montrent que la diagonale des cumulants d'ordre quatre des observations contient toute l'information utile pour estimer les fréquences 3-D (f_{1k}, f_{2k}, f_{3k}) pour $k = 1, 2, \cdots, K$:

$$c_{4y}(\mathbf{h}) = -\sum_{k=1}^{K} c_k^4 \exp\left[j 2\pi \mathbf{h} f_k^T\right] \quad (8)$$

Cette dernière relation est à la base de la méthode ESPRIT 3-D [12] fondée sur les statistiques d'ordre supérieur, rappelée dans cet article. En effet, l'équation (9) permet d'obtenir les fréquences en décomposant une matrice des cumulants $\mathbf{C}_y$ en sous-espaces signal et bruit. Cette dernière a une structure Toeplitz Bloc Bloc Toeplitz (TBBT) et se décompose sous la forme suivante :

$$\mathbf{C}_y = \mathbf{S}\boldsymbol{\Psi}(\mathbf{S})^H \quad (9)$$

où, $\boldsymbol{\Psi}$ est une matrice diagonale et $\mathbf{S}$ est la matrice de Vandermonde 3-D formée à partir des composantes fréquentielles. Sous certaines conditions sur la taille de la matrice $\mathbf{C}_y$, son rang est exactement le nombre de fréquences K Ainsi, la décomposition en éléments propres de la matrice estimée des cumulants $\hat{\mathbf{C}}_y$ permet d'accéder à la matrice de Vandermonde 3-D S et d'estimer les fréquences (f_{1k}, f_{2k}, f_{3k}) pour $k = 1, 2, \cdots, K$

4 Borne de Cramér-Rao

Dans ce paragraphe, nous allons développer l'expression analytique de la borne de Cramer-Rao asymptotique pour le vecteur des paramètres du signal utile non bruité $\boldsymbol{\theta}_x$. Pour cela on considère les hypothèses additionnelles suivantes:

A1 : la densité spectrale $S_b(f)$ du bruit additif est continue et ne présente pas de maxima localisés aux fréquences $f_k, k = 1, \cdots, K$.

A2 : les vecteurs paramètres $\boldsymbol{\theta}_x$ et $\boldsymbol{\theta}_b$ n'ont aucun élément commun.

Sous ces conditions, nous montrons tout d'abord que les BCR exactes pour un estimateur sans biais $\hat{\boldsymbol{\theta}} = [\hat{\boldsymbol{\theta}}_x, \hat{\boldsymbol{\theta}}_b]^T$ est une matrice diagonale par bloc donnée par :

$$BCR(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} BCR(\hat{\boldsymbol{\theta}}_b) & 0 \\ 0 & BCR(\hat{\boldsymbol{\theta}}_x) \end{bmatrix} \quad (10)$$

Le (k,l)[ième] élément de la BCR associée au vecteur $\boldsymbol{\theta}_x$ est donné par la relation suivante :

$$\left[BCR(\hat{\boldsymbol{\theta}}_x)\right]^{-1}_{kl} = 2\,\mathrm{Re}\left\{\frac{\partial \mathbf{x}^H(\boldsymbol{\theta}_x)}{\partial(\boldsymbol{\theta}_x)_k}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta}_b}\frac{\partial \mathbf{x}(\boldsymbol{\theta}_x)}{\partial(\boldsymbol{\theta}_x)_l}\right\} \quad (11)$$

Où $\mathrm{Re}\{.\}$ désigne la partie réelle de la quantité complexe en question et $\boldsymbol{\Gamma}_{\boldsymbol{\theta}_b}$ est la matrice d'auto corrélation du bruit. Le vecteur des observations non bruitées $\mathbf{x}(\boldsymbol{\theta}_x)$ est :

$$\begin{aligned}\mathbf{x}(\boldsymbol{\theta}_x) = [&\boldsymbol{x}(0,0,0,\boldsymbol{\theta}_x),\cdots,\boldsymbol{x}(\boldsymbol{M}_1-1,0,0,\boldsymbol{\theta}_x),\cdots,\\ &\boldsymbol{x}(0,\boldsymbol{M}_2-1,0,\boldsymbol{\theta}_x),\cdots,\boldsymbol{x}(\boldsymbol{M}_1-1,\boldsymbol{M}_2-1,0,\boldsymbol{\theta}_x),\\ &\cdots,\boldsymbol{x}(\boldsymbol{M}_1-1,\boldsymbol{M}_2-1,\boldsymbol{M}_3-1,\boldsymbol{\theta}_x)]^T\end{aligned} \quad (12)$$

Pour le problème considéré, la BCR asymptotique est donnée par la limite suivante :

$$AsBCR(\hat{\boldsymbol{\theta}}_x) = \lim_{M\to\infty} \mathbf{K}_M\, BCR(\hat{\boldsymbol{\theta}}_x)\mathbf{K}_M \quad (13)$$

où $M = M_1 M_2 M_3$ et $\mathbf{K}_M$ est une matrice de normalisation, diagonale par bloc, de taille $5K \times 5K$ définie par

$$\mathbf{K}_M = \mathbf{I}_K \otimes \mathbf{D} \quad (14)$$

avec $\mathbf{I}_K$ la matrice identité de taille K, $\mathbf{D} = diag(\sqrt{M}, \sqrt{M}, M_1\sqrt{M}, M_2\sqrt{M}, M_3\sqrt{M})$ et $\otimes$ désigne le produit de Kronecker.

En développant les dérivées du vecteur $\mathbf{x}(\boldsymbol{\theta}_x)$, nous montrons que l'expression (11) s'écrit sous la forme suivante :

$$BCR(\hat{\boldsymbol{\theta}}_x) = \frac{1}{2}\left[\mathrm{Re}\left\{\mathbf{G}^H\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta}_b}\mathbf{G}\right\}\right]^{-1} \quad (15)$$

où $\mathbf{G}$ est la matrice de taille $M \times 5K$ donnée par la concaténation des vecteurs gradients $\mathbf{g}(\mathbf{m},\boldsymbol{\theta}_x) = \dfrac{\partial \mathbf{x}(m_1,m_2,m_3,\boldsymbol{\theta}_x)}{\partial \boldsymbol{\theta}_x}$

où

$$\boldsymbol{g}(\boldsymbol{m},\theta_x) = [\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_K]^T$$

$$\boldsymbol{g}_k = [1, \boldsymbol{jc_k}, \boldsymbol{jc_k}\,2\pi\boldsymbol{m}_1, \boldsymbol{jc_k}\,2\pi\boldsymbol{m}_2, \boldsymbol{jc_k}\,2\pi\boldsymbol{m}_3]^T\,\boldsymbol{e}^{\boldsymbol{j}(2\pi\boldsymbol{m}\boldsymbol{f}_k^T+\varphi_k)} \quad (16)$$

$$\begin{aligned}\boldsymbol{G} = [&\boldsymbol{g}(0,0,0,\theta_x),\cdots,\boldsymbol{g}(\boldsymbol{M}_1-1,0,0,\theta_x),\cdots,\\ &\boldsymbol{g}(0,\boldsymbol{M}_2-1,0,\theta_x),\cdots,\boldsymbol{g}(\boldsymbol{M}_1-1,\boldsymbol{M}_2-1,0,\theta_x),\\ &\cdots,\boldsymbol{g}(\boldsymbol{M}_1-1,\boldsymbol{M}_2-1,\boldsymbol{M}_3-1,\theta_x)]^{\boldsymbol{T}}\end{aligned} \quad (17)$$

En utilisant les deux équations (11) et (15), l'équation (13) devient :

$$\boldsymbol{AsBCR}(\hat{\boldsymbol{\theta}}_x) = \frac{1}{2}\,\mathrm{Re}\left\{\left[\lim_{\boldsymbol{M}\to\infty}\mathbf{K}_{\boldsymbol{M}}\mathbf{G}^{\boldsymbol{H}}\boldsymbol{\Gamma}^{-1}_{\boldsymbol{\theta}_b}\mathbf{G}\mathbf{K}_{\boldsymbol{M}}\right]\right\}^{-1} \quad (18)$$

L'expression analytique de la matrice $\mathbf{G}^H$ obtenue à partir des équations (16) et (18), et la structure TBBT de la matrice d'autocorrélation du bruit additif $\boldsymbol{\Gamma}_{\theta_b}$, nous permet de montrer que la matrice des BCR asymptotiques est donnée par:

$$AsBCR(\hat{\boldsymbol{\theta}}_x) = \begin{bmatrix} [AsBCR]_1 & & & & \\ & \ddots & & & \\ & & [AsBCR]_k & & \\ & & & \ddots & \\ & & & & [AsBCR]_K \end{bmatrix} \quad (19)$$

où chaque sous bloc $[AsBCR]_k$ est donné en fonction de la densité spectrale du processus bruit additif $b(\mathbf{m})$ comme suit :

$$[AsBCR]_k = \frac{1}{2}\begin{bmatrix} S_b(w_k) & 0 & 0 & 0 & 0 \\ 0 & 10\frac{S_b(w_k)}{c_k^2} & -6\frac{S_b(w_k)}{c_k^2} & -6\frac{S_b(w_k)}{c_k^2} & -6\frac{S_b(w_k)}{c_k^2} \\ 0 & -6\frac{S_b(w_k)}{c_k^2} & 12\frac{S_b(w_k)}{c_k^2} & 0 & 0 \\ 0 & -6\frac{S_b(w_k)}{c_k^2} & 0 & 12\frac{S_b(w_k)}{c_k^2} & 0 \\ 0 & -6\frac{S_b(w_k)}{c_k^2} & 0 & 0 & 12\frac{S_b(w_k)}{c_k^2} \end{bmatrix} \quad (20)$$

Ainsi, pour $k = 1,\cdots,K$, les expressions des BCR asymptotiques du $\left[\hat{c}_k, \hat{\varphi}_k, \hat{f}_{1k}, \hat{f}_{2k}, \hat{f}_{3k}\right]$ sont les éléments diagonaux de la matrice $[AsBCR]_k$. On remarque que les BCR relatives aux fréquences et à la phase sont inversement proportionnelles aux rapports signal sur bruit (RSB) locale $RSB_k = c_k^2 / S_b(f_k)$.

5 Simulation numériques

5.1 Estimation des fréquences par ESPRIT 3-D dans le cas de BABG

Dans cette partie, nous allons considérer le scénario suivant : Le signal harmonique est constitué de deux modes dont les paramètres sont :

K=2, $f_1 = (f_{11}, f_{21}, f_{31}) = (0.21,0.21,0.21)$,

$f_2 = (f_{12}, f_{22}, f_{32}) = (0.22,0.22,0.22)$,

Les amplitudes des sinusoïdes complexes sont des unitaires, les phases sont uniformément distribuées dans l'intervalle $]0,\ 2\pi[$.

Le bruit additif est un iid à distribution normale de moyenne nulle et de variance σ_b^2 choisit de telle sorte que le rapport signal sur bruit soit 10 dB. La figure 1 montre la distribution de la première composante estimée $\hat{f}_{11}$ pour 100 essais Monté Carlo. La méthode utilisée est celle présentée dans [4].

5.2 Estimation des fréquences par ESPRIT 3-D dans le cas de BACG

Dans cette partie, nous allons considérer le scénario suivant : Le signal harmonique est constitué toujours de deux modes dont les paramètres sont :

K=2, $f_1 = (f_{11}, f_{21}, f_{31}) = (0.21,0.21,0.21)$,

$f_2 = (f_{12}, f_{22}, f_{32}) = (0.22,0.22,0.22)$.

Les amplitudes des sinusoïdes complexes sont choisies de telle sorte à fixer un niveau du rapport signal sur bruit à 10 dB. Les phases sont uniformément distribuées dans l'intervalle $]0,\ 2\pi[$.

Le bruit additif coloré est la sortie d'un filtre AR 3-D à support quart d'espace (quarter space region of support). L'ordre du modèle AR 3-D est (1,1,1) dont les paramètres transverses sont donnés dans le tableau 1 ci-dessous excité en entrée par un iid à distribution normale de moyenne nulle et de variance $\sigma_b^2 = 1$. La figure 2 montre la distribution de la première composante estimée $\hat{f}_{11}$ par ESPRIT 3-D en utilisant les autocorrélations [4]. La figure 3 montre les résultats en utilisant les cumulants d'ordre quatre [12].

a_{000}	a_{001}	a_{010}	a_{011}
1	-0.78	-0.23	0.1794

a_{100}	a_{101}	a_{110}	a_{111}
-0.65	0.5070	0.1495	-0.1166

Tableau 1 : les paramètres transverses du modèle AR 3-D (1, 1,1) utilisé pour générer le BACG.

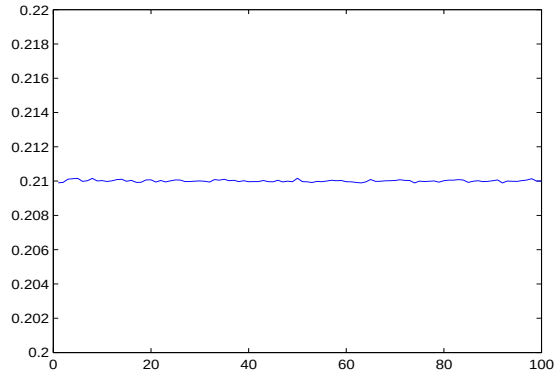


Figure 1 : *Distribution d'estimation de la composante fréquentielle 3-D* $\hat{f}_{11}$ *pour 100 Monté Carlo essais; pour un BABG en utilisant les autocorrélations.*

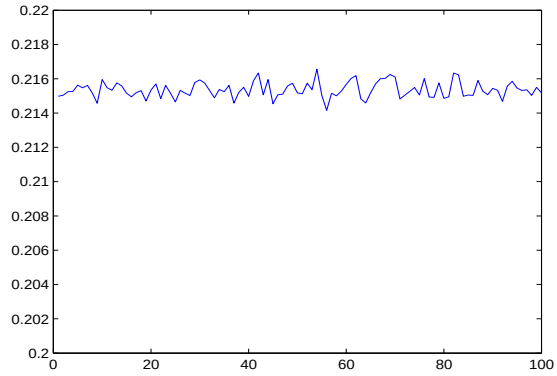


Figure 2 : *Distribution d'estimation de la composante fréquentielle 3-D* $\hat{f}_{11}$ *pour 100 Monté Carlo essais pour un BACG en utilisant les autocorrélations.*

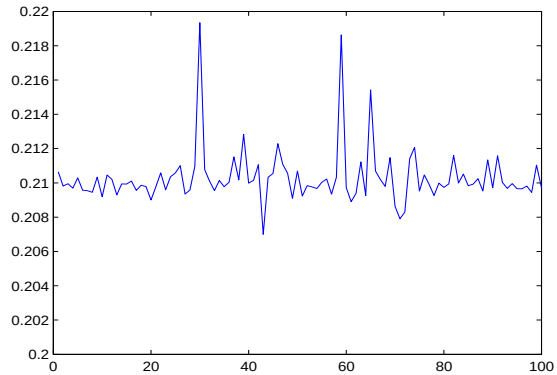


Figure 3: *Distribution d'estimation de la composante fréquentielle 3-D* $\hat{f}_{11}$ *pour 100 Monté Carlo essais pour un BACG en utilisant les cumulants d'ordre 4.*

A partir de ces figures, il apparaît que l'estimation de la composante fréquentielle 3-D $\hat{f}_{11}$ pour un BABG en utilisant les statistiques du second ordre est non biaisée

(figure 1). En outre, une fois le bruit est coloré (BACG), les autocorrélations sont inefficaces. Le recours aux statistiques d'ordre supérieur (cumulants d'ordre 4) est une bonne alternative, en effet, la moyenne de $\hat{f}_{11}$ pour la figure 3 est : 0.2103 et celle de la figure 2 est 0.2154.
L'existence des deux pics dans la figure 3 peut s'expliquer par l'échec de la méthode de mise en triplet des fréquences estimées (appariement des fréquences 3-D). Attribuer $\hat{f}_{21}$ à $\hat{f}_{11}$, va nous pousser à améliorer la méthode d'appariement automatique des fréquences .

5.3 BCR asymptotique vs EQM

Dans cette partie, nous allons considérer le deuxième scénario de la section 5.2 avec le rapport signal sur bruit virant de -15 dB à 15 dB. La taille des observations est fixé à $32\times32\times32$. La figure 4 montre la borne de Cramér-Rao asymptotique et l'erreur quadratique moyenne de l'estimée $\hat{f}_{11}$ pour 100 essais Monté Carlo avec ESPRIT 3-D en utilisant les autocorrélations et en fin la figure 5 illustre les performances de l'erreur quadratique moyenne (EQM) par rapport à la BCR asymptotique en utilisant les cumulants d'ordre quatre.

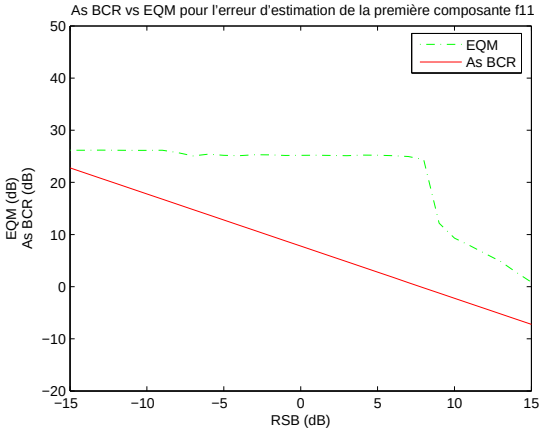


Figure 4 : *BCR asymptotique vs EQM pour l'erreur d'estimation de la composante fréquentielle 3-D* $\hat{f}_{11}$ *en utilisant l'autocorrélation* .

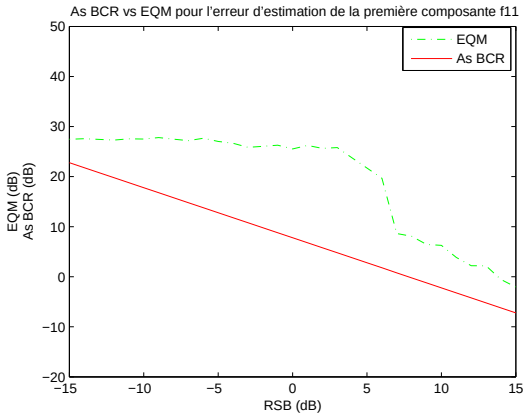


Figure 5 : *BCR asymptotique vs EQM pour l'erreur d'estimation de la composante fréquentielle 3-D* $\hat{f}_{11}$ *en utilisant des cumulants d'ordre 4.*

Pour des valeurs de RSB négatifs ou petits, (c'est-à-dire la contribution du bruit est forte), l'EQM est au dessus de la BCR asymptotique, la qualité d'estimation est mauvaise, et lorsque le rapport RSB est grand (c'est-à-dire la contribution du bruit est négligeable) l'EQM approche la BCR ainsi la qualité de l'estimateur de fréquences 3-D par la méthode ESPRIT 3-D est efficace. L'utilisation des statistiques d'ordre supérieures en l'occurrence les cumulants d'ordre quatre améliorent la qualité d'estimation par rapport à l'utilisation de l'autocorrélation, mais le prix à payer pour cela réside dans la surcharge de calcul induite. En plus, la taille de l'échantillon doit être grande (asymptotique) pour que la qualité des cumulants soit consistante, au contraire si la taille de l'échantillon est réduite, l'estimateur basé sur l'autocorrélation peut donner des résultats meilleurs que les cumulants d'ordre 4.

5.4 Conclusion

Dans ce travail, nous avons présenté une méthode d'estimation des fréquences 3-D dans le cas sinusoïdes complexes noyées soit dans un bruit additif blanc gaussien ou coloré gaussien.
Le développement des expressions théoriques de la borne de Cramér-Rao asymptotique des paramètres du modèle en particulier des fréquences 3-D est présenté.
En terme de perspective on envisage améliorer la méthode d'appariement des fréquences estimées.

Références

[1] Richard Roy, Thomas Kailath. ESPRIT : Estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoust. Speech Signal Processing*, 7(37) : 984-995, July 1989.
[2] Yingbo. Hua. Estimating two-dimensional frequencies by matrix enhancement and matrix

pencil. *IEEE Transactions on Signal Processing*, 9(40) : 2267-2280, September 1992.

[3] Stéphanie Rouquette and Mohamed. Najim. Estimation of frequencies and damping factors by 2-D ESPRIT type methods. *IEEE Transactions on Signal Processing*, 1(49) : 237-245, Jan. 2001.

[4] Brahim Aksasse, Mohamed El Ansari, Yannick Berthomieu, and Mohamed Najim. High resolution 3D spectral estimation. Dans *Proc EUSIPCO02*, vol. II, pages 391-394, Sept 3-6, 2002, Toulouse, France.

[5] Martin Haardt and Josef A. Nossek. 3D unitary ESPRIT for joint 2D angle carrier estimation. Dans *Proc. IEEE ICASSP97*, pages 255-258, April 21-24, 1997, Munich, Germany.

[6] Ananthram Swami and Jerry M. Mendel. Cumulant-based approach to the harmonic retrieval and related problem. *IEEE Trans. Acoust., Speech, Signal, Processing*, 5(39) : 1099-1109, May 1991.

[7] Hosny M. Ibrahim, Reda R. Gharieb. Estimating two-dimensional frequencies by a cumulant-based FBLP method. *IEEE Transactions on Signal Processing*, 1(47) : 262-266, January 1999

[8] Youssef Stitou, Flavious Turcu, Mohamed Najim, and Larbi Radouane. 3-D Texture characterization based on Wold decomposition and higher order Statistics. Dans *Proc IEEE ICASSP05*, vol. 2, pages 165-168, March 18-23,2005, Philadelphia, USA

[9] Petre Stoica, Andreas Jacobsson. and Jian Li, Cisoid parameter estimation in the colored noise case: Asymptotic Cramér-Rao Bound, maximum likelihood, and nonlinear least squares. *IEEE Transactions on Signal Processing*, 8(45) : 2048-2059, August 1997.

[10] Amit Mitra, and Petre. Stoica. The Asymptotic Cramér-Rao Bound for 2-D superimposed exponential signals. *Multidimensional Systems and Signal Processing*, (13) : 317-331, 2002.

[11] Josepf M. Francos. Cramér-Rao Bound on the estimation accuracy of complex-valued homogeneous Gaussian random fields. *IEEE Transactions on Signal Processing*, 3(50) : 710-724, March 2002.

[12] Youssef Stitou, Brahim Aksasse, and Mohamed Najim. Cumulant Matrix Pencil Method for three-Dimensional Frequency Estimation in colored Gaussian Noise. Dans *Proc. Second International Symposium on Communication, Control and Signal Processing (ISCCSP 2006)*, 13-15 March 2006, Marrakech, Morocco.

Benchmark de métriques de qualité sur bases de données d'images compressées

M. Nauge M.-C. Larabi C. Fernandez

Institut XLim, dépt. SIC, Université de Poitiers

Bât. SP2MI, Téléport 2, Bvd Marie et Pierre Curie, BP 30179
86962 Futuroscope Chasseneuil Cedex France

{nauge, larabi, fernandez}@sic.univ-poitiers.fr

Résumé

Il est utile de pouvoir quantifier les dégradations perçues afin de juger l'intérêt de nouveaux algorithmes de compression, tatouage ou des techniques de transmission. Nous proposons d'évaluer les performances d'un lot de métriques de qualité d'images couleurs en visant la corrélation avec le jugement humain. Le but escompté est de faciliter le choix d'une métrique parmi les nombreuses disponibles, en fournissant des scores de performance standards et exhaustifs. Pour cette étude quatre bases de données d'images sont utilisées et analysées afin d'élargir au maximum le jeu d'images tests et de mesurer la flexibilité des métriques.

Mots clefs

Métrique de qualité, système visuel humain, qualité d'expérience, évaluation objective, tests subjectifs.

1 Introduction

On note l'émergence des problématiques liées à l'augmentation de la qualité de service et de la qualité d'expérience entre l'homme et les machines. La preuve est l'explosion des recherches dans le domaine des interfaces plus intuitives, plus naturelles, plus rapides, qui nous comprennent facilement, partout, tout le temps. Il semble indispensable de disposer d'outils et de méthodologies capables de mesurer la qualité de ces nouveaux systèmes.

Pour mesurer la qualité des images numériques, il existe deux méthodes. La première consiste à effectuer des tests subjectifs, où un jeu d'images est présenté à un panel d'évaluateurs chargé de donner leurs avis sur la qualité perçue, généralement sur une échelle graduée de valeurs allant de Très mauvais à Très bonne. Pour garantir une certaine fiabilité des résultats, des recommandations strictes sont à respecter. L'ITU-R BT.500 [1] définit et formalise les contraintes pour effectuer les campagnes de tests subjectifs. Elles concernent le calibrage du système de visualisation, la distance et le positionnement de chaque observateur, leur nombre ainsi que les outils pour le dépouillement des résultats. Ces expérimentations sont délicates, coûteuses en temps et en argent. Il existe une seconde méthode pour effectuer les mesures de qualité : il s'agit des tests objectifs par l'utilisation de métriques de qualité. Il en existe 3 types : avec référence, avec référence réduite et sans référence. Les métriques avec référence utilisent l'intégralité de l'image sans dégradation pour effectuer les comparaisons. Les métriques sans référence utilisent seulement l'image dégradée pour juger la qualité. C'est une tâche facile pour l'homme mais très complexe pour une machine. Les métriques avec référence réduite extraient un minimum d'attributs de l'image sans distorsion, puis la comparaison s'effectue sur l'image dégradée avec ce minimum d'informations.

Dans cette étude, notre démarche vise à récolter un maximum de métriques de qualité et de les évaluer de manière rigoureuse sur un grand nombre d'images. Nous souhaitons faciliter le choix de métriques de qualité pour toute personne désireuse de tester l'apport de son nouvel algorithme en termes de réduction de dégradation perceptible. L'idée est également de faciliter les démarches de comparaison de nouvelles métriques de qualité en les comparant de manière équitable sur un grand nombre de bases de données d'images. Nous proposons dans une première partie de faire un tour d'horizon des métriques actuellement disponibles et utilisées pour l'étude. Leur comparaison est effectuée dans une troisième partie. Mais avant cela, une analyse des différentes bases de données d'images utilisées est détaillée dans une deuxième partie. La partie quatre est dédiée aux expérimentations suivie de ses conclusions et perspectives.

2 Description des métriques utilisées

Il existe une grande variété de métriques de qualité d'images numériques. Les travaux de Pederson [2] font état de plus de 111 métriques existantes. Depuis cette étude, nous pouvons très facilement ajouter à ce nombre plusieurs dizaines de métriques et ceci est dû à l'effervescence que connaît cette problématique. On peut les classer en 3 grandes catégories aux frontières plus ou moins floues : mathématiques, mesures pondérées par quelques propriétés du Système Visuel Humain (SVH) et la modélisation complète du SVH . Les premières métriques telles

MSE, SNR, RMSE, PSNR sont purement mathématiques, basées sur une comparaison pixel à pixel dans l'espace RVB. La métrique Peak Signal to Noise Ratio (PSNR) est la plus connue et la plus utilisée, grâce à son implémentation simple et son exécution rapide. Cependant le PSNR montre une faible corrélation avec le jugement humain sur certaines images (figure 1). De nombreux travaux visent à développer des métriques plus proches de la perception humaine.

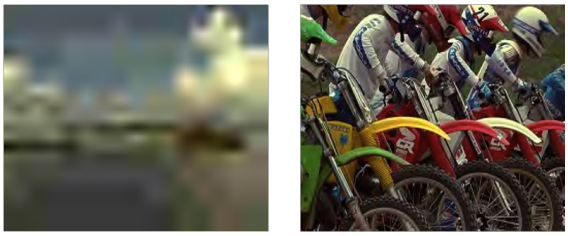


Figure 1 – *deux images avec un PSNR de 24dB*

On y retrouve les métriques VDP [3], Pdiff [4], NQM [5] qui tentent de modéliser finement le Système Visuel Humain (SVH) et le système de visualisation. Mais ces méthodes sont complexes et couteûses en temps d'exécution. La troisième catégorie de métriques tente de trouver un équilibre en effectuant des mesures pondérées par le SVH. On peut citer PSNR-HVS , PSNR-HVSM [6] qui effectuent des mesures proches de PSNR après avoir effectuées une pondération par une fonction de sensibilité aux contrastes (CSF) et un effet de masquage simplifié. Le but est de mimer les capacités changeantes de perception de stimuli par l'homme, en fonction de la fréquence, l'orientation, la couleur et la présence d'autres stimuli sur une image.
Il existe PSNR-E76, PSNR-E94 et PSNR-E00 [7] qui effectuent la comparaison pixel à pixel dans un domaine plus perceptuel. DeltaE2000 est une mesure de distance couleur à partir d'un espace CIELAB, spécialement conçue pour s'adapter aux différences de perception du SVH. Typiquement, les différences de couleurs dans les tons bleus sont moins perceptibles que les différences dans les tons verts. Les métriques S-Cielab [8], SHAME I et SHAME II [9] utilisent également une mesure de distance couleur dans l'espace CIELAB, après avoir effectuées un prétraitement de filtrage spatial, afin de considérer la distance d'observation comme un facteur agissant sur la détection d'artefact.
Il existe d'autres métriques telles qu'UQI [10], SSIM [11], MSSIM [12] et ses nombreuses dérivées qui ne prennent en considération que les changements structurels dans l'image, sans se soucier des conditions de visualisation ou des caractéristiques poussées du SVH. L'idée sous-jacente est que l'homme est naturellement très sensible aux changements de structures dans une image.
D'autres métriques exploitent des statistiques de scènes naturelles, par exemple IFC [13] ainsi que VIF et VIFP [14] qui utilisent conjointement l'analyse statistique de scène et pondération par CSF. Il existe VSNR [15] et WSNR [16] qui exploitent quelques propriétés du SVH. L'originalité réside dans l'utilisation d'ondelette pour effectuer les mesures. Etant donnée la variété des métriques, il semble intéressant de les comparer. Afin d'effectuer des comparaisons de métriques et de juger leurs corrélations avec le jugement humain, il est indispensable de disposer de bases de données d'images dont on connait les scores subjectifs d'humains. Ces scores sont les MOS (Mean Opinion Score).

3 Bases d'images utilisées

Il y a peu de bases de données d'images permettant de vérifier les performances des métriques. Ces bases sont LIVE [17], Toyama [18], IVC [19] et TID2008 [20]. Le tableau 1 présente les caractéristiques de chaque base utilisée pour l'étude.

Tableau 1 – *Caractéristiques des bases de données*

Car.	LIVE 1	Toyama	TID2008	IVC
Distors.	JPG, J2K	JPG, J2K	JPG, J2K	JPG, J2K
Codeur Jpg	Matlab imwrite	cjpeg	Non précisé	Non précisé
Codeur jp2k	Kakadu v2.2	Jasper v1.7	kakadu	Non précisé
Méthode	Single stimulus	Single stimuli	Double stimuli	Double stimuli
Image	Couleurs RGB avec 24 bits/pixels			
Résolution	entre 768x512 et 634x438	768x512	812x384	512x512
Nb img.	29	14	25	10
écran	CRT 21-inch (1024x768) (non calibrés)	CRT 17-inch (1024x768)	LCD et TFT 17 et 19 inches 1152x864	Non précisé.
Distance déobservation	2-2.5H (hauteur écran)	4H (hauteur image)	Trés varié	6H (hauteur écran)
Eclairage ambiant	Bureau	Faible	Trés varié	Normalisé
Nb. observateurs	J2K [20-25] JPG 20	16	654	15
Type observateur	Etudiants Univ.Texas	Non expert , étudiants	Non précisé	Non précisé
Ecart type et IC	Calculable	Fournis	Incalculable	Incalculable

Tout d'abord nous remarquons une différence d'exhaustivité des informations disponibles pour certaines bases. On peut critiquer l'absence d'informations relatives à l'écart type des MOS qui rend impossible certains calculs nécessaires à l'évaluation des performances de métriques.
Bien que LIVE1 ne les fournissent pas, les notes de chaque observateur sont fournies, ce qui permet de calculer les informations manquantes. Nous pouvons remarquer que les recommandations de l'ITU-R relatives aux protocoles d'expérimentation ne sont pas toujours respectées. Par exemple les distances d'observations ne sont pas assurées. Les conditions d'éclairage ne sont pas contrôlées. Les dispositifs d'affichage ne sont ni identiques ni réglés pour chaque expérimentation. Dans de telles conditions allons-nous réellement tester les performances des métriques ou les conditions de l'évaluation subjective ?
En ce qui concerne le respect des recommandations, certaines études se veulent rassurantes et montrent que les différences ne sont pas notables. La base TID2008 qui a eu recours a une très large expérimentation (3 laboratoires de

pays différents, ainsi que des dispositifs d'affichages TFT et CRT mélangés, associés à des distances d'observation et des conditions d'éclairage variées) affirme que les résultats obtenus entre chaque laboratoire sont corrélés à 97%. Une autre étude a tenté de vérifier l'impact de la différence de culture (Japon/France) ainsi que l'incidence du type d'affichage (CRT/LCD). Les résultats numériques démontrent une corrélation à plus de 95%. Donc la combinaison de toutes ces contre-indications semble être négligeable.

Abordons maintenant le choix des images et de la magnitude des distorsions. Pour tester les performances des métriques de qualité, il est important d'avoir des images variées, représentatives de la diversité des images échangées à travers le monde. Nous pouvons noter que les 3 bases LIVE, Toyama et TID2008 utilisent les mêmes images sources (12 images communes). Ce panel d'images est tout de même intéressant car il contient des images d'objets manufacturés, de visages, d'animaux, de paysages naturels, différentes prises de vue avec des premiers et arrières plans plus ou moins distincts. Mais si ce choix d'images initial n'est pas correct, les trois bases subissent le même discrédit. Bien que les images sources soient les mêmes, la magnitude des distorsions et les codeurs sont différents. Par exemple la base TID2008 a des distorsions qui rendent le contenu des images indiscernable tandis que toutes les images de Toyama restent très correctes. Tandis que LIVE propose des distorsions réparties de manière plus homogène en terme de magnitude. On se rend compte que les échelles, qui ont été proposées aux utilisateurs, n'ont pas le même sens pour une base ou pour une autre. Quand les valeurs donnent un état « Bad » pour une image de la base Toyama, il s'agit finalement d'un état « Good » de la base TID2008. Une métrique performante avec la base Toyama montre une grande justesse de mesure pour les images très peu dégradées, nécessitant une analyse très précises des dégradations. Mais si cette métrique donne de mauvais résultats avec la base Toyama qui crée d'importantes dégradations, cela sous entend qu'elle n'est pas très robuste aux importantes distorsions. L'idée est d'exploiter la complémentarité des bases de données pour juger les métriques.

Si certains sont sceptiques sur la diversité des images de ces bases de données et qu'ils espèrent trouver d'autres images avec la base IVC, il faudra être très prudent. Bien que cette base affirme respecter de manière rigoureuse le protocole d'évaluation (environnement normalisé), le choix des images sources peut laisser perplexe. Les images semblent éloignées des images actuelles. La dynamique des couleurs et la résolution des images sources sont très basses, bien en dessous des capacités d'acquisition des capteurs grand public. Il est très difficile de disposer de bases de données liant respect des protocoles, qualité de contenu, et exhaustivité des résultats (détailler tous les paramètres de l'évaluation, fournir les scores subjectifs qui ont permis le calcul du MOS etc.). Cependant il semble nécessaire et suffisant d'utiliser ces bases comme référence actuelle pour effectuer des tests de performance de métrique.

Les bases d'images sont importantes et il semble qu'une vague de prise en compte de l'aspect subjectif de la qualité des traitements déferle sur le monde scientifique. De plus en plus de laboratoires envisagent de disposer d'une salle permettant de réaliser des tests subjectifs. Nous conseillons de veiller à respecter les standards existants afin de minimiser les inquiétudes des futurs utilisateurs. Il est important de veiller à l'exhaustivité des résultats obtenus afin d'assurer une transparence des résultats et permettre plus de flexibilité des analyses. Puisqu'il est intéressant d'utiliser plusieurs bases d'images, il faut également veiller à faciliter leurs utilisations. Pour réaliser cette étude utilisant seulement 4 bases, il a fallu consacrer beaucoup de temps pour uniformiser les informations du fait de la variété des formats de fichier contenant les résultats et des différences de hiérarchie des dossiers. Dans le cas de cette étude, nous avons normalisé toutes les bases de données. Cela passe par une spécification de la hiérarchie des dossiers contenant les images, une convention de nomination de fichier, le stockage des résultats dans des formats de fichier standard et non propriétaires. Le respect des standards garantit flexibilité et interopérabilité des différents composants acteurs de tous projets.

4 Expérimentation

4.1 Procédure

Afin d'évaluer équitablement les performances des métriques de qualité, nous suivons le plan de test du VQEG [21]. La méthode consiste à disposer d'un maximum d'images dont on connait les MOS subjectifs d'un panel d'observateurs (les MOS sont obtenus en respect des recomandations de [1]). Nous utilisons les bases d'images LIVE, Toyama, IVC et TID2008 pour les dégradations jpg et jp2k. La compression JPEG introduit des effets de bloc, tandis que JPEG2000 a tendance à ajouter du flou à l'image. Ces deux types de dégradation permettent de tester la robustesse des métriques. Nous exécutons 27 métriques de qualité avec référence sur chaque image de chaque base afin d'obtenir leurs prédictions (MOSp). Les prédictions MOSp sont classées par base et par type de distorsion. Les analyses permettant de tester les métriques portent sur trois facteurs : un facteur de corrélation grâce au calcul de la corrélation de Pearson, un facteur de précision avec un calcul de racine de l'erreur quadratique moyenne (RMSE), et un facteur de cohérence par un calcul de taux de rejet (OR). L'analyse des résultats de corrélation de Pearson pour chaque métrique permet d'étudier l'existence de relation entre les valeurs MOS subjectives et les MOSp des métriques. Le plan de test du VQEG préconise d'appliquer une régression non linéaire sur chaque série de MOSp avant d'effectuer les mesures de performance. En appliquant ce prétraitement la majorité des métriques affichent des scores de corrélation proches de 99%, ce qui rend leurs comparaisons délicates. La section « 4.7 Costs and Benefits of the logistic transformation » d'une version antérieure du plan de test du VQEG [22] permet d'expliquer ces fortes

corrélations. Il est expliqué que la regression non linéaire peut introduire un gain important de corrélation quand les métriques ont une corrélation inférieure à 80% sans regression. Or beaucoup de métriques affichent des résultats sous ce palier. Les figures 2 et 4 font donc apparaître des scores de corrélation sans regression afin de ne pas introduire un biais trop significatif. Plus les valeurs sont proches de 1 ou -1 plus les résultats sont corrélés.

Pour les calculs d'OR et de RMSE, il faut mesurer les « Perror » représentant la différence entre la valeur subjective MOS et la valeur prédite MOSp. Cependant la plage de valeurs MOSp est très variable d'une métrique à l'autre. Typiquement la plage de prédictions MOSp de PSNR est de [17-45] tandis qu'elle est de [0.34-0.99] pour SSIM pour la base LIVE_jp2k. Une normalisation est appliquée sur chaque série de MOSp afin d'effectuer des comparaisons sur une échelle unique de [0.0-1.0]. L'étude du RMSE (figure 5) permet de mesurer avec précision les écarts de prédiction qui sont meilleures quand les valeurs sont proches de zéro. L'objectif de l'OR est de vérifier que les mesures sont contenues dans un intervalle acceptable. Les valeurs proches de zéro indiquent qu'il y a peu de prédictions trop éloignées des MOS subjectifs. Les écarts autorisés pour les métriques sont définis pour chaque image, et sont relatifs à la variabilité des scores des évaluateurs. Si pour une même image, les évaluateurs donnent des scores variés, l'écart autorisé pour les métriques sera important. L'objectif est de pénaliser les métriques qui donnent de mauvaises prédictions là où les observateurs sont unanimes sur la qualité d'une image. Les OR sont restreints aux bases LIVE et Toyama car elles sont les seules à fournir suffisament d'informations. L'utilisation combinée des facteurs, corrélation, RMSE et OR permet de juger les métriques avec équité et permet de faciliter la prise de décision dans le choix de métriques pour un type d'utilisation.

4.2 Résultats et discussion

Les résultats des figures 2 et 4 montrent que la métrique PSNR_HVSM a la meilleure corrélation avec le jugement humain avec des scores de 94% sur TID2008_jpeg et 96% sur TID2008_jp2k. La magnitude de distorsion de la base TID_2008 est très importante, il est donc interressant d'utiliser cette métrique dans des contextes où les images ont une plage importante de dégradation. Pour les contextes où les dégradations sont minimes, mais que l'on souhaite tout de même détecter d'infimes dégradations, il faut s'intérresser aux résultats obtenus sur la base Toyama. Pour les dégradations JPEG dans ce contexte c'est la métrique VIF qui surpasse les autres avec un score de 0.98 avec près de 5% de corrélation de plus que la seconde métrique IFC. C'est également VIF qui l'emporte avec un score de 0.949 sur Toyama_jp2k. Il semble que cette métrique se révèle la plus précise sur les 2 types de distorsions, elle est donc également assez robuste. Pour des contextes d'utilisation où une grande robustesse est attendue, il est interressant d'observer le comportement moyen des métriques sur toutes les bases de données. C'est VIF avec sont 0.849 qui est en tête suivi de près par PSNR_HVSM sur les dégradations du type JPEG. Pour les dégradations jp2k le trio de tête est composé de VIF (0.915), VIFP (0.905) et PSNR_HVSM (0.907). Dans ce cas il semble difficile de départager le 0.905 de VIFP du 0.907 de PSNR_HVSM. Bien que ces métriques soient fortement corrélées, elles n'ont pas la même précision de prédiction. On peut utiliser le RMSE pour départager les métriques ayant des scores de corrélation trop proches.

Les résultats de RMSE (figure 5) montrent que c'est la métrique VIFP qui est la plus stable sur l'intégralité des bases d'images. Les RMSE montrent également que peu de métriques ont une erreur moyenne inférieure à 20%. Le classique PSNR reste tout même assez précis sur l'ensemble des bases d'images. On remarque également que les erreurs de prédictions sont les plus importantes sur la base Toyama et IVC. Cette difficulté de précision peut s'expliquer sur la base Toyama connaissant sa faible magnitude de distorsions. Il est donc difficile d'être précis sur ces faibles dégradations. Les mesures pixel à pixel des différentes versions de PSNR sont donc efficaces malgré leur faible complexité.

Pour affiner le jugement, il est indispensable d'étudier l'OutlierRatio (figure 3) afin de vérifier que les mesures restent dans un intervalle de confiance acceptable. Cette mesure d'OR accentue de manière significative les écarts entre les différentes bases et les différentes métriques. On remarque encore une fois que la base Toyama est vraiment très exigente sur la précision des prédictions. Même la très reconnue SSIM affiche un taux de rejet proche des 70%. Alors que ce taux est en dessous des 20% sur la base LIVE. On peut donc s'interroger sur de tels écarts. La principale raison de ces forts taux de rejet vient du fait que les écarts autorisés fournis sur la base Toyama sont très restrictifs.

Finalement la modélisation complète du SVH n'est pas la solution car la très complexe HDR_VDP ne s'est jamais démarquée. Mais l'approche structurelle uniquement de la famille des SSIM n'est pas suffisante non plus. Il faut également noter que le PSNR bien que très critiquable et pouvant facilement être mis en échec se révèle en moyenne relativement performant malgré sa faible complexité. Il semble qu'ajouter peu de paramètres du SVH sur des mesures très simples augmente la qualité des prédictions. Typiquement PSNR-HVSM n'introduit qu'un filtrage CSF et un effet de masquage simplifié afin de garantir un temps de calcul limité et une robustesse aux différents types de bases de données et de distorsions.

5 Conclusion

Pour valider l'apport de nouveaux algorithmes de traitement d'images, il est nécessaire d'utiliser plusieurs métriques de qualité sur plusieurs bases de données d'images. Nous conseillons les métriques PSNR, PNSR_HVSM [23], VIFP et MSSIM [24] pour garantir une certaine variété de résultats tout en maintenant une bonne corrélation avec le

jugement humain. PSNR est utile pour sa popularité dans le monde scientifique, et son exécution rapide. PSNR_HVSM pour son intégration de composantes du SVH tout comme VIFP qui introduit en plus quelques informations statistiques de scènes naturelles. Et MSSIM pour son approche pertinente basée sur la perte de structure. Il faut pondérer ces conseils, car aucune métrique n'associe à la fois précision et robustesse. De plus ces analyses reposent uniquement sur l'étude des distorsions introduites par compression. De nouvelles distorsions ou de nouvelles métriques donneraient lieu à de nouveaux résultats. Nous souhaitons enrichir et maintenir à jour cette étude avec plus de métriques et de bases de données. Bien que ce travail soit long et fastidieux, il permet entre autre de connaître à tout instant les performances de chaque métrique par une étude standard et indépendante. Elle permet également de faciliter l'utilisation et le choix des métriques en fonction des applications visées. Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-VERS-002 dans le cadre du projet Caiman.

Références

[1] Methodology for the subjective assessment of the quality of television pictures. Rapport technique RECOMMENDATION ITU-R BT.500-11, ITU-R.

[2] Marius Pedersen et Jon Yngve Hardeberg. Survey of full-reference image quality metrics. Dans *GCIS'2009 Global Congress on Intelligent Systems*, Gjovik, Norway, Juin 2009.

[3] R. Mantiuk, K. Myszkowski, et H.-P. Seidel. Visible difference predicator for high dynamic range images. Dans *IEEE International Conference on Systems, Man and Cybernetics*, pages 2763–2769, Octobre 2004.

[4] H. Yee. A perceptual metric for production testing. *Journal of Graphics Tool*, pages 33–40, 2004.

[5] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, et A. C. Bovik. Image quality assessment based on a degradation model. Dans *IEEE transactions on image processing*, pages 636–650, 2000.

[6] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, et M. Carli.. Two new full-reference quality metrics based on hvs. Dans *Second International Workshop on Video Processing and Quality Metrics*, Scottsdale USA, 2006.

[7] G. Sharma, W. Wu, et E. N. Dalal. The ciede2000 color-difference formule : Implementation notes, supplementary test data, and mathematical observations. *Color Research and Application*, Février 2004.

[8] X. Zhang et B.A. Wandell. A spatial extension of cielab for digital color image reproduction. Dans *Soc. Inform. Display 96 Digest*, pages 731–734, San Diego, 1996.

[9] M. Pedersen et J.Y. Hardeberg. Shame : A new spatial hue angle metric for perceptual image difference. Dans *Color Research and Application*, Naples, FL, USA, Mai 2009.

[10] Z. Wang et A. C. Bovik. A universal image quality index. Dans *IEEE Signal Processing Letters*, pages 81–84, 2002.

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, et E. P. Simoncelli. Image quality assessment : From error visibility to structural similarity. Dans *IEEE Transactios on Image Processing, vol. 13, no. 4*, pages 600–612, Avril 2004.

[12] Z. Wang, E. P. Simoncelli, et A. C. Bovik. Multiscale structural similarity for image quality assessment. Dans *37th IEEE Asilomar Conference on Signals, Systems and Computers*, Novembre 2003.

[13] H. R. Sheikh. *Image Quality Assessment Using Natural Scene Statistics*. PhD thesis, University of Texas at Austin, 2004.

[14] H. R. Sheikh et A. C. Bovik. Image information and visual quality. Dans *IEEE Transactions on Image Processing*, pages 430–444, 2006.

[15] D. M. Chandler et S. S. Hemami. Vsnr : A wavelet-based visual signal-to-noise ratio for natural images. Dans *IEEE Transactions on Image Processing*, Septembre 2007.

[16] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, et A. C. Bovik. Image quality assessment based on a degradation model. *IEEE Transactions on Image Processing*, 9(4) :636–650, Avril 2000.

[17] H.R. Sheikh, Z. Wang, L. Cormack, et A.C. Bovik. Live image quality assessment database release 1. http ://live.ece.utexas.edu/research/quality.

[18] Y. Horita, Y. Kawayoke, et Z. M. Parvez Sazzad. Image quality evaluation database. http ://mict.eng.utoyama.ac.jp/mict/index2.html.

[19] P. Le Callet et F. Autrusseau. Subjective quality assessment irccyn/ivc database, 2005. http ://www.irccyn.ec-nantes.fr/ivcdb/.

[20] N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola, M. Carli, et F. Battisti. Color image database for evaluation of image quality metrics, 2008. http ://www.ponomarenko.info/tid2008.htm.

[21] Final report from the video quality experts group on the validation of objective models of multimedia quality assessement. Rapport technique PHASE I 2008, VQEG.

[22] Final report from the video quality experts group on the validation of objective models of video quality assessement. Rapport technique PHASE II 2003, VQEG.

[23] Ni. Ponomarenko. Psnrhvsm code, 2006. http ://ponomarenko.info.

[24] M. Gaubatz. Metrix mux, 2007. http ://foulard.ece.cornell.edu/gaubatz/metrix_mux.

Bdd	UQI	Shamel	IFC	MSE	SCielab	VIFP	PSNR	MSSIM	NQMRFZ3	Pdiff	PSNR_HVS	PSNR_E76	ShameII	NQM	DSSIM	VSNR	SVD	HdrVdp	PSNR_E94	VIF	PSNR_HVSM	RMSE	SSIM	PSNRLum	WSNR	SNR	PSNR_E00
Toyama	0,77	-0,463	**0,858**	-0,229	-0,24	0,753	0,356	0,794	0,389	0,709	0,621	0,445	-0,417	0,652	-0,657	**0,809**	-0,128	0,334	0,392	**0,898**	0,735	-0,259	0,621	0,434	0,437	0,296	0,384
LIVE1	**0,847**	-0,797	**0,828**	**-0,899**	-0,368	**0,935**	**0,886**	**0,884**	0,752	0,704	**0,911**	**0,885**	-0,781	**0,853**	**-0,844**	**0,923**	-0,216	0,784	**0,89**	**0,935**	**0,918**	**-0,881**	**0,909**	**0,865**	0,731	**0,888**	**0,892**
Tid2008	0,793	-0,761	0,78	**-0,916**	-0,148	**0,918**	**0,889**	**0,929**	0,73	0,643	**0,943**	**0,895**	**-0,811**	**0,869**	**-0,859**	**0,9**	-0,328	**0,831**	**0,888**	**0,932**	**0,944**	**-0,897**	**0,901**	**0,818**	0,66	**0,849**	**0,926**
IVC	**0,819**	-0,583	0,921	-0,617	-0,105	0,79	0,591	**0,821**	0,534	0,181	0,662	0,544	-0,389	0,429	-0,5	0,654	-0,257	0,63	0,565	**0,922**	0,695	-0,638	0,7	0,341	0,265	0,617	0,546
Toutes	**0,807**	-0,651	**0,847**	-0,665	-0,215	**0,849**	0,681	**0,857**	0,601	0,559	0,784	0,692	-0,599	0,701	-0,715	**0,821**	-0,232	0,645	0,684	**0,922**	**0,823**	-0,669	0,783	0,615	0,523	0,663	0,687

Figure 2 – *Scores de corrélations sur images dégradées par JPEG*

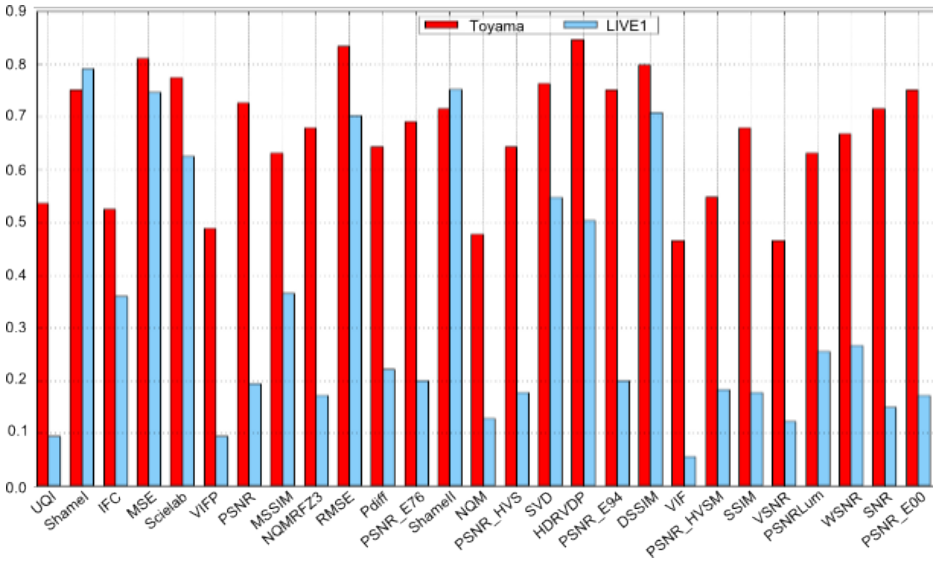


Figure 3 – *Scores des OR sur images dégradées par JPEG*

Bdd	UQI	Shamel	IFC	MSE	SCielab	VIFP	PSNR	MSSIM	NQMRFZ3	Pdiff	PSNR_HVS	PSNR_E76	ShameII	NQM	DSSIM	VSNR	SVD	HdrVdp	PSNR_E94	VIF	PSNR_HVSM	RMSE	SSIM	PSNRLum	WSNR	SNR	PSNR_E00
Toyama	0,631	-0,701	0,786	-0,788	0,058	**0,927**	**0,843**	**0,825**	0,614	0,767	**0,866**	0,778	-0,658	**0,84**	-0,782	**0,912**	-0,179	0,552	0,797	**0,949**	**0,893**	**-0,809**	**0,847**	**0,878**	0,673	0,787	0,767
LIVE1	**0,822**	-0,522	0,816	-0,793	-0,301	**0,907**	**0,843**	**0,8**	0,78	0,672	**0,864**	0,787	-0,575	**0,871**	-0,726	**0,903**	-0,199	0,737	**0,82**	**0,903**	**0,883**	-0,794	**0,857**	**0,866**	0,79	0,743	0,795
Tid2008	**0,913**	-0,762	**0,82**	**-0,842**	0,172	**0,941**	**0,866**	**0,936**	**0,884**	0,64	**0,959**	**0,821**	-0,751	**0,934**	**-0,8**	**0,93**	-0,252	**0,925**	**0,847**	**0,916**	**0,96**	**-0,835**	**0,864**	**0,866**	**0,837**	**0,831**	0,89
IVC	**0,8**	-0,649	**0,882**	-0,709	-0,052	**0,848**	0,77	**0,77**	0,755	0,363	**0,875**	0,649	-0,528	0,673	-0,548	**0,803**	-0,091	0,69	0,708	**0,893**	**0,893**	-0,731	0,76	0,661	0,633	0,734	0,678
Toutes	0,791	-0,659	**0,826**	-0,783	-0,031	**0,905**	**0,83**	**0,833**	0,758	0,611	**0,891**	0,758	-0,628	**0,829**	-0,714	**0,887**	-0,18	0,726	0,793	**0,915**	**0,907**	-0,792	**0,832**	**0,818**	0,733	0,774	0,783

Figure 4 – *Scores de corrélations sur images dégradées par jp2k*

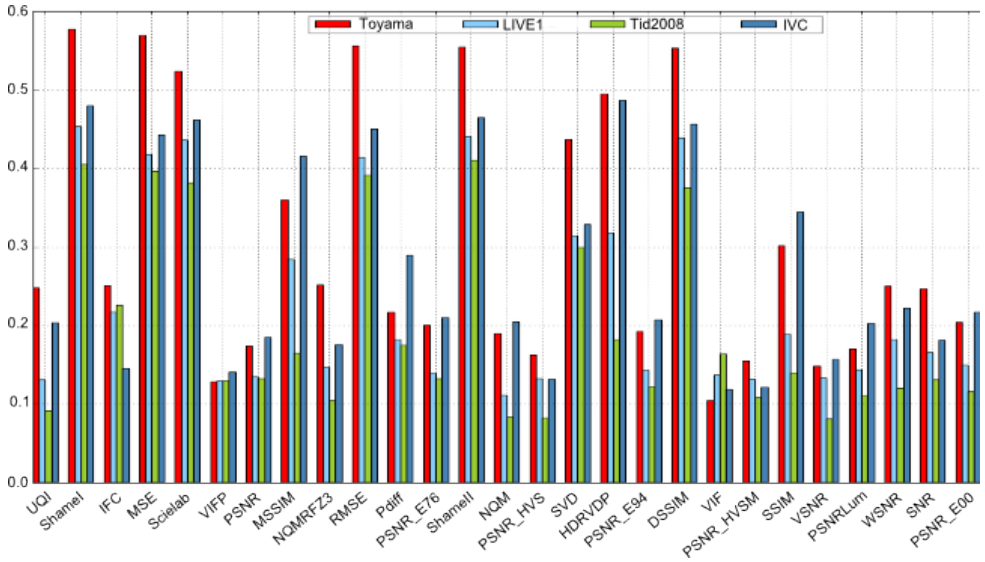


Figure 5 – *Scores de RMSE sur images dégradées par jp2k*

Descripteurs visuels robustes pour l'identification de locuteurs dans des émissions télévisées de talk-shows

Félicien Vallet[1,2] Slim Essid[1] Jean Carrive[2] Gaël Richard[1]

[1]Télécom ParisTech CNRS/LTCI
46 rue Barrault,
75634 Paris cedex 13,
France
{firstname.lastname}@telecom-paristech.fr

[2]Institut national de l'audiovisuel
4 avenue de l'Europe,
94366 Bry-sur-Marne cedex,
France
{fvallet, jcarrive}@ina.fr

Résumé

Dans cet article, nous proposons une nouvelle méthode multimodale pour l'identification de locuteurs dans une émission de talk-show à base de machines à vecteurs supports. Notre étude met en évidence l'efficacité de descripteurs visuels spécifiques pour ce type de contenu vidéo, résultat de l'assemblage par un réalisateur des images prises par plusieurs caméras. Ces descripteurs sont motivés par des connaissance a priori sur l'approche suivie par le réalisateur dans la sélection des plans appropriés. Leur utilisation conjointement aux MFCCs (coefficients cepstraux à fréquence Mel) montre une amélioration significative du score d'identification (+8%) par rapport à plusieurs systèmes de référence dont un système audio standard.

Mots clefs

Système multimedia, analyse d'image, identification de locuteurs.

1 Introduction

L'Institut national de l'audiovisuel (Ina) a pour principales missions de regrouper, conserver et mettre à disposition les archives de la radio et de la télévision française. La classification et la segmentation automatique de ses contenus sont des étapes essentielles pour l'Ina dans de nombreux domaines tels que la recherche et l'indexation de vidéos, la structuration ou la génération automatique de résumés. Dans les scènes audiovisuelles, comme celles proposées dans un talk-show télévisé, le but d'une segmentation automatique est d'obtenir des segments pertinents tels que « passage musical », « extrait de film », « invité principal à l'écran » ou « arrivée d'un nouvel invité ».

De ce point de vue, l'identification de locuteurs est une tâche importante pouvant contribuer à une segmentation sémantique. Notre but est l'identification de locuteurs a priori inconnus. Il n'y a pas de données disponibles pour chaque locuteur pouvant être utilisées pour apprendre des classifieurs de façon totalement supervisée. Par conséquent, notre approche peut être considérée « semi-supervisée » en ce sens que les exemples sont collectés en ligne en demandant à l'utilisateur de notre système, par exemple un documentaliste de l'Ina traitant un talk-show, de sélectionner arbitrairement un court (4 à 15 s) extrait vidéo de chaque locuteur intervenant dans le talk-show. Ces courts extraits (un unique segment par locuteur) sont ensuite utilisés pour l'apprentissage de classifieurs pour l'identification de locuteurs sur l'intégralité de l'émission (environ 3 h). Il est important de noter que la sélection manuelle d'extraits par l'utilisateur est plus simple qu'il n'y paraît. Tout d'abord, le nombre total de locuteurs à chercher est connu a priori grâce à une notice, disponible pour chaque talk-show, détaillant la liste des invités. Ensuite, l'utilisateur peut rapidement localiser les extraits des différents locuteurs en utilisant le *slider* temporel et/ou le mode avance rapide.

Les systèmes de segmentation en locuteurs (*speaker diarization*) [1] sont des méthodes alternatives à notre approche. Cependant, ceux-ci étant totalement non-supervisés, l'utilisateur aura toujours à associer un locuteur à chaque cluster proposé avec l'inconvenient que certains locuteurs puissent être introuvables, ces systèmes n'étant pas parfaits. Par conséquent, notre approche ne semble pas plus coûteuse en termes d'efforts pour l'utilisateur.

Dans cet article, l'accent est mis sur l'utilisation de descripteurs visuels robustes augmentant l'ensemble d'attributs audio habituellement utilisé pour cette tâche, i.e. les coefficients cepstraux à fréquence Mel (MFCCs) [2], afin d'améliorer le taux de détection. Plusieurs études dans le domaine de la segmentation en locuteurs ont proposé des approches multimodales [3, 4]. D'autres travaux dans le champ multimodal ont montré l'intérêt de mesurer la synchronie audio-visuelle pour la détection du locuteur actif [5, 6, 7]. Cependant, dans chaque cas les auteurs ont utilisé des bases de données très particulières : corpus de news (National Institute of Standard and Technology Rich Transcription : NIST RT[1]), vidéos de meetings en mode multi-camera (Augmented Multi-party In-

[1]http ://www.nist.gov/index.html

teraction : AMI[2]), vidéos de locuteurs prononçant des nombres (Clemson University Audio Visual Experiments : CUAVE[3]), etc. Le contenu exploité dans cette étude a la particularité d'être monté, un réalisateur de télévision sélectionnant des plans pris par plusieurs cadreurs. De plus, ces plans présentent de grandes variations de cadrage et d'angles de vue. De fait, nous proposons une nouvelle approche d'identification des locuteurs mettant en évidence l'utilité de combiner des attributs audio classiques comme les MFCCs avec des descripteurs visuels spécifiques et robustes sur les talk-show télévisés fournis par l'Ina.
Dans la section 2 de cet article, nous détaillons les motivations et l'extraction des descripteurs visuels et en section 3, nous rappelons le principes des machines à vecteur support (SVM) utilisées pour la classification. Dans la section 4, nous présentons notre étude expérimentale et discutons des résultats avant de donner des perspectives pour la suite de ce travail dans la section 5.

2 Extraction de descripteurs visuels robustes

2.1 Motivations pour la conception des descripteurs

Nous traitons un type de contenu particulier : le talk-show télévisé. Contrairement aux bases de données utilisées en biométrie et segmentation en locuteurs, le contenu vidéo est monté, c'est-à-dire que plusieurs caméras sont utilisées pour le tournage et qu'à chaque instant, les images d'une seule sont diffusées. Le plan est choisi par le réalisateur qui, généralement, essaie de suivre le locuteur à l'écran. Par conséquent, bien que le cadrage varie (de plan large à gros plan), la plupart du temps « on voit qui l'on entend ». Comme décrit plus bas, nous extrayons des descripteurs visuels à partir des images des personnes à l'écran, en supposant qu'elles sont les locuteurs actifs. Il est important de mentionner que, les émissions étant réalisées en direct, notre tâche est compliquée par les conditions sonores bruyantes et la grande spontaneité de parole des intervenants.

2.2 Suivi de visage et de costume

Dans notre étude, les mouvements de caméra, le cadrage, les angles de vue et les conditions variables d'éclairage rendent très délicate l'utilisation d'un détecteur de visage comme, par exemple, dans [8]. Afin que le téléspectateur ne confonde pas les participants présents sur un plateau télévisé, leurs costumes sont en général précautionneusement choisis. Notre hypothèse étant qu' « on voit qui l'on entend », nous supposons que l'information portée par le costume peut aider à l'identification des locuteurs et présente l'avantage de pouvoir être extraite de façon plus robuste que les attributs de visage. Cette hypothèse est renforcée par l'importante corrélation entre couleur dominante du costume et tours de locuteurs comme présenté dans la Figure 1.

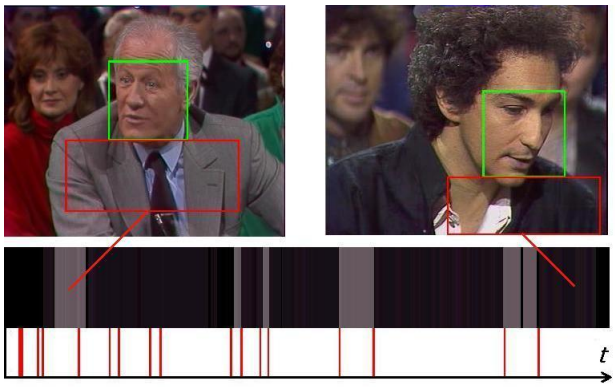


Figure 1 – *Couleur dominante du costume et tours de locuteur (en rouge) pour un segment de parole de 2min*

Nous décidons donc, comme dans [9], d'utiliser le costume comme attribut pour détecter automatiquement l'apparition d'une personne à l'écran. La présence d'un individu est attestée par la recherche de visages dans chaque trame. Nous utilisons l'algorithme de Viola et Jones [10] (disponible dans la librarie OpenCV [11]) pour détecter les visages. Nous déterminons ensuite les régions des costumes en traçant des rectangles sous les visages détectés comme dans [9].
Deux classifieurs sont utilisés pour détecter les visages de face et de profil pour chaque image. De plus, nous limitons le nombre de détections dans chaque trame et indiquons une taille minimale pour les visages, pour ne garder finalement que la plus grande des régions d'intérêt proposées. Ce processus doit empêcher, autant que possible, la détection de visages dans le public (à l'arrière plan et donc souvent plus petits).
Cependant, la détection de visages trame à trame introduit de nombreuses fausses alarmes et non-détections. Pour s'affranchir de ce problème nous utilisons une heuristique simple exploitant les propriétés temporelles de la vidéo. Ainsi, nous proposons de faire l'économie d'un suivi de visage puisqu'une analyse de flux optique est effectuée pour l'extraction de descripteurs de mouvement (cf Section 2.4).
Après avoir implémenté un détecteur de changement de plan (*cuts*) basé sur l'intersection d'histogrammes de couleur [12], nous extrayons des points d'intérêt sur les rectangles contenant visages et costumes avec l'algorithme de Shi et Tomasi [13]. Ensuite, ces points d'intérêt sont suivis entre deux *cuts* grâce à l'algorithme de Lucas et Kanade [14].
Le suivi est initialisé avec un nombre maximum de 300 points d'intérêt au temps t_s (correspondant à la première trame contenant un visage après le dernier *cut* détecté) et arrêté au temps t_e, soit à la fin du plan courant ou avant le *cut* suivant si plus d'un tiers des points suivis sont perdus entre deux trames (indiquant généralement la présence d'un changement de plan non-détecté). Cette procédure est répétée tout le long de l'émission et les erreurs de détection

[2]http ://corpus.amiproject.org
[3]http ://www.ece.clemson.edu/speech/cuave.htm

de visage sont corrigées entre t_s et t_e de la manière suivante.

Si une trame f ne présente pas de visages détectés entre t_s et t_e, les regions d'intérêt du visage et du costume de la trame $f-1$ sont utilisées pour déduire celles manquantes par translation (déduites du déplacement des points d'intérêt suivis). De plus, si entre t_s et t_e le déplacement entre les trames f et $f+1$ des régions d'intérêt est trop important par rapport à celui des points d'intérêt, nous supposons que le visage détecté à $f+1$ est différent de celui détecté à f. Il correspond souvent à un visage au second plan qui peut ponctuellement devenir prépondérant pour la trame $f+1$ à cause de variations du cadrage. Par conséquent, les régions d'intérêt correspondantes pour le mauvais visage et le mauvais costume sont supprimées et d'autres valides sont créées par translation de celles de la trame f comme expliqué précédemment. Le procédé d'extraction proposé est résumé dans la Figure 2.

Bien qu'une évaluation directe de la procédure précédente soit délicate en raison de l'absence de vérité terrain appropriée, il est important de noter que celle-ci est implicitement évaluée par les descripteurs déduits par la suite et utilisés pour l'identification de locuteurs (voir Section 4).

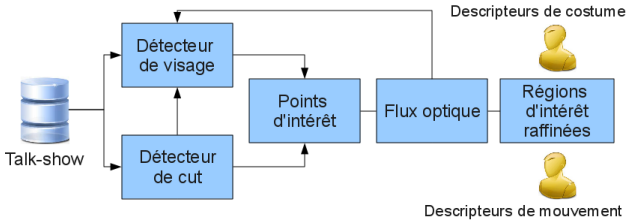


Figure 2 – *Procédé d'extraction des descripteurs*

2.3 Descripteurs de couleur du costume

Une fois obtenue une description robuste des occurrences des costumes des personnes à l'écran dans une vidéo, nous proposons deux attributs calculés sur les régions d'intérêt correspondantes et reposant sur le MPEG-7 Dominant Color Descriptor [15]. Celui-ci propose une description compacte de 1 à 8 couleurs dominantes pour une image ou région d'intérêt, une couleur dominante étant un vecteur de trois composantes RGB $[x_R, x_G, x_B]$, chacune quantifiée sur 32 bins.

Le premier descripteur calculé est basé sur l'association des deux couleurs principales, c'est-à-dire les deux couleurs avec la plus grande proportion P_i de pixels, où i est l'indice des couleurs dominantes pour l'image. Dans le cas où une seule couleur dominante est détectée, celle-ci est dupliquée. Notre second descripteur est la couleur dominante moyenne qui est une moyenne pondérée des n couleurs recouvrant au moins 40% des pixels de la region d'intérêt selon :

$$x_{C_{avg}} = \frac{\sum_{i=1}^{n} P_i . x_{C_i}}{\sum_{i=1}^{n} P_i} \quad (1)$$

avec

$$\sum_{i=1}^{n} P_i \geq 40\% \quad (2)$$

x_{C_i} étant la valeur de la i$^{\text{ème}}$ couleur dominante. Ce ratio de 40% s'est avéré être une estimation robuste de la surface couverte par le costume dans la region d'intérêt, en prenant en compte les conditions potentiellement bruitées survenant par exemple lorsque les mains du locuteur entrent dans le rectangle englobant le costume.

2.4 Descripteurs de mouvement du locuteur

Inspiré de [4], nous proposons de calculer plusieurs descripteurs de mouvement basés sur l'analyse du flux optique. Nous supposons que chaque locuteur possède ses propres gestuelles et expressions qui, décrites avec les bons attributs, peuvent être très discriminantes (par exemple le mouvement des mains souvent visible dans la région d'intérêt du costume). Pour caractériser ces particularités de façon robuste et efficace, nous proposons de déduire du flux optique des descripteurs de mouvement pour l'image globale, ainsi que pour les régions d'intérêt du visage et de la poitrine (qui est la même que celle du costume) comme montré dans la Figure 3 a).

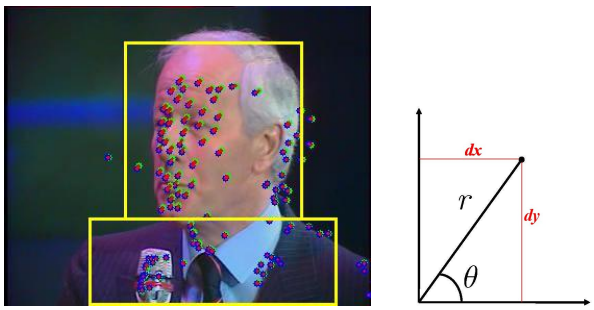


Figure 3 – *a) Mouvement entre deux trames consécutives pour une sélection de points d'intérêt. b) Amplitude de mouvement r et orientation θ pour un déplacement (dx, dy)*

Les amplitudes et orientations de vitesse et d'acceleration sont calculées comme les dérivées première et seconde des points d'intérêt de l'image globale et des régions d'intérêt, celles-ci étant évaluées comme les coordonnées r et θ dans un système polaire. Nous proposons également de calculer l'amplitude relative présentée comme le rapport des amplitudes pour les régions d'intérêt du visage et de la poitrine sur celle de l'image globale. Le Tableau 1 récapitule les différents descripteurs proposés dans cette section et donne leurs acronymes et dimensions.

Acronyme	dim	Description
DomColMoy	3	couleur dominante moyenne du costume
DomCol2	6	2 couleurs dominantes principales du costume
OrientVit	3	Orientation de vitesse pour image / visage / costume
OrientAcc	3	Orientation d'acceleration pour image / visage / costume
IntVit	5	Amplitude de vitesse absolue et relative pour image / visage / costume
IntAcc	5	Amplitude d'acceleration absolue et relative pour image / visage / costume

Tableau 1 – *Description du set d'attributs proposé*

2.5 Interpolation pour trames sans visage

Pour un certain nombre de trames, aucun visage n'est détecté à cause des choix de conception présentés dans la Section 2.2. Par conséquent, afin d'obtenir des descripteurs continus, nous interpolons les descripteurs sur les trames sur lesquelles aucun visage n'a été détecté. Cela est réalisé en choisissant aléatoirement pour ces trames la valeur des descripteurs précédents ou suivants. Cette stratégie peut sembler assez approximative mais donne de bons résultats (voir Section 4).

3 Classification SVM

En raison de leur efficacité à résoudre un grand nombre de problèmes de classification, les classifieurs SVM sont devenus très populaires dans de nombreux domaines de recherche. Nous conseillons au lecteur de se référer aux ouvrages de référence [16] et ne rappelons ici succintement que les principes de base.

Dans les problèmes bi-classes, l'algorithme d'apprentissage SVM recherche l'hyperplan $\mathbf{w}.\mathbf{x} + b = 0$ qui sépare les exemples d'apprentissage $\mathbf{x}_1, ..., \mathbf{x}_n$ assignés aux classes $y_1, ..., y_n$ ($y_i \in \{-1, 1\}$) tel que $y_i(\mathbf{x}_i.\mathbf{w} + b + \xi_i) - 1 \geq 0, \forall i$, sous la contrainte que la distance $\frac{2}{||\mathbf{w}||}$ entre l'hyperplan et les plus proches exemples soit maximale, ξ_i étant des variables d'écart positives prenant en compte les éventuels points abérrants ou *outliers*.

Lors de la résolution de ce problème d'optimisation la somme des ξ_i est pénalisée par un facteur de coût C (à définir) afin de contrôler le nombre total d'*outliers* autorisés. Il est possible d'utiliser différents facteurs de coût C_+ et C_-, respectivement associés aux classes positives et négatives, dans le cas d'ensembles d'apprentissage déséquilibrés et cela afin d'éviter que la solution ne soit biaisée par la sur-représentation d'une classe par rapport à l'autre [17].

Les données n'étant pas linéairement séparables dans l'espace des d descripteurs initiaux, une fonction noyau $k(\mathbf{x}, \mathbf{y})$ peut être utilisée pour projeter les données dans un espace de beaucoup plus grande dimension dans lequel les deux classes deviennent linéairement séparables. Un vecteur de la base de test est alors classé suivant le signe de la fonction $f(\mathbf{x}) = \sum_{i=1}^{n_s} \alpha_i y_i k(\mathbf{s}_i, \mathbf{x}) + b$, où $\mathbf{s}_i$ sont les vecteurs supports, α_i les multiplicateurs de Lagrange, et n_s le nombre de vecteurs supports. Dans cette étude nous utilisons le noyau gaussien $k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{||\mathbf{x}-\mathbf{y}||^2}{2d\sigma^2}\right)$.

Nous testons également les performances obtenues à l'aide de SVM transductifs (TSVM) [18]. Dans ce cas, en plus de l'ensemble d'exemples étiquetés utilisé avec le SVM, on fournit au classifieur un ensemble d'exemples dont l'étiquette est non renseigné. L'algorithme alloue ensuite de façon itérative une étiquette à chacun de ces exemples. En plus des paramètres C et σ, l'utilisateur doit spécifier la proportion p des données de test à assigner à chaque classe.

4 Etude expérimentale

4.1 Description

L'évaluation des descripteurs visuels proposés est effectuée sur une émission de 3.5 h, appartenant au corpus « Le Grand Échiquier » (plus de 50 émissions de talk-show des années 1980). Cette base de données présente des caractéristiques qui l'ont rendue célèbre parmi plusieurs projets européens et nationaux. Chaque émission est consacrée à un invité principal entouré d'invités secondaires. Les interviews sont ponctuées d'extraits de film, de passages musicaux ou de scènes de théâtre. Les données sont disponibles au format MPEG2 et la piste audio est rééchantillonnée à 12.8 kHz en raison d'une largeur de bande très étroite dans les fichiers originaux. Seules les sections de parole sont traitées, représentant plus d'1.5 h.

Comme le Tableau 2 l'indique, les tours de locuteurs à identifier sont de longueurs très variables. Il est également bon de noter que deux personnes se partagent environ 70% du temps total de parole : le présentateur et l'invité principal.

temps de parole total	5745 s
nombre de locuteurs	10
nombre de sections de parole	64
nombre de tours de locuteurs	1048
durée moyenne d'un tour	5.3 s
écart type de la durée d'un tour	7.1 s
tour de parole le plus long	92 s
tour de parole le plus court	0.2 s

Tableau 2 – *Caractéristiques des données audio*

La piste audio est extraite du fichier vidéo, convertie en mono en moyennant canaux droit et gauche puis sous-échantillonnée. Ensuite, nous extrayons les 13 premiers coefficients MFCCs, en incluant le coefficient cepstral d'ordre 0. Les descripteurs vidéo étant calculés toutes les 40 ms (25 trames/sec), nous effectuons une intégration temporelle, 4 trames audio consécutives étant moyennées afin que descripteurs audio et vidéo soient disponibles à la même fréquence (25 Hz). Nous supposons également qu'elles peuvent être considérées comme synchrones (ce qui sera discuté par la suite). Les vecteurs de descripteurs sont finalement formés par simple concaténation.

En plus des descripteurs visuels robustes décrits dans la Section 2, nous extrayons l'histogramme de couleur YUV et le MPEG-7 ColorLayout Descriptor [15] sur chaque trame pour servir d'attributs de référence.
La sélection des courts extraits d'apprentissage par le documentaliste est simulée par le choix aléatoire pour chaque locuteur de segments longs de 4 à 15 s. La classification SVM, effectuée en mode *un contre un*, est réalisée à l'aide de la toolbox *SVM^{light}* développée dans [19] avec un noyau gaussien. Le paramètre σ est fixé pour toutes les expériences par 5 validations croisées sur la base d'apprentissage tandis que les facteurs de coût C_-/C_+, traitant des déséquilibres à l'intérieur de l'ensemble d'apprentissage, sont fixés par le rapport du nombre d'exemples positifs sur celui d'exemples négatifs. 100 validations croisées sur les segments de parole candidats assurent la validité des résultats, c'est-à-dire que le tirage aléatoire des segments d'apprentissage et de test est effectué 100 fois.
Les résultats présentés dans la section 4.2 sont obtenus à l'aide de la fonction *detection error rate* (DER) utilisée pour les évaluations NIST Rich Transcription (RT). Cette fonction est spécialement conçue pour la tâche de classification de locuteur. Cependant, en raison des spécificités de notre corpus, nous proposons d'utiliser parallèlement une nouvelle métrique. En effet, la répartition du temps de parole étant très déséquilibrée entre les n locuteurs, l'identification correcte des deux principaux suffirait à l'obtention de bons résultats avec la fonction NIST RT. Par conséquent, la métrique que nous proposons pondère le volume de parole de chaque personne de sorte qu'elle est plus sensible à la correcte identification d'un locuteur qu'à la longueur totale de ses interventions :

$$\text{New DER} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{nombre de trames mal-classées pour le locuteur } i}{\text{nombre total de trames pour le locuteur } i} \quad (3)$$

4.2 Résultats et discussion

Les résultats du Tableau 3 sont obtenus après intégration temporelle de la sortie des classifieurs SVM sur des fenêtres de 0.5 s avec un recouvrement de 50%. Ils montrent qu'une amélioration significative est obtenue avec notre ensemble de descripteurs visuels robustes en comparaison avec un système basé sur les MFCCs uniquement. De plus, les performances de ces descripteurs dépassent largement celles des associations d'attributs visuels classiques et de MFCCs. En fait, ces dernières dégradent fortement le score de référence obtenu avec les MFCCs. Ceci peut être expliqué par l'ajout d'informations bruitées dues au manque de focus des descripteurs visuels calculés globalement. L'addition de la couleur dominante moyenne du costume entraîne une amélioration de +8%, validant l'hypothèse de caractérisation d'un locuteur par son costume.
Les attributs visuels proposés attestent d'une grande robustesse. En effet, ils fonctionnent de façon collaborative avec les MFCCs, en assurant une grande discrimination au niveau local mais également en permettant aux MFCCs de dominer la description grâce à leur plus grande dimension en cas de non-détection de visage ou d'interpolation incorrecte. Nous remarquons également que l'ensemble d'attributs donnant le meilleur score NIST DER ne donne pas le meilleur score avec la métrique que nous proposons. Cela était attendu, sachant que certains locuteurs parlent moins de 10 s alors que d'autres interviennent plus de 30 mn.

Ensemble d'attributs	NIST DER	New DER
MFCC-DomColMoy-OrientVit	29.5	46.0
MFCC-DomColMoy-IntVit	27.9	44.5
MFCC-DomColMoy-OrientAcc	27.4	44.3
MFCC-DomColMoy-IntAcc	28.7	42.4
MFCC-DomColMoy	27.5	43.4
MFCC-YUVHistCol	52.1	53.7
MFCC-ColorLayout	59.8	63.6
MFCC	35.4	52.1

Tableau 3 – *Erreur de classification pour différents sets d'attributs*

La synchronisation des descripteurs audio et vidéo ne s'avère pas être critique ici. En effet, une étude préliminaire nous a montré une relative insensibilité du système de classification à l'introduction d'un retard entre modalité audio et vidéo.

Enfin, l'utilisation de TSVM n'a pas permis une amélioration des résultats de classication. Alors qu'il n'y a pas de raisons qu'un classifieur TSVM correctement calibré se comporte plus mal qu'un SVM classique, le paramètre p, relatif à la répartition du temps de parole pour chaque locuteur peut s'avérer trop restrictif pour notre approche. En effet, nous ne pouvons attendre de l'utilisateur de notre système qu'il soit capable de faire d'importantes suppositions quant au temps de parole de chaque locuteur en raison de la structure a priori inconnue de l'émission.

5 Conclusion et perspectives

Nous avons montré que l'extraction de descripteurs visuels basés sur le mouvement et la couleur dominante du costume de la personne à l'écran améliore significativement le taux d'identification de locuteurs dans des émissions de talk-show. Combinés avec des attributs classiques MFCCs, ces deux ensembles assurent une très bonne discrimination entre locuteurs. Ils s'avèrent d'ailleurs être particulièrement efficaces pour le traitement d'émissions télévisées. Ce type de réalisation et le déséquilibre entre les interventions de chaque locuteur rendent cette tâche particulièrement difficile. Dans un travail futur, nous combinerons notre approche avec un système de segmentation en locuteurs pour réduire encore l'intervention humaine dans le processus d'identification de locuteur.

Références

[1] S.E. Tranter et D.A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14 (5) :1557–1565, 2006.

[2] Lawrence Rabiner et Biing-Hwang Juang. *Fundamentals of speech recognition*. Prentice Hall PTR, 1993.

[3] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, et R. Kasturi. Audio segmentation and speaker localization in meeting videos. Dans *International Conference on Pattern Recognition*, 2006.

[4] Gerald Friedland, Hayley Hung, et Chuohao Yeo. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. Dans *International Conference on Acoustics, Speech and Signal Processing*, 2009.

[5] Harriet J. Nock, Giridharan Iyengar, et Chalapathy Neti. Speaker localisation using audio-visual synchrony : An empirical study. Dans *International conference on image and video retrieval*, 2003.

[6] Malcolm Slaney et Michele Covell. Facesync : A linear operator for measuring synchronization of video facial images and audio tracks. Dans *NIPS*, 2000.

[7] John Hershey et Javier Movellan. Audio-vision : Using audiovisual synchrony to locate sounds. *Advances in Neural Information Processing (MIT Press)*, 1999.

[8] Ming-Yu Chen et Alexander Hauptmann. Searching for a specific person in broadcast news video. Dans *International Conference on Acoustics, Speech and Signal Processing*, 2004.

[9] Gaël Jaffré et Philippe Joly. Costume : A new feature for automatic video content indexing. Dans *International conference on Adaptivity, Personalization and Fusion of Heterogeneous Information*, 2004.

[10] Paul Viola et Michael Jones. Robust real-time object detection. Dans *International Workshop on Statistical and Computational Theories of Vision-Modeling, Learning, Computing and Sampling*, 2001.

[11] G. Bradski et A. Kaehler. *Learning OpenCV : Computer Vision with the OpenCV Library*. O'Reilly Media, 2008.

[12] Michael J. Swain et Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, 7 :11–32, 1991.

[13] J. Shi et C. Tomasi. Good features to track. Dans *Conference on Computer Vision and Pattern Recognition*, 1994.

[14] Bruce Lucas et Takeo Kanade. An iterative image registration technique with an application to stereo vision. Dans *International Joint Conference on Artificial Intelligence*, 1981.

[15] B.S. Manjunath, Philippe Salembier, et Thomas Sikora, éditeurs. *Introduction to MPEG-7 - Multimedia Content Description Interface*. Wiley, 2002.

[16] Bernhard Scholkopf et Alexander J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT-Press, 2001.

[17] K. Morik, P. Brockhausen, et T. Joachims. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. Dans *International Conference on Machine Learning*, 1999.

[18] Thorsten Joachims. Transductive inference for text classification using support vector machines. Dans *International Conference on Machine Learning*, 1999.

[19] Thorsten Joachims. *Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning*. MIT-Press, 1999.

Optimization design of orthogonal filter banks for image coding via multi-objective genetic algorithm

A. Boukhobza[1] *A. Taleb Ahmed* [2] *N. Taleb* [3] *A. Bounoua* [3]

[1] Department of Electronics, UHBC University, Chlef, Algeria
[2] LAMIH CNRS FRE 3304, UVHC University, Valenciennes, France.
[3]RCAM Laboratory of research, UDL University, Sidi bel abbes, Algeria

aek_boukhobza@yahoo.ca, abdelmalik.taleb-ahmed@univ-valenciennes.fr,
ne_taleb@univ-sba.dz, nacerbounoua@yahoo.fr

Abstract

In this paper, we propose an optimisation method based on a multi-objective Genetic Algorithm (GA) for the design of orthogonal filter banks in an image coding scheme. A parameterization is used to achieve perfect reconstruction orthonormal FIR filter banks with a first order regularity. We search the optimal parameter set according to the coding gain, frequency selectivity and the group delay characteristics. Particularly, to design near linear phase filter banks, the group delay flatness in the filter pass-band is introduced as an objective to be minimised. We formulate the optimization problem as multi-objective and we use the Non-dominated Sorting Genetic Algorithm approach (NSGAII) to solve this problem by searching solutions that achieve the best compromise between the different objectives criteria, these solutions are known as Pareto Optimal Solutions. From experimental results, our new optimal filter banks are shown to outperform significantly the Daubechies orthogonal filter banks for the majority of test images.

Keywords

orthogonal filter banks, multiobjective optimization, wavelet image coding.

1 Introduction

In a wavelet image coding scheme, the choice of filter banks is a key problem which affects both the system and the performances of compression. Generally, optimal filter banks are selected for image coding systems from a library of filter banks designed for signal processing purposes using some metrics related to such systems [1].
While the most suitable filter banks for image coding belong to the biorthogonal class of filter banks, the orthogonal Daubechies filter banks belong to the class of wavelet filter banks used most often in image coding applications. Orthogonal filter banks have some interesting properties, such as energy preservation, that are often used in the design of quantization procedures and bit allocations algorithms.
These properties make the orthogonal filter banks very attractive, but in the case of wavelet FIR filter banks, orthogonality is non-compatible with phase-linearity, which seems to be relevant too. A solution of this drawback is to design orthogonal filter banks as symmetric as possible. This can be achieved by the optimization of the group delay flatness in the orthogonal filter banks. To introduce flexibility in the design, the group delay flatness can be considered only in passbands instead of the full band since phase distortion is not important in stopbands. This allows the investigation of a larger region of solution space where better performing filter banks can be obtained.
In this work, we are interested in the design of orthogonal filter banks (FBs) for image coding. The optimal FBs design approach presented in this work is a continuation and improvement of earlier work in the field. Perfect reconstruction (PR) filter banks with a first order regularity requirement are constructed using the parameterization proposed in the reference [3]. In addition, the design of optimal PR-orthogonal FBs for image coding should consider several criteria of practical significance related to such application, namely: energy compaction capabilities or coding gain, frequency selectivity, and phase linearity.
In fact, the design problem requires simultaneous optimization of three objective functions with different individual optima. Practically, there it is no possible solution that satisfies maximally all the objective functions. To solve this problem, a multi-objective genetic approach called NSGAII [6] is used to find *Pareto optimal solutions* that make all possible tradeoffs among competing objectives through evolution.
This article is organized as follows. In section 2, we define the design criteria of filter banks in image coding applications. The optimization problem formulation and the multi-objective GA used for designing filter banks are presented in section 3. In section 4 we evaluate the performances of compression of the optimized filter bank for a set of test images. Finally, section 5 concludes the paper with a summary of our work.

2 Design criteria

2.1 PR condition

Figure 1 shows a two channel filter bank where $H_0(z)$ and $H_1(z)$ are the analysis low and high-pass filters, respectively, and $G_0(z)$ and $G_1(z)$ are the synthesis filters.

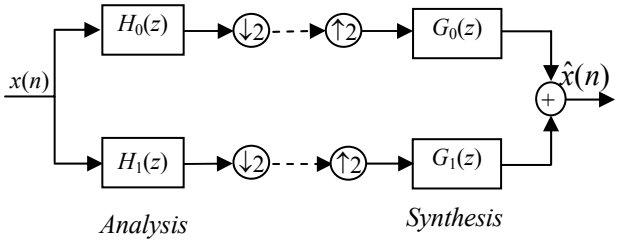


Figure 1 - *Two channel filter bank*

To design a PR orthogonal filter bank, $\hat{x}(n) = x(n)$, the analysis and synthesis filters have to satisfy the following equations:

$$\begin{aligned} h_1(n) &= (-1)^{(n+1)} h_0(L-n-1) \\ g_0(n) &= h_0(L-n-1) \qquad n = 0, \dots, L-1 \\ g_1(n) &= h_1(L-n-1) \end{aligned} \tag{1}$$

Where L is the length of the filters. The above equations therefore establish an important property of an orthogonal system: In an orthogonal two-channel filter bank, all filters are obtained from a single prototype filter.

The lattice parameterization described by Vaidyanathan [2] offers the opportunity to design orthogonal wavelet filters via unconstrained parameters. Imposing the even length filters with $L=2N$ and using polyphase decomposition, $H_0(z)$ and $H_1(z)$ in the analysis filter bank are then given by:

$$\begin{aligned} \begin{bmatrix} H_0(z) \\ H_1(z) \end{bmatrix} &= \begin{bmatrix} \sum_{n=0}^{N-1} h_0(2n) z^{-2n} + z^{-1} \sum_{n=0}^{N-1} h_0(2n+1) z^{-2n} \\ \sum_{n=0}^{N-1} h_1(2n) z^{-2n} + z^{-1} \sum_{n=0}^{N-1} h_1(2n+1) z^{-2n} \end{bmatrix} \\ &= \underbrace{\begin{pmatrix} H_{00}(z^{-2}) & H_{01}(z^{-2}) \\ H_{10}(z^{-2}) & H_{11}(z^{-2}) \end{pmatrix}}_{\boldsymbol{H}_{\mathrm{p}}(z^{-2})} \begin{pmatrix} 1 \\ z^{-1} \end{pmatrix} \end{aligned} \tag{2}$$

A paraunitary filter bank with FIR filters of length $L=2N$ can be reached by the following parameterization [2, 3]:

$$\begin{aligned} \boldsymbol{H}_{\mathrm{p}}(z^{-2}) &= \left[\prod_{\mathrm{i}=1}^{\mathrm{N}-1} \{ \boldsymbol{I} + (\mathrm{z}^{-2} - 1) \boldsymbol{v}_{\mathrm{k}} \boldsymbol{v}_{\mathrm{k}}^{\mathrm{T}} \} \right] \boldsymbol{V}_0 \\ &= \left[\prod_{\mathrm{i}=1}^{\mathrm{N}-1} \boldsymbol{V}_{\mathrm{k}}(\mathrm{z}^{-2}) \right] \boldsymbol{V}_0 \end{aligned} \tag{3}$$

Where

$$\boldsymbol{v}_{\mathrm{k}} = \begin{bmatrix} \cos\theta_{\mathrm{k}} \\ \sin\theta_{\mathrm{k}} \end{bmatrix}, \boldsymbol{V}_0 = \begin{bmatrix} \cos\theta_0 & -\sin\theta_0 \\ \sin\theta_0 & \cos\theta_0 \end{bmatrix}$$

The polyphase matrix can be expressed as [3]:

$$\boldsymbol{H}_{\mathrm{p}}(z^{-2}) = \boldsymbol{X}_{\mathrm{k}}(\mathrm{z}^{-2})\, \boldsymbol{V}_0 \tag{4}$$

Where

$$\boldsymbol{X}_{\mathrm{k}}(\mathrm{z}^{-2}) = \begin{bmatrix} \sum_{m=1}^{K+1} c_m^K z^{-2(m-1)} & \sum_{m=1}^{K+1} d_m^K z^{-2(m-1)} \\ \sum_{m=1}^{K+1} f_m^K z^{-2(m-1)} & \sum_{m=1}^{K+1} g_m^K z^{-2(m-1)} \end{bmatrix}$$

The values of the polynomial coefficients are calculated iteratively using the following equation:

$$\begin{cases} c_m^K = \cos^2\theta_{\mathrm{k}}\, c_{m-1}^{K-1} + \sin^2\theta_{\mathrm{k}}\, c_m^{K-1} + \sin\theta_{\mathrm{k}} \cos\theta_{\mathrm{k}} (f_{m-1}^{K-1} - f_m^{K-1}) \\ d_m^K = \cos^2\theta_{\mathrm{k}}\, d_{m-1}^{K-1} + \sin^2\theta_{\mathrm{k}}\, d_m^{K-1} + \sin\theta_{\mathrm{k}} \cos\theta_{\mathrm{k}} (g_{m-1}^{K-1} - g_m^{K-1}) \\ f_m^K = \cos^2\theta_{\mathrm{k}}\, f_{m-1}^{K-1} + \sin^2\theta_{\mathrm{k}}\, f_m^{K-1} + \sin\theta_{\mathrm{k}} \cos\theta_{\mathrm{k}} (c_{m-1}^{K-1} - c_m^{K-1}) \\ g_m^K = \cos^2\theta_{\mathrm{k}}\, g_{m-1}^{K-1} + \sin^2\theta_{\mathrm{k}}\, g_m^{K-1} + \sin\theta_{\mathrm{k}} \cos\theta_{\mathrm{k}} (d_{m-1}^{K-1} - d_m^{K-1}) \end{cases} \tag{5}$$

Where:

$$\begin{cases} c_1^1 = \sin^2\theta_1,\, c_2^1 = \cos^2\theta_1,\, d_1^1 = -\sin\theta_1 \cos\theta_1,\, d_2^1 = \sin\theta_1 \cos\theta_1 \\ f_1^1 = -\sin\theta_1 \cos\theta_1,\, f_2^1 = \sin\theta_1 \cos\theta_1,\, g_1^1 = \cos^2\theta_1,\, g_2^1 = \sin^2\theta_1 \\ c_{m-1}^{K-1} = d_{m-1}^{K-1} = f_{m-1}^{K-1} = g_{m-1}^{K-1} = 0, \qquad \text{for } \mathrm{m} < 2 \\ c_m^{K-1} = d_m^{K-1} = f_m^{K-1} = g_m^{K-1} = 0, \qquad \text{for } \mathrm{m} > K \end{cases}$$

and $\mathrm{K} = 2, \dots, \mathrm{N}-1$ and $\mathrm{m} = 1, \dots, \mathrm{K}+1$.

Therefore eq. (4) implies that $L=2N$ sequence $h_i(n)$ which satisfies the PR condition is parametrized by N free parameters $\{\theta_{\mathrm{k}}\}$.

The regularity constraint is the crucial distinction between wavelet transforms and perfect reconstruction filter banks. It is related to the number of zeros of $H_0(z)$ at $z = -1$ [1]. In image coding, some regularity is desired and higher regularity does not appear to yield significant improvements for coding quality [1]. The following relation provides a constraint on θ_0 which guarantees the presence of one a zero at $z = -1$ in $H_0(z)$:

$$\left. \begin{bmatrix} H_0(z) \\ H_1(z) \end{bmatrix} \right|_{z=-1} = \begin{bmatrix} cos\,\theta_0 & -sin\,\theta_0 \\ sin\,\theta_0 & cos\,\theta_0 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ \beta \end{bmatrix}, \tag{6}$$

Where β is a constant. A possible solution is $\theta_0 = -\pi/4$, for a given $\beta = -\sqrt{2}$ [3]. This relation reduces the number of free parameters by one. It is shown in [3] that the first order regularity condition, yields:

$$\begin{aligned} h_0(2n) &= \tfrac{1}{\sqrt{2}} (c_{n+1}^{N-1} - d_{n+1}^{N-1}), \\ h_0(2n+1) &= \tfrac{1}{\sqrt{2}} (c_{n+1}^{N-1} + d_{n+1}^{N-1}), \end{aligned} \quad \mathrm{n} = 0, \dots, \mathrm{N}-1 \tag{7}$$

So, the filter $H_0(z)$ with first order regularity can be designed via the following stages:

1. Generate N-1 angles $\{\theta_k, k = 1, \dots, N-1\}$,
2. Compute the polynomial coefficients $c_m^{N-1}, d_m^{N-1}, f_m^{N-1}, g_m^{N-1}$ using eq.(5).
3. Compute the coefficients of $H_0(z)$ using eq.(7).

After computing the coefficients of the low pass filter $H_0(z)$, all the remaining filters of filter banks are deduced from this filter by using eq.(1).

2.2 Coding gain

The Coding Gain (CG) measures the energy concentration capability of filter banks and is a widely accepted general measure of coding performance [4, 5]. By modelling a natural image as a one-dimensional Markovian source with a correlation factor ρ and by assuming uncorrelated quantization errors, Katto and Yasuda [5] derived a filter dependent coding gain:

$$CG(\rho) = 10\log10\left(\prod_{k=0}^{M-1}(A_k B_k)^{-\frac{1}{\alpha_k}}\right) \quad (8)$$

Where: $A_k = \sum_i \sum_j h_k'(i) h_k'(j)\rho^{|j-i|}$, $B_k = \sum_i g_k'(i)^2$

For orthogonal filters we have $B_k = \sum_i g_k'(i)^2$=1. Consequently, we obtain:

$$CG(\rho) = 10\log10\left(\prod_{k=0}^{M-1}(A_k)^{-\frac{1}{\alpha_k}}\right) \quad (9)$$

Where h_k' and g_k' are the k[th] analysis and synthesis filter of the M channel nonuniform filter bank equivalent to the N_d (M= N_d +1) level tree structured filter bank respectively (e.g., figure 2), α_k is the corresponding subsampling ratio, and ρ is the correlation factor.
In addition, we have:

$$H_i'(z) = \begin{cases} H_1(z) & if\ i = 0 \\ H_1\left(z^{\alpha_i/2}\right)\prod_{k=0}^{i-1} H_0\left(z^{\alpha_k/2}\right) & if\ 1 \le i \le M-2 \\ \prod_{k=0}^{i} H_0\left(z^{\alpha_k/2}\right) & if\ i = M-1 \end{cases} \quad (10)$$

Where:

$$\alpha_k = \begin{cases} 2^{i+1}\ if & 0 \le i \le M-2 \\ 2^i\ if & i = M-1 \end{cases}$$

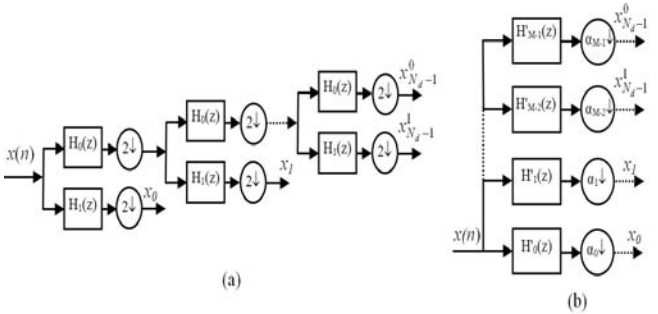


Figure 2 - *M-band filter banks (M= N_d +1): a) N_d level tree structured filter bank (b) the equivalent M channel nonuniform filter bank.*

In our work, a correlation factor ρ=*0.95* is used and a six-level dyadic tree-structured subband decomposition is adopted here since experimentally this number of levels provides the best performance for a wide range of image types and is often used in the evaluation of wavelet image coding algorithms.

2.3 Measure of Symmetry

As mentioned, linear-phase and PR are mutually exclusive in the orthonormal filter bank design. But severe phase nonlinearities are known to create undesired degradations in image and video applications. Therefore, a measure that indicates the level of nonlinearity in the filter-phase response is included as a parameter in the optimal filter design.
Nonlinear-phase is related to the asymmetry of the impulse response. As a measure of filter symmetry, we use the group delay flatness. In the case of symmetry, the group delay is simply a constant. Otherwise, the mean squared Error of the group delay can be used to evaluate group delay flatness:

$$E_{gd} = \frac{2}{\pi}\int_0^{\pi/2} (\tau(\omega) - \tau_0)^2 d\omega \quad (11)$$

Where $\tau(\omega)$ is the group delay of the lowpass filter, defined as $-d\theta/d\omega$, with $\theta(\omega)$ being the phase of $H_0(\omega)$, the discrete time Fourier transform of h_0. Also, τ_0 is the average group delay over the interval $[0, \pi/2]$. The integral is evaluated only over the passband interval since the group delay behavior over the stop band is of little importance. Obviously, given a number of coefficients, the more group delay flatness we have, the closer E_{gd} is to zero. Eq. (11) can be approximated as a summation:

$$E_{gd} = \frac{1}{K}\sum_{n=0}^{K-1}\left(\tau\left(\frac{n\pi}{2K}\right) - \tau_0\right)^2 \quad (12)$$

Where we have K points uniformly distributed over $[0, \pi/2]$ and τ_0 is the mean value defined as $\tau_0 = \frac{1}{K}\sum_{n=0}^{K-1}\tau\left(\frac{n\pi}{2K}\right)$.

2.4 Frequency selectivity

The advantage of frequency selectivity in image coding is that the coarse quantization into unimportant subbands is less expensive, since errors will be confined to the band where they occur. A generally used criterion in subband coding theory is to make the two analysis filters approximate the ideal low-pass and high-pass filters, respectively.
To quantify the frequency selectivity of filter, we define the filter bank Transition Band Energy (TBE) as:

$$TBE = \int_0^{\pi} |H_0(\omega)H_1(\omega)|^2\, d\omega \quad (13)$$

Where $H_i(e^{j\omega})$ is the frequency response of filters $H_i(z)$.

Using parseval's relation we obtain:

$$TBE = \pi \sum_{n=0}^{L-1} |h_0(n) * h_1(n)|^2 \quad (14)$$

This function is a measure of the deviation from an ideal lowpass and highpass filter pair [4]. If the overlap between the filters is zero, which is only possible for ideal filters, then the *TBE* is zero.

3 A multi-objective Genetic Algorithm for the design of filter banks

A general multi-objective optimization problem consists of a number of objectives to be optimized simultaneously and is associated with a number of inequality and equality constraints. Such a problem can be stated as follows:

$$\begin{aligned} &minimise\ (or\ maximise) f_i(x) \quad i = 0, \dots, N \\ &subject\ to: \begin{cases} T_j(x), & j = 1, \dots, N \\ S_k(k) \le 0, & k = 1, \dots, N \end{cases} \end{aligned} \quad (15)$$

The f_i are the objective functions, N is the number of objectives, x is a vector whose p components are the design or decision variables, T_j and S_k are the constraint functions. Generally, the objectives under consideration conflict with each other, and optimizing a particular solution with respect to a single objective can degrade results with respect to the other objectives. Generally, it is difficult to combine the above objectives both to formulate a single objective function. A reasonable solution to a multi-objective problem is to investigate a set of solutions, each of which satisfies the objectives at an acceptable level without being dominated by any other solution. Such solutions form a trade-off space and are known as the *Pareto optimal solutions*.

In a minimization problem, for M objective functions, a feasible solution **X** is dominated by feasible solution **Y** if:

$$\begin{aligned} &\forall i = 1, \dots, M \quad f_i(X) \ge f_i(Y)\ and \\ &\exists j = 1, \dots, M \quad f_j(X) > f_j(Y) \end{aligned} \quad (16)$$

A solution is said to be Pareto optimal if it is not dominated by any other solution in the solution space. Furthermore, that solution is said to be *nondominated solution*. The set of all feasible nondominated solutions are called the *Pareto optimal set*. The corresponding objective function in the objective space constitutes a *Pareto front*.

In our design, we search the angles $\{\theta_k, k = 1, \dots, N-1\}$ that maximise the coding gain and minimise the two individual objective functions, namely, the transition band energy and the group delay error. Our multi-objective optimization problem is formulated as follows:

$$min_{\theta_k}\left(Objf_1, Objf_2, Objf_3\right), and \begin{cases} Objf_1 = TBE \\ Objf_2 = \dfrac{1}{CG_{dB}^2} \\ Objf_3 = E_{gd} \end{cases} \quad (17)$$

In our case, A set of angles are treated as chromosomes which are optimized by a multi-objective genetic algorithm to obtain a set of filter banks that minimise all the prescribed objective functions with a satisfactory level. This algorithm is based on the NSGA II approach [6]. This algorithm is based on the Non-dominated Sorting Genetic Algorithm known as NSGAII. The particularity of this approach is that, in addition to the Pareto nondomination principle used in the conventional multi-objective genetic algorithms, a crowding operator is used to maintain diversity in the population and an elitism mechanism is introduced to prevent the loss of better solutions once they are found during the genetic evolution.

4 Experimental results

Before evaluation of our results, we give some important parameters used in our simulation work. In our genetic algorithm, real valued angles are used for chromosome construction. A simulated binary crossover operator with a distribution index of 40 and a probability of 0.9 is used [7]. Also, a polynomial mutation of distribution index 20 and of probability of 0.01 is applied [7]. The population size n_{pop} is set to 100 and chromosomes of the initial population are obtained by randomly generating angles between [0, 2π]. The maximum generation (G_{max}) is set to 500 generations.

In this work, our design method is applied to the design of an orthogonal filter bank of length 8 (e.g., N=4). Figure 3 shows a 3D scatter plot of a set of Pareto optimal solutions obtained for the optimized filter bank. The NSGAII algorithm produces a set of Pareto optimal solutions which are considered as candidates of the final decision making solution. To select a final solution $\boldsymbol{\theta}^{opt}$ from this set of solutions, we use the following relation [8]:

$$\boldsymbol{\theta}^{opt} = \arg\min_{\boldsymbol{\theta}^j} \max_{i=\{1,2,3\}} \omega_i (Objf_i(\boldsymbol{\theta}^j) - \bar{f}_i) \quad (18)$$

We set:

$$\omega_i = \frac{1}{\frac{1}{n_{pop}} \sum_{j=1}^{n_{pop}} Objf_i(\boldsymbol{\theta}^j)}$$

Where $\boldsymbol{\theta}^j, j = 1 \dots n_{pop}$ is the set of generated Pareto optimal solutions and $\bar{\boldsymbol{f}} = \{\overline{f_1}, \overline{f_2}, \overline{f_3}\}$ is the aspiration level. Note that the $\boldsymbol{\theta}^{opt}$ can be approximately given as the one which gives the closest Pareto optimal solution to the given aspiration level. By choosing the aspiration level (0.05, 0.010, 0.01), we have selected the optimal set as $\theta^{\text{opt}} = \{4.14392765, 2.75161017, 5.1338803\}$.

In our work, it is very interesting to design filter banks that can perform well for any test image. Therefore, a series of images which have different frequency contents has been selected for evaluation of filter banks coding performance.

In experimentation, the SPIHT codec [10] is employed for evaluation of the performance of the optimized filter banks. Six levels of wavelet decomposition have been employed.

Images	Filter banks	Compression ratios				
		128:1	64 : 1	32 :1	16 :1	8:1
Lena	Db4	27,51	30,26	33,24	36,59	40,18
	Opt4	27,53	30,34	**33,34**	36,66	40,21
Barbara	Db4	23,44	24,48	27,41	30,81	35,87
	Opt4	23,41	24,49	27,58	31,09	**36,21**
Goldhill	Db4	26,20	28,03	29,59	32,36	35,97
	Opt4	26,14	28,01	**29,71**	32,46	36,05
Finger	Db4	20,18	21,95	23,58	26,58	30,01
	Opt4	20,19	22,05	23,79	27,02	**30,70**
Bike	Db4	18,19	19,35	21,95	25,44	30,46
	Opt4	18,22	19,37	22,02	**25,56**	30,56
Target	Db4	16,49	17,71	21,93	27,79	35,76
	Opt4	16,69	18,02	22,65	28,10	**36,93**
Straw	Db4	18,73	20,02	21,45	24,00	27,24
	Opt4	18,71	19,99	21,48	24,16	**27,47**
Aerial	Db4	21,19	23,06	24,82	27,74	31,89
	Opt4	21,31	23,25	**25,37**	27,97	32,13

Table 1- *PSNR (in dB) versus compression ratio for our optimized filter bank* "**Opt4**" *and the Daubechies orthogonal filter bank "**Db4"**.*

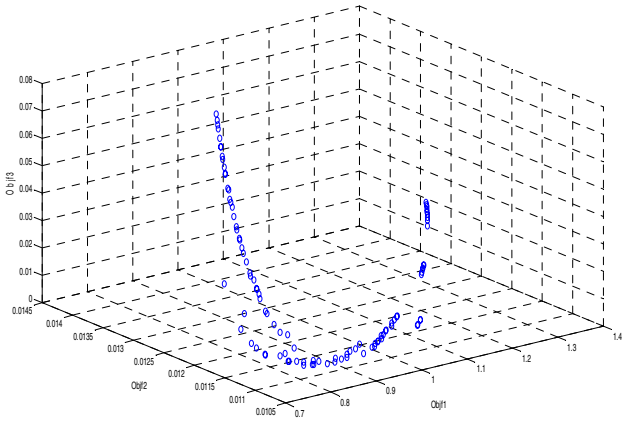


Figure 3 - *3-D scatter plot of the Pareto optimal solutions obtained by using the NSGAII.*

	TBE	CG_{dB}	E_{gd}
db4	0.8	9.664	0.143
FB_{opt}	0.681	9.738	0.0289

Table 2- *Comparison between Characteristics of filter banks*

To qualify the effectiveness of our design method, we compare the performance of our optimized filter bank labeled "**Opt4**" with that of the Daubechies popular orthogonal filter bank "**Db4**". Table 1 presents the PSNR values generated by these two filter banks compared for different compression ratio where the best result is highlighted in each case.

For the eight test images, our optimal filter bank Opt4 outperforms the Db4 filter bank in the majority of cases and the greatest degree of improvement is occurred for compression ratios less than 64:1. For some images such as "Finger" and "Target" the improvement is very significant, e.g., 0.69 and 1.17dB. In the cases where the optimized filter bank is worse, the degradation is very small. Statistically, we have obtained in average an improvement of 0,20dB.

To justify the improvement of performance obtained with our filter banks, we compare their characteristics to those of the filter bank db4 in Table 2. It is clear that our optimized filter bank provide a significant improvement in term of all considered criteria.

To assess the compression performance of our optimized filter bank, we present an example of a test image which has been compressed at rate where the distortion becomes visible. Figure 4 shows the image "Barbara" compressed at a compression ratio 32:1 for the Db4 filter and Opt4 filters. For this compression rate, all compressed images suffer from ringing artifacts. One can clearly see these artifacts surrounding the contours of the images. The image compressed using the optimized filter Opt4 has less ringing effect in comparison to that of the Db4 filter which can be observed in the zoomed portions of such images.

5 Conclusion

In this work, a method based on genetic algorithms was presented for the optimization of filter banks for a lossy image coding scheme. The problem of optimization is to find a set of filter bank coefficients that satisfy multiple objectives which are used to measure the effectiveness of filter banks in such scheme. The problem was formulated as multi-objective and solved using the NSGAII algorithm. From simulation results, it is shown that our optimized filter banks outperform significantly the Db4 filter bank for the majority of tested cases.

By sacrificing the high degree of regularity of orthogonal wavelet filter banks, superior image compression performance can be achieved with filters that exhibit good energy compaction and near linear phase characteristics.

While our interest in this work is in orthogonal filter banks, because of the availability of orthonormality, regularity and perfect reconstruction, the importance of orthonormality for image compression should be given greater consideration for biorthogonal filter banks.

(a) (b)

(a) 4×zoom (b) 4×zoom

Figure 4 - *Image Barbara compressed at 32:1 with SPIHT: (a)* Db4 *(b)* Opt4

References

[1] M. Vetterli and J. Kovačević. Wavelet and subband coding. *Englewood Cliffs*, New Jersey, 1995.

[2] P. P. Vaidyanathan. Multirate Systems and Filter Banks. *Englewood Cliffs*, NJ: Prentice-Hall, 1993.

[3] L. K. Shark, C. Yu. Design of optimal shift-invariant orthonormal wavelet filter banks via genetic algorithm. *Journal of signal processing*, 83 (2003) 2579-2591, Elsevier, 2003.

[4] K. Deb, A. Pratab, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.,* 6(2):182-197, April 2002.

[5] J. Katto and Y. Yasuda. Performance evaluation of subband coding and optimization of its filter coefficients. *Proc. SPIE Symposium on Visual Comm. and Image Process.*, 1605, pages 95-106, 1991.

[6] M. Lightstone, E. Majani, and S. K. Mitra, Low bit-rate design considerations for wavelet-based image coding. *Multidimensional Systems and Signal Processing*, 1997 (8):111-128, 1997.

[7] M. M. Raghuwanshi and O. G. Kakde. Survey on multiobjective evolutionary and real coded genetic algorithms. In proc. of the 8th Asia Pacific Symposium on Intelligent and Evolutionary Systems, pages 150-161, 2004.

[8] Y.B. Yun, H. Nakayama, M. Arakawa. Multiple criteria decision making with generalized DEA and an aspiration level method. *European Journal of Operational Research*, 158 (2004) 697-706, Elsevier, 2003.

[9] M. Eskicioglu and P. S. Fisher. Image Quality Measures and Their Performance. *IEEE Trans. Comm.*, 43(12): 2959-2965, December 1995.

[10] A. Said and W.A. Pearlman. A new, fast, efficient image codec based on set partitioning in hierarehical trees. *IEEE Trans. Circuits Syst. Video Technol.*, 6(3): 243-250, June 1996.

Analyse de comportements dans les points de vente

R. Sicre[12] H. Nicolas[1]

[1] LaBRI (Laboratoire Bordelais de Recherche en Informatique)
Université de Bordeaux, 351 Cours de la libération, 33405 Talence Cedex - France

[2] MIRANE S.A.S.
16 rue du 8 mai 1945, 33150 Cenon - France

{sicre, nicolas}@labri.fr

Résumé

Cet article présente une nouvelle méthode permettant d'analyser, en temps réel, les comportements humains lors de l'acte d'achat. En particulier, nous cherchons à détecter l'intérêt d'une personne pour certains produits et des interactions telles qu'une personne saisissant des produits dans un point de vente.
Le système est basé sur un modèle de comportement. Le module d'analyse vidéo détecte le mouvement, suit les objets (personnes) dans la vidéo et décrit le mouvement local de ces objets. Les comportements définis dans le modèle sont ensuite reconnus. Enfin, nous testons le système sur des jeux de données réels.

Mots clefs

Vision par ordinateur, Analyse de comportement, Détection d'événements, Vidéosurveillance, Marketing.

1 Introduction

De plus en plus d'applications qui utilisent la vision par ordinateur sont développées dans divers domaines tels que la vidéosurveillance [8], surveillance de trafic routier [7], les jeux vidéos, le marketing, etc.
Dans le domaine du marketing, de plus en plus de points de vente s'équipent de systèmes d'affichage vidéo. Ces systèmes ont pour but de jouer des clips publicitaires les uns après les autres. Ils permettent un nouveau type de communication avec les consommateurs par le biais de clips d'animation, films, etc. Cependant l'impact de ces systèmes d'affichage est plus faible que l'on espérait. Ceci est dû au fait que les gens sont habitués à avoir de nombreux affichages publicitaires. C'est pourquoi le contenu et la localisation de ces systèmes ont besoin d'être étudiés. De nos jours, plusieurs logiciels sont disponibles sur le marché. En ce qui concerne la localisation des systèmes d'affichage, certains logiciels permettent de détecter et suivre les clients lors de leurs parcours dans le point de vente afin d'identifier leurs habitudes. D'autres logiciels calculent l'audience des systèmes d'affichage basée sur la détection des visages.

L'étude présentée dans cet article se situe dans le même contexte. Nous souhaitons adapter les clips affichés au comportement des personnes présentes afin d'améliorer l'impact des vidéos. Le système nécessite une phase d'analyse vidéo, suivie d'une phase de reconnaissance de comportement. Plus précisément, le système détecte, en temps réel, des personnes prenant des produits dans des zones connues. La détection d'un tel événement a pour résultat, par exemple, l'affichage d'un clip correspondant à l'objet saisi.
Apres une présentation des travaux précédents, l'article présente notre système. D'abord le modèle de comportement, puis l'analyse vidéo basée objet et la reconnaissance de comportement (voir figure 1). Nous présentons des résultats et proposons les travaux futurs en conclusion.

2 Travaux précédents

Les travaux précédents qui ont pour but de décrire les comportements humains, le font dans des contextes très variés [4] [12] [6]. Les êtres humains sont considérés comme des objets déformables. Le but de l'analyse de comportement est de reconnaitre des échantillons de mouvement afin d'en tirer des conclusions de haut niveau. Il y a beaucoup de problèmes à résoudre. Ceci est dû au fait que nous cherchons à mettre en correspondance des activités du monde réel et des données perçues par des systèmes de traitement des vidéos. Les buts sont ici de sélectionner des propriétés pertinentes générées par des méthodes d'analyse vidéo et de gérer leurs incertitudes.
L'analyse de comportement se fait généralement en deux étapes : description et reconnaissance des actions. La première étape vise à définir un modèle qui décrit chaque action pertinente dans notre contexte applicatif.
Ensuite il existe deux possibilités. D'abord, il y a une phase d'entrainement utilisant des données étiquetées pour ensuite reconnaitre les nouvelles données, basée sur cet entrainement. Les méthodes utilisées sont les modèles de Markov cachés, les réseaux de neurones, les Machines à vecteurs de support (SVM), etc.

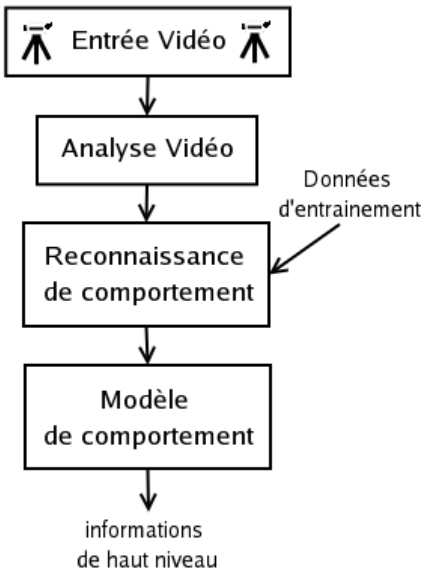


Figure 1 – *Diagramme fonctionnel du système*

Deuxièmement, un modèle logique est généré qui ne nécessite pas de phase d'entrainement. Cependant ces méthodes ne sont pas très flexibles et dépendent fortement de connaissances de la scène.

3 Modèle de comportement

Cette section présente le modèle qui définit le comportement des personnes lors de l'acte d'achat. Dans les points de vente les clients se déplacent, comparent les prix, prennent des produits, etc. Nous utilisons six états pour décrire le comportement des personnes présentes. La succession d'état décrit un scénario exécuté par une personne.

Enter : Une nouvelle personne entre dans la scène.
Exit : La personne sort de la scène.
Interested : La personne est proche des produits, probablement intéressée.
Interact : La personne interagit avec les produits, prend des produits.
Stand by : La personne est dans la scène, mais elle n'est proche ni des produits, ni d'une bordure de l'image. Cette personne peut se déplacer ou être arrêtée.
Inactive : La personne a quitté la scène.

4 Analyse vidéo basée objet

Afin de détecter les six états, nous avons besoin d'informations concernant chaque personne dans la scène. Nous voulons donc connaitre la position et le contour de chaque personne pour chaque image (ou frame) de la vidéo. C'est pourquoi, nous utilisons un procédé de détection de mouvement et de suivi d'objets. Nous supposons que les zones où se situent les produits sont connues.

Ensuite, la reconnaissance d'événements se fait grâce à la description et la classification du mouvement local d'une personne, de sa position relative aux zones de produits et de sa surface de recouvrement avec ces zones de produits.

4.1 Détection de mouvement - suivi d'objets

La détection de mouvement et le suivi d'objets déterminent le contour et la position de chaque objet (ou personne) en mouvement dans la scène pour chaque frame. La méthode utilisée se divise en deux phases : tout d'abord, la détection de mouvement identifie les régions en mouvement qui n'appartiennent pas à l'arrière-plan. Puis ces régions sont suivies le long de la vidéo. Les méthodes les plus rapides sont sélectionnées pour faire face aux contraintes de temps-réel de notre application.

La détection de mouvement utilise un modèle de l'arrière-plan basé pixel. Une mixture de gaussiennes est associée à chaque pixel afin de caractériser l'arrière-plan. Le modèle est mis à jour en ligne. Une distribution Gaussienne est mise en correspondance avec la valeur courante de chaque pixel. Si une Gaussienne appartient à l'arrière-plan, le pixel est classé de même. Sinon, il est considéré comme avant-plan (voir figure 2). Il est intéressant de noter que notre méthode permet aussi de tester si des pixels détectés comme appartenant à l'avant-plan correspondent en réalité à une ombre. Cette méthode nous permet d'effacer ces ombres et améliore grandement les résultats. Plus de détails concernant cette méthode se trouve dans [9] et [13]. Des filtres morphologiques sont appliqués sur le résultat de la détection afin d'effacer du bruit et d'améliorer la forme des régions détectées.

En pratique, un objet détecté peut être recouvert par plusieurs régions non connectées, car l'algorithme ne détecte pas certaines parties de la personne (voir figure 2 première colonne). C'est pourquoi, nous devons fusionner des régions. Les régions détectées sont représentées par leur boite englobante. Lorsque deux boites se superposent, les régions sont fusionnées. De plus, étant donné que nous suivons des personnes, nous supposons que les personnes sont situées dans des rectangles significativement plus hauts que larges, bien que le ratio dépende de chaque personne ainsi que de la position de la caméra. Par conséquent, deux régions situées dans le même axe vertical peuvent être considérées comme couvrant la même personne et ainsi être fusionnées. Dans certains cas, par exemple lorsqu'une seconde personne se trouve derrière la première, la fusion sera incorrecte. Cependant dans ce cas précis, la deuxième personne n'aura aucun intérêt dans notre contexte applicatif. Plus précisément, la fusion se produit lorsque le centre de gravité de la région supérieure se trouve entre les extrémités de la région inferieure. Chaque fusion est vérifiée par la phase de suivi.

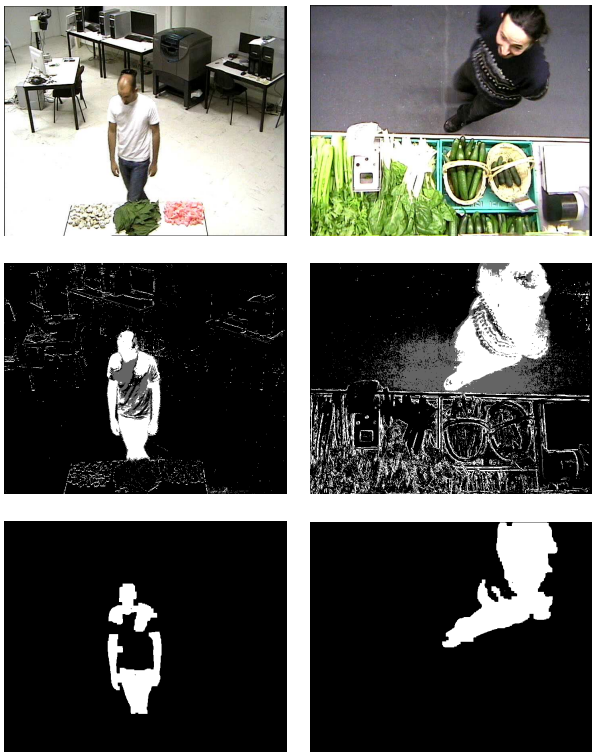


Figure 2 – *Détection de mouvement pour des vidéos dans les jeux MALL1 et LAB1. Les images de la premières ligne sont prises de la vidéo, la deuxième ligne correspond à la détection de mouvement avec detection des ombres (en gris) et la dernière est le résultat final après filtrage.*

Le suivi d'objets calcule d'abord pour chaque région un descripteur basé sur sa position, taille, surface, moment de couleur du premier et second ordre. Ces descripteurs sont ensuite mis en correspondance d'une image à l'autre, en utilisant un système de vote qui détermine les régions les plus similaires dans deux ensembles de régions. Finalement, les correspondances sont vérifiées grâce aux objets déjà suivis.

4.2 Description de la prise de produits

Cette section se focalise sur les interactions entre les personnes suivies et les zones de produits (personne prenant des produits). Lors de la prise d'un produit, une personne tend d'abord le bras, puis saisi un produit et finalement emmène le produit. Les différentes phases de cet événement de « prise de produit » correspondent à des mouvements locaux de cette personne, qui sont observables. Tout comme dans [3] et [10], nous utilisons un descripteur basé sur le mouvement local, mais aussi sur les interactions avec les zones de produits pour caractériser la prise de produit. Ce descripteur est utilisé dans la phase de reconnaissance de comportement (voir section 4) et est défini comme suit.

Le descripteur du mouvement local est calculé pour chaque frame et pour chaque personne suivie (voir figure 3). D'abord, la taille du masque de chaque personne est normalisée à une taille standard de 120x120 pixels, en gardant le rapport hauteur / largeur. Ensuite, le flow optique est calculé en utilisant l'algorithme de Lucas Kanade [5]. Cet algorithme nous retourne deux matrices contenants les valeurs des vecteurs de mouvement selon les axes x et y. Nous séparons les valeurs négatives des valeurs positives des deux matrices et obtenons 4 matrices avant d'appliquer un filtrage Gaussien sur chacune d'entre elles pour réduire le bruit. Une cinquième matrice est générée qui représente la silhouette de la personne suivie. Puis nous réduisons la dimensionnalité de ces matrices afin d'améliorer la vitesse du système. Chaque matrice est divisée en une grille 2x2. Nous intégrons les valeurs de chaque cellule de la grille selon un histogramme radial contenant 18 portions de 20 degrés chacune. Les matrices sont désormais représentées par un vecteur à 72 dimensions (2x2x18). Le descripteur complet contient lui 360 dimensions (5x72).
Pour prendre en compte les informations temporelles, nous utilisons 15 frames autour de l'image courante. Ces 15 frames sont divisées en trois groupes de 5 et représentent respectivement : le passé, le présent et le futur. Après avoir appliqué une Analyse en Composante Principale (PCA) sur les descripteurs de mouvement local de chaque frame du groupe, nous conservons les 50 premières composantes pour le groupe présent et les 10 premières pour les groupes passé et futur. Ce descripteur du contexte de mouvement local possède 70 dimensions (10+50+10), il est ajouté au descripteur précédent.

Le descripteur d'interaction utilise des informations provenant de la phase de suivi d'objets. Six valeurs sont calculées :

- La surface d'une personne recouvrant une zone de produits
- Un booléen qui est mis à 1 lorsque cette surface dépasse la taille théorique d'une main ou lorsque une personne est proche d'une zone de produits et que du mouvement est détecté dans cette zone.
- La surface totale recouverte par la personne dans l'image.
- La taille (hauteur) de la personne.
- La position du maximum de la personne selon l'axe y.
- La position du minimum de la personne selon l'axe y.

Les mesures concernant la taille et la position de la personne ont des variations intéressantes lorsqu'une personne saisit des produits. La surface recouverte par une personne dans l'image a tendance à augmenter lorsque celle-ci prend un produit, suivant la taille des produits. Ces mesures composent le descripteur d'interaction, qui possède 90 dimensions (6x15), car nous conservons les mesures des 15 frames.

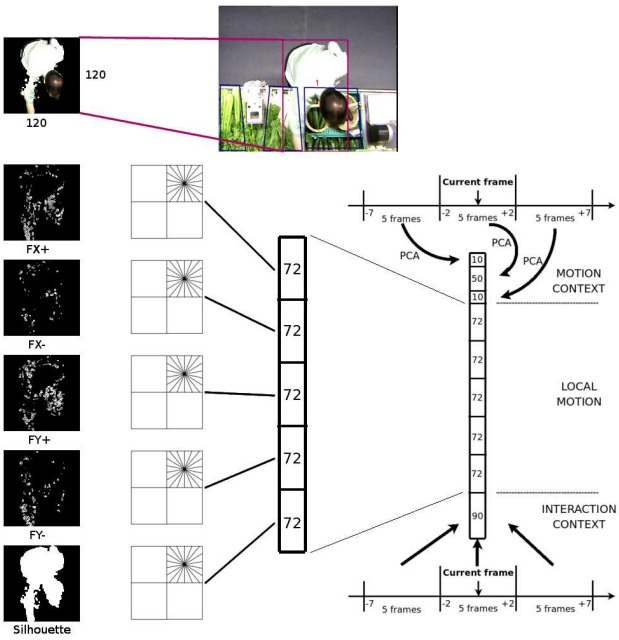


Figure 3 – *Diagramme représentant le descripteur*

5 Reconnaissance de comportement

Cette section présente le procédé de reconnaissance de comportement. Basés sur le modèle de comportement, nous déterminons les six états et créons une Machine à Etat Fini (MEF).

En utilisant les informations de l'analyse vidéo, nous devons définir pour chaque frame l'état dans lequel se trouve chaque objet (ou personne) suivi.

Enter est détecté lorsqu'une nouvelle personne est connectée à une bordure de l'image.

Exit est détecté lorsqu'une personne déjà suivie est connectée à une bordure de l'image.

Interested est détecté lorsque le contour d'une personne est connecté à une zone de produits.

Stand by est détecté lorsqu'une personne est dans la scène et qu'elle n'est ni connecté à une zone de produits ni à une bordure de l'image.

Inactive est détecté lorsque le système ne suit plus une personne, ou perd sa trace. Cet événement se produit lorsqu'une personne quitte la scène, ou est occulté par quelque chose dans la scène ou par une autre personne.

Interact est détecté en utilisant les SVM [2] sur le descripteur de mouvement local et d'interaction. Tout d'abord, il y a deux phases lors de la classification de données : l'entrainement et le test. Les données sont composées de plusieurs instances. Une instance est elle-même composée d'une valeur cible (target value) et de plusieurs attributs. Dans notre cas, la valeur cible est 0 ou 1, selon si une prise de produit a lieu ou non. Les attributs sont toutes les valeurs contenues dans le descripteur. Les SVM créent un modèle qui prédit les valeurs cibles à partir des attributs, en résolvant le problème d'optimisation suivant :

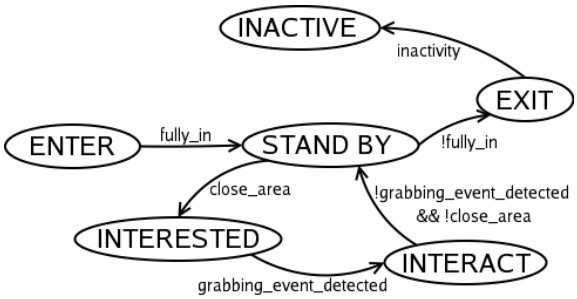


Figure 4 – *Portion de la machine à état fin. Toutes les transitions ne sont pas écrites pour une meilleure lecture.*

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i$$

avec $y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i$ et $\xi_i \geq 0$

Le vecteur d'entrainement x_i est ajusté dans un espace à plus haute dimension par la fonction ϕ. Dans cet espace, les SVMs déterminent un hyperplan séparant les données et maximisant la variance. $\boldsymbol{C > 0}$ est le paramètre de pénalité (penalty parameter) de l'erreur. Le système utilise un noyau Gaussien (RBF).

$$K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \gamma > 0$$

Bien que l'on puisse utiliser divers noyaux, nous choisissons le noyau RBF. En effet, il gère les relations non-linéaires entre les attributs et les valeurs cibles, contrairement aux noyaux linéaires. Les noyaux RBF ont aussi moins d'hyper-paramètres et de difficultés numériques que les noyaux polynomiaux ou sigmoid.

Une Machine à Etats Finis est créée afin d'organiser et de hiérarchiser les six états [1] [11]. La machine à état est synchrone et déterministe. La machine est synchrone, car elle itère à chaque nouvelle frame. Nous calculons le nouvel état en partant de l'état précédent et en testant toute les conditions de transition. Si une condition est satisfaite, le système se déplace dans le nouvel état correspondant. Sinon, le système reste dans le même état. La machine est déterministe, car pour chaque état, Il ne peut y avoir plus d'une transition valide pour chaque entrée possible. Une MEF modélise le comportement d'une personne. Nous sauvegardons le chemin de la personne à travers la MEF pour déterminer le scénario joué. La figure 4 montre un chemin possible à travers la MEF.

6 Résultats

6.1 Description des données

Nous utilisons plusieurs jeux de données pris avec la même caméra. Une partie des jeux de données ont été acquis dans notre laboratoire (LAB1, 2 et 3). Les autres ont été tournés dans un hypermarché (MALL1 et MALL2). Les jeux LAB1 et MALL1 possèdent cinq et six vidéos respectivement et contiennent beaucoup d'interactions avec les produits. Deux et quatre personnes différentes font leurs courses respectivement (voir figure2). LAB2 est un jeu dans lequel il n'y a pas d'interactions avec les produits et possède 3 acteurs. Les produits pris ont diverses tailles, couleurs, formes. De plus, tous les produits sont identiques dans les zones de produits. LAB3 et MALL2 sont deux jeux de données ou plusieurs personnes (deux à quatre) interagissent simultanément (voir figure 5).

6.2 Tests sur les états

Nous lançons des tests sur les données pour comprendre le comportement du système dans divers cas. Comme le système génère un état pour chaque personne et pour chaque frame de la vidéo, nous comparons ces résultats à la vérité terrain. Nous calculons ensuite le pourcentage d'états corrects (voir tableau 1). Le jeu de données LAB2, qui n'a pas d'interaction avec les produits obtient 89% et a un meilleur résultat que LAB1 avec 76%. Nous notons que l'état interact (prise de produits) est plus difficile à détecter que les autres états, pour les vidéos prises en laboratoire.

6.3 Tests sur la prise de produits

Pour reconnaître la prise de produits, nous utilisons un procédé de validation croisée (cross validation) (voir tableau 2). En particulier, pour reconnaître les événements d'une vidéo d'un jeu de données, nous utilisons toutes les autres vidéos de ce jeu comme entrainement et ensuite nous calculons la précision et le rappel.
Le jeu de données LAB1 offre de meilleurs pourcentages que MALL1. Cette différence est principalement due au bruit, qui est plus important dans l'hypermarché, ainsi qu'a des mouvements de caméra lors de la capture des vidéos. Cependant MALL1 offre de très bons taux de rappel et précision pour la détection de prise d'objet (voir tableau 2). Ceci est du à la position de la caméra, la caméra est directement au dessus des produits et plus proche que dans les vidéos tournées au Laboratoire. Les produits sont donc plus grands dans l'image et cela facilite leur détection. Dans le cadre d'une application concrète, un compromis devra être fait concernant la distance entre les zones de produits et la caméra. Une position proche des produits offrira de bons résultats pour la détection de prise d'objets. Cependant, un champ de vision trop réduit limitera les possibilités d'analyse du comportement des personnes.
Comme on le voit sur le tableau 2, nous avons comparé les résultats en utilisant deux descripteurs: le descripteur d'interaction seul (DI) et le descripteur de mouvement local avec le descripteur d'interaction (DMI). Le DMI a des performances similaires au DI en ce qui concerne la précision, mais offre un meilleur rappel. Pour les jeux de données avec plusieurs personnes, DMI a de meilleurs résultats en moyenne que DI. Cependant, le descripteur de mouvement local est plus bruité, cela est du à des erreurs de suivi.
La limite la plus importante du système réside dans le fait qu'il ne gère pas les phases d'occlusions. Cependant, la reconnaissance de prise de produits est robuste et offre de bons résultats pour les vidéos avec plusieurs personnes.
Pour conclure, nous pouvons faire quelques remarques générales concernant les produits. Ceux de petite taille sont logiquement plus difficiles à détecter lorsqu'ils sont pris. De plus, les produits clairs ont tendance à générer des fausses détections dues aux ombres des personnes lorsqu'elles sont proches. La méthode de détection des ombres atteint ses limites dans ce cas précis.

6.4 Temps d'exécution

Notre application doit générer une réaction en temps réel, dès qu'un événement est détecté. Le système est testé sur un ordinateur avec un Pentium 4, 3 Ghz et 1 Go de RAM. L'application analyse 6 à 10 images par secondes pour des résolutions de 704x576 ou 640x480 respectivement. La détection de mouvement est la phase la plus couteuse en temps de calcul.

7 Conclusion

Cet article présente un nouveau type d'application, utilisant la vision par ordinateur dans le domaine du marketing. Le système améliore les interactions entre les clients et les systèmes d'affichage, dans un point de vente. Celui-ci détecte, suit, et analyse les comportements des clients tels que les intérêts et les interactions avec les divers produits. Le système offre des résultats intéressants, 73% des frames sont correctement libellées pour des vidéos prises en environnement réel. Les interactions (prises d'objets) sont détectées avec une précision de 0.79 et un rappel de 0.85. Cette évaluation nous permet de comprendre le comportement du système pour augmenter son efficacité. Un prototype sera bientôt mis en place pour être testé sur une longue durée.
Notre méthode peut être améliorée avec un algorithme de gestion des occlusions. Il serait aussi intéressant de caractériser de nouveaux comportements et scénarios en utilisant la même technique.

Références

[1] F. Bremond, G. Medioni, "Scenario recognition in airborne video imagery", *Proc. Int. Workshop Interpretation of Visual Motion*, pp 57-64. 1998.

Dataset	Video	Frames	Correctness
MALL 1	1	327	70,03%
	2	444	74,77%
	3	434	66,13%
	4	336	70,83%
	5	164	76,22%
	6	232	79,74%
	mean		**72,95%**
LAB 1	1	545	85,87%
	2	672	74,40%
	3	704	76,28%
	4	771	60,57%
	5	518	92,66%
	mean		**76,13%**
LAB 2	1	476	87,61%
	2	342	82,46%
	3	143	91,61%
	4	303	95,71%
	mean		**89,35%**

Tableau 1 – *Tableau représentant le pourcentage d'état correctement détecté (correctness) pour chaque frame des vidéos.*

Figure 5 – *Quelques images provenant des jeux de données LAB3 et MALL2.*

Jeux	Video	Frames	Rappel I	Precision I	Rappel MI	Precision MI
MALL1	1	327	0,5306	0,5977	0,8163	0,6667
	2	444	0,9307	0,9592	0,9505	1
	3	434	0,6667	0,7368	0,7525	0,5802
	4	336	0,6905	0,9063	0,7976	0,8701
	5	164	0,5	1	0,5	1
	6	232	0,8475	0,9434	0,8983	0,9815
	mean		**0,6943**	**0,8572**	**0,7859**	**0,8498**
LAB1	1	545	0,7299	0,7692	0,8321	0,8085
	2	672	0,6585	0,648	0,6748	0,6288
	3	704	0,6774	0,9333	0,7473	0,9392
	4	771	0,7203	0,7687	0,7552	0,7347
	5	518	0,9818	0,75	0,9818	0,7941
	mean		**0,7536**	**0,7738**	**0,7982**	**0,7811**
MALL2 MP	1	215	0,8929	0,8333	0,8214	0,902
	2	208	0,6462	0,8235	0,7846	0,8947
	3	735	0,7151	0,7278	0,9101	0,72
	4	153	1	0,5106	0,5417	0,9286
	5	382	0,8333	0,8404	0,6583	0,8404
	6	504	0,7273	0,8571	0,8561	0,8828
	mean		**0,8025**	**0,7655**	**0,762**	**0,8614**
LAB3 MP	1	212	0,7282	0,8929	0,8738	0,9375
	2	211	0,9063	0,8969	0,7604	0,9012
	3	300	0,784	0,7538	0,856	0,7643
	4	303	0,7561	0,5636	0,7317	0,6383
	5	259	0,8158	0,6596	0,8026	0,6854
	mean		**0,7981**	**0,7534**	**0,8049**	**0,7853**

Tableau 2 – *Tableau représentant le rappel et la précision pour la détection de prise de produits. Deux descripteurs sont testés sur plusieurs vidéos : descripteur d'interactions (I) et le descripteur de mouvement local et d'interaction (MI).*

[2] C. Chang and C. Lin, "LIBSVM: a library for support vector machines", http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[3] A. A. Efros, A.C. Berg, G. Mori, J. Malik, "Recognizing action at a distance", *Int. Conf. on Computer Vision*, 2003.

[4] W. Hu, T. Tan, L. Wang, S. Maybank "A survey on visual surveillance of object motion and behaviours," *IEEE Transaction on systems, man, and Cybernetics*, pp 334 – 352, 2004.

[5] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision", *in Proc. 7th IJCAI*, pp.674–679, 1981.

[6] T. Moeslund, A. Hilton, V. Kruger, "A survey of advances in vision-based human motion capture and analysis", *Computer vision and image understanding*, pp 90-126, 2006.

[7] B. T. Morris and M. M. Trivedi, "A survey of vision-based trajectory learning and analysis for surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 8, pp. 1114–1127, 2008.

[8] PETS: Performance Evaluation of Tracking and Surveillance, http://winterpets09.net/

[9] C. Stauffer, W. Grimson, "Adaptive background mixture models for real-time tracking" *Proc. Computer Vision and Pattern Recognition*, pp. 246–252, 1999.

[10] D. Tran, A. Sorokin, "Human activity recognition with metric learning", *Euro. Conf. on Computer Vision*, 2008.

[11] F. Wagner, *Modeling Software with Finite State Machines: A Practical Approach*, Auerbach Publications, ch. 4, 2006.

[12] A. Yilmaz, O. Javed, M. Shah, "Object tracking: a survey", *ACM Computing Surveys*, 2006.

[13] Z. Zivkovic, F. van der Heijden "Efficient adaptive density estimation per image pixel for the task of background subtraction" *Pattern Recognition Letters*, vol. 27, no. 7, pages 773-780, 2006.

Transmission robuste de vidéo basée ondelette à travers un canal MIMO

J. Abot[1] C. Perrine[1] C. Olivier[1] Y. Pousset[1]

[1] XLIM-SIC (Laboratoire XLIM département Signal Image Communication)

CNRS UMR 6172, Université de Poitiers,
Téléport 2, Bvd P. et M. Curie, BP 30179, 86962 Futuroscope-Chasseneuil Cedex – France

{abot, perrine, olivier, pousset}@sic.univ-poitiers.fr

Résumé

Cet article présente l'évolution d'un codeur pour images fixes basé ondelette vers le codage vidéo. Ce codeur associe une transformation en ondelette classique, une quantification vectorielle par sous-bande et une modulation de type MAQ-M. L'exploitation conjointe des éléments de la chaîne de transmission assure une qualité de service dans des conditions de transmission difficiles. Les redondances temporelles sont exploitées, dans le cadre du codage vidéo, par un algorithme spécifique d'estimation et de compensation de mouvement adapté à nos contraintes de codage. Nous proposons ainsi une méthode de codage permettant d'atteindre une bonne robustesse des vecteurs mouvement. Ce codeur nommé V-WTSOM présente une hiérarchisation importante des informations. Nous exploitons cette hiérarchisation en transmission à travers un canal MIMO incluant une possibilité de précodage diagonal. La décomposition en sous-canaux SISO parallèles et indépendants permet d'envisager des stratégies d'allocation basées sur le contenu et améliorant globalement la qualité de service.

Mots clefs

Transformée en ondelette, vecteur mouvement, MIMO, précodage, V-WTSOM

1 Introduction

De nos jours, nous assistons à un développement important des technologies sans fil telles que la téléphonie mobile ou les réseaux ad'hoc. Ces technologies impliquent des transmissions d'informations dont la qualité et la quantité ne cessent d'augmenter à l'image de la HD. Toutefois la bande passante sur les canaux radiomobiles devient de plus en plus limitée. On comprend alors l'importance que revêt l'étape de compression dans une chaîne de transmission vidéo. A l'heure actuelle, de nombreux standards de codage vidéo (H.264, MPEG-4) permettent d'obtenir un compromis intéressant en termes de débit/distorsion sur des canaux présentant peu de perturbations [1]. Tous les standards vidéos actuels sont basés sur la transformée DCT (Discrete Cosines Transform). Aujourd'hui, aucun codeur basé ondelette n'a fait l'objet d'une standardisation. Pourtant différentes techniques de codage vidéo basées ondelettes existent dont [2, 3] sont des exemples. L'intérêt de ces méthodes est de compresser davantage en apportant une scalabilité accrue. Mais ces méthodes s'occupent rarement des problèmes de robustesse lors de la transmission. Comme source de problème, nous pourrions citer, par exemple, le codage à longueur variable qui peut être responsable de désynchronisations de flux en réception. Des méthodes de robustesse [4, 5] ont été envisagées pour H.264 donnant naissance au profil « extended ». Toutefois lorsque le Taux d'Erreur Binaire (TEB) devient trop important il devient nécessaire d'utiliser des Codes Correcteurs d'Erreur (CCE) ayant pour conséquence de diminuer le débit utile. Afin de faire face à cette problématique, nous avons proposé dans [6] une stratégie de codage conjoint (codeur WTSOM pour Wavelet Transform Self-Organized Map pour images fixes). Cette stratégie a été pensée dès la conception du codeur pour s'adapter aux contraintes apportées par le canal de transmission et non les subir. Ainsi, il devient possible de transmettre des images sur des canaux présentant des TEB très importants ($\geq 10^{-3}$) et cela, sans CCE. Le codeur vidéo V-WTSOM présenté dans ce papier, s'appuie comme pour les standards sur l'estimation et la compensation de mouvement. Il exploite aussi les qualités du codeur WTSOM. Pour les codeurs standards, les Vecteurs Mouvement (VM) sont très sensibles aux erreurs de transmission. Nous proposons pour V-WTSOM une technique d'estimation et de codage des VM permettant d'assurer la robustesse au prix de certaines distorsions. L'objectif étant de décoder une vidéo dans des conditions où les standards n'arrivent plus à décoder. Par exemple, V-WTSOM présente un intérêt pour la visioconférence. L'objectif de ce codeur n'est pas de concurrencer les standards mais d'assurer des transmissions exploitables lorsque les standards ne peuvent plus assurer la qualité de service requise en minimisant le débit utile de la source. Les conditions de transmission défavorables se retrouvent en particulier dans les environnements complexes tels que les environnements urbains riches avec multitrajets et/ou

incluant la mobilité. La technologie MIMO (Multiple Input Multipe Output) couplée à la modulation OFDM (Orthogonal Frequency Division Multiplex) permet d'exploiter la diversité du canal. Cette association a déjà été adoptée pour les normes Wi-Fi (IEEE802.11n), LTE et Wi-Max (802.11e). Nous utilisons cette association dans le cadre de la stratégie de transmission V-WTSOM afin d'exploiter la hiérarchisation du flux de données. L'apport de ce papier est donc de proposer une stratégie de codage robuste des VM permettant une bonne adéquation avec une stratégie de transmission basée MIMO.
Dans la seconde partie, nous présenterons le codeur d'images fixes WTSOM ainsi que son extension à la vidéo. Dans une troisième partie, nous évoquerons la technique MIMO et en particulier le principe du précodage diagonal et son intérêt pour la stratégie de transmission. La quatrième partie présentera des résultats de simulation avant de terminer par une conclusion.

2 Codeur V-WTSOM

2.1 Codage en GOP

Le codeur V-WTSOM exploite les redondances temporelles par le codage en GOP (Group Of Pictures), à l'image des standards. On trouve 3 types d'images : Intra (I), Prédite (P) et Unidirectionnelle (U). Le codage de ces images obéit à une hiérarchie illustrée sur le schéma suivant (fig. 1) :

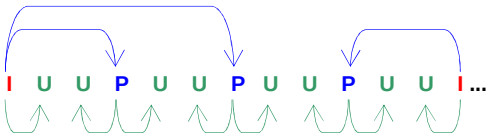


Figure 1 - *Codage du GOP dans V-WTSOM*

Sur la figure 1, la pointe d'une flèche indique une image courante. Ces images nécessitent une image référence (désignée par le départ d'une flèche) pour leur reconstruction. Les images Unidirectionnelles sont des images nécessitant une seule référence afin d'être reconstruite, à la différence des images Bidirectionnelles des standards, qui nécessitent une interpolation liée à deux références.

2.2 Codage des images Intra

Les images Intra sont codées suivant l'algorithme WTSOM pour images fixes [6] dont nous rappelons brièvement le principe (fig. 2). L'image subit une transformation en ondelette de niveau 3. Seules les sous-bandes les plus informatives sont conservées (5 à 7 suivant la QoS désirée). Pour justifier ce choix, nous nous appuyons sur l'entropie calculée dans chaque sous-bande [6]. L'impact visuel lié à cette suppression d'information est jugée acceptable compte tenu de l'application visée. Les sous-bandes conservées subissent une Quantification

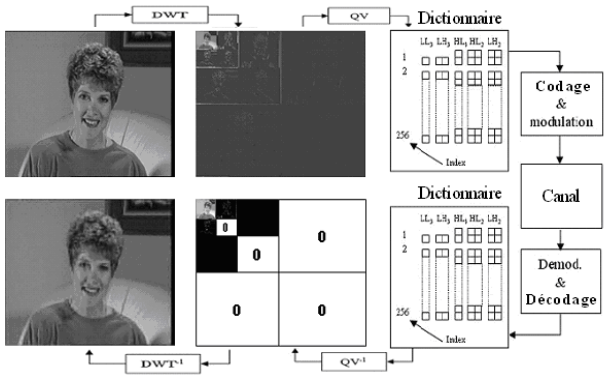


Figure 2 - *Principe de codage de WTSOM fixe*

Vectorielle (QV) dont le format des vecteurs de quantification est adapté au contenu de la sous-bande. Réaliser une QV implique d'utiliser des dictionnaires de vecteurs. Pour les construire, nous utilisons les cartes topologiques de Kohonen issues de l'algorithme Self-Organizing Map (SOM). Cet algorithme présente la particularité de générer des dictionnaires auto-organisés dans le sens où deux vecteurs voisins dans le dictionnaire vont présenter le plus de similitudes possibles. La stratégie WTSOM tire sa robustesse de cette particularité. En superposant une constellation MAQ-M aux dictionnaires auto-organisés, nous pouvons minimiser l'impact d'une erreur de transmission (en considérant M égal au nombre de vecteurs du dictionnaire). Dans cette configuration, un symbole erroné en réception produit un vecteur décodé très proche du vecteur original. La distance euclidienne sera d'autant plus grande que l'erreur sera importante. Cette stratégie permet d'atteindre de bonnes performances en termes de robustesse. Dans le cadre d'une transmission, nous considérons que le récepteur dispose des dictionnaires. Le débit atteint pour un codage type [6] est de 0,3125 bpp.

2.3 Codage des images P et U

Le codage des images P et U s'inspire des techniques employées par les standards MPEG et H.26X. Le codage de ces images est réalisé en exploitant les redondances temporelles par le biais d'un algorithme d'estimation et de compensation de mouvement par block-matching (BMA pour Block-Matching Algorithm) [7].
Rappelons que l'estimation de mouvement par BMA implique d'utiliser une image référence et une image courante. L'image courante est découpée en blocs, l'idée consiste alors à rechercher où les blocs de l'image courante se situent dans l'image référence. Un bloc est désigné comme étant optimal par minimisation d'un critère, généralement le SAD (Sum of Absolute Difference). Le déplacement des blocs est représenté par un VM créant ainsi un champ de VM pour l'image complète. Appliquer ce champ de VM sur l'image référence permet de reconstruire une image compensée en

mouvement. Toutefois, un champ de VM ne peut pas coder à lui seul une image car celle-ci présenterait de multiples distorsions (effet bloc important). Ainsi un terme d'erreur (résidu) entre l'image courante originale et l'image compensée en mouvement est calculé. Ce résidu doit subir un codage intra avant sa transmission.

2.3.1 Estimation de mouvement

Dans le standard H.264, le champ de VM a pour but de donner une estimation de l'image (image compensée en mouvement). Un résidu va contenir les corrections à apporter à l'image. En outre, l'estimation et la compensation de mouvement par le champ de VM permettent de réduire l'entropie du résidu. Par la suite, le résidu est compressé (découpage en blocs, DCT, quantification, codage à longueur variable). La faible entropie présentée par le résidu couplée à un codage à longueur variable permet ainsi d'être efficace d'un point de vue compression. Par contre la méthode est sensible aux erreurs de transmission (désynchronisation du flux). Il faut alors mettre en place des outils de robustesse comme par exemple des CCE ou des marqueurs de resynchronisation. Pour V-WTSOM, l'approche est différente. L'objectif de ce codeur est d'être robuste aux erreurs de transmission. Ainsi, nous avons opté pour un codage à longueur fixe (absence de désynchronisation du flux en cas d'erreur). Cette approche rend obsolète la notion d'entropie du résidu car la quantité d'information, faible ou importante, n'aura pas d'impact sur le taux de compression des résidus. La principale difficulté consiste alors à réaliser un codage efficace mais respectant la contrainte suivante sur les taux de compression au sein du GOP, i.e :

$$T_c(I) < T_c(P) \leq T_c(U)$$

où T_c est le taux de compression des images I, P ou U.

2.3.2 Codage des VM

La problématique du codage des VM est récurrente dans le codage vidéo. Les VM peuvent être vus comme des pointeurs vers une zone de l'image référence. En complément le résidu permet de corriger localement la distorsion éventuellement présente lors de la compensation de mouvement. On comprend alors que la moindre erreur de transmission sur les VM engendre un pointeur vers une autre zone de l'image référence. Cette zone n'est alors plus optimale et le résidu devient inefficace afin de corriger la distorsion. Les VM sont, par conséquent, très sensibles aux erreurs de transmission qui modifient leur valeur. Dans le standard H.264, les VM subissent une prédiction spatiale, puis sont codés sans perte à l'aide d'un codeur entropique (CABAC). Dans le schéma de codage V-WTSOM, cette sensibilité aux erreurs doit être maîtrisée afin d'assurer la robustesse. Afin d'être en accord avec le schéma de WTSOM, nous souhaitons que notre codage soit adapté à une transmission MAQ-M. Une étude a permis de conclure qu'un codage sur 4 bits des VM représente un bon compromis entre qualité/distorsion et robustesse. L'algorithme BMA va venir lire le contenu du dictionnaire, et ne va tester que les blocs candidats désignés par les VM du dictionnaire. Se pose alors le problème de la construction de ce dictionnaire. Utiliser une carte topologique de Kohonen pour la construction de ce dictionnaire ne donne pas de résultats cohérents relativement à une mesure de qualité. Notre proposition est basée sur une simple étude statistique des VM (fig. 3) sur une base de sept vidéos.

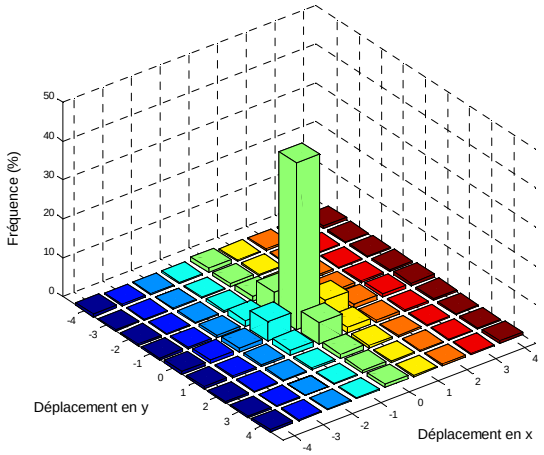


Figure 3 - *Distribution des VM après estimation exhaustive sur un ensemble de sept vidéos*

La figure 3 montre la distribution des VM lorsque l'estimation se fait de manière exhaustive (tous les blocs de la fenêtre de recherche sont testés). Le VM (0,0) présente une fréquence plus importante que les autres VM. Il correspond à tous les blocs qui ne bougent pas d'une image à l'autre. Les VM les plus courants sont aussi dans le voisinage de (0,0). Ainsi les 16 VM les plus fréquents représentent à eux seuls 77,3% de tous les VM d'une estimation exhaustive soit une bonne représentation. La concentration des VM les plus fréquents autour de (0,0) impliquera qu'en cas d'erreur de transmission, l'impact visuel sera minimisé par le choix d'un bloc très voisin du bloc optimal. Enfin, considérer un nombre de VM supérieur à 16 reviendrait à rendre notre algorithme plus sensible aux erreurs de transmission. En effet, plus de VM implique des erreurs d'impact potentiellement plus important à cause de la distance euclidienne plus grande entre deux extrémités du dictionnaire. Finalement, du point de vue de la transmission des VM, il est nécessaire d'organiser le dictionnaire de manière à respecter l'ordonnancement spatial des VM en cas d'erreur.

2.3.3 Traitement des résidus

Le résidu bénéficie d'un codage Intra adapté à ces caractéristiques. Etant donné que, de part sa nature, il

contient moins d'information qu'une image Intra classique, nous pouvons envisager de le coder avec des vecteurs présentant une résolution plus importante. De plus, suivant la qualité désirée nous pouvons coder plus ou moins de sous-bandes. Afin d'obtenir des résultats acceptables, l'expérience montre qu'il n'est pas nécessaire de coder plus de 3 sous-bandes (LL3, HL3 et LH3) pour les résidus des images P et seulement LL3 pour les résidus des images U. Le codage et l'utilisation du résidu permet au codeur V-WTSOM de faire preuve d'une grande scalabilité.

2.4 Exemple de transmission

Nous présentons un exemple de transmission réalisée par le biais d'un canal gaussien et d'une modulation MAQ-16. La passage de la MAQ-256 à la MAQ-16 pour la transmission des éléments codés sur 8 bits est compatible mais provoque une baisse de PSNR de l'ordre de 1,5dB [6]. Les vidéos utilisées sont au format CIF en niveaux de gris. Les caractéristiques du codage V-WTSOM pour ces simulations sont données dans le tableau 1 :

Codage I	Codage P		Codage U		Tc
	VM_P	Résidu_P	VM_U	Résidu_U	
LL3 : vecteurs 1x1 HL3 : vecteurs 2x1 LH3 : vecteurs 1x2 HH3 : vecteurs 2x2 HL2 : vecteurs 4x4 LH2 : vecteurs 4x4 HH2 : vecteurs 4x4	Bloc BMA : 16x16 pixels	LL3 : vecteurs 1x1 HL3 : vecteurs 2x2 LH3 : vecteurs 2x2	Bloc BMA : 16x16 pixels	LL3 : vecteurs 2x2	0,114 bpp
Information codée sur 8 bits			Information codée sur 4 bits		

Tableau 1 - *Exemple de configuration de codage pour le codeur V-WTSOM*

Un des apports de ce travail réside dans la robustesse des VM au bruit de transmission, sans protection particulière. La figure 4 est un exemple de transmission permettant de mettre en avant la robustesse des VM. Dans cet exemple, seuls les flux codant les VM ont été bruités (fig. 4).

(a) *(b)*

Figure 4 - *Illustration de la robustesse des VM (Foreman_40) ; (a) TEB des VM = 0 ; PSNR moyen = 25,4dB ; (b) TEB = 1,4.10^{-1} ; PSNR moyen = 23,7dB ; Tc = 0,114 bpp*

On remarque que de légères distorsions apparaissent. Elles s'expliquent par le fait que le bruit peut changer la valeur d'un VM. Cela signifie qu'au lieu de choisir le bloc optimal pour décoder une zone de l'image, l'algorithme sélectionne un bloc dans son voisinage immédiat (relativement à l'importance du bruit). On peut espérer que localement les différences entre blocs sont minimes ce qui a pour effet de minimiser l'impact d'une erreur de transmission.

3 Stratégie de transmission MIMO

3.1 Principe du MIMO

Face aux contraintes propres du canal de transmission telles que la mobilité ou les phénomènes de multitrajets, des solutions multi-antennaires ont été développées. Le principe de base des systèmes multi-antennes baptisés MIMO (Multiple Input Multiple Output) [8] consiste à exploiter au maximum la diversité de transmission qu'elle soit spatiale, temporelle ou fréquentielle (fig. 5).

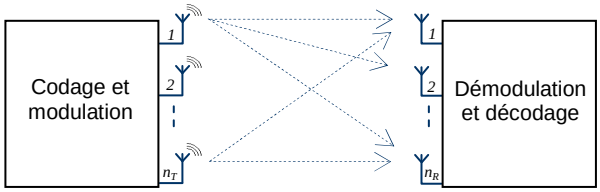


Figure 5 - *Principe d'un système MIMO*

Ces systèmes permettent d'envisager deux stratégies. On peut optimiser le débit en décomposant le flux à transmettre en n_T sous-flux transmis simultanément. On parle alors de démultiplexage spatial (LST pour Layered Space Time). On augmente ainsi le débit d'un facteur n_T. On peut également favoriser la robustesse de la transmission en transmettant n_T fois le flux originale. On parle alors de Codage Spatio-Temporel (CST). Dans ce cas de figure, le débit correspond au débit d'un lien SISO mais la robustesse du lien est améliorée. Des solutions existent également permettant de mixer ces deux approches.

3.2 Utilisation de précodeurs MIMO

Il existe beaucoup de configurations de systèmes MIMO. Ils peuvent être classés en deux familles suivant que l'on connaisse l'état du canal à l'émission ou à la réception. Cette connaissance est obtenue par le biais de séquence d'apprentissage. Si l'on dispose uniquement de l'état du canal à la réception, on parlera de système en boucle ouverte (OL-MIMO pour Open Loop MIMO). Dans ce cas, nous ne pouvons envisager qu'une stratégie de transmission aveugle. La solution la plus efficace consiste alors à transmettre avec une puissance équivalente sur chacune des antennes émettrices. Au contraire, si l'on a connaissance de l'état du canal à l'émission, on parle de système MIMO en boucle fermée (CL-MIMO pour Closed Loop MIMO). Nous pouvons alors envisager d'adapter la puissance d'émission sur chaque antenne afin d'optimiser la transmission de l'information. C'est le rôle des précodeurs diagonaux MIMO. Connaissant l'état du canal, les précodeurs peuvent décomposer le canal MIMO

en sous-canaux virtuels SISO (Single Input Single Output). Ces sous-canaux parallèles et indépendants présentent alors des caractéristiques propres (capacité, TEB, SNR…). Les précodeurs cherchent à optimiser la transmission de l'information suivant un critère et fournissent ainsi une pondération sur chacune des voies virtuelles SISO. L'utilisation de précodeurs diagonaux permet ainsi d'envisager une allocation intelligente de l'information et se prête donc particulièrement bien à la transmission d'un flux hiérarchisé.

3.3 Stratégie d'allocation V-WTSOM/MIMO

Le flux issu du codeur V-WTSOM est un flux hiérarchisé. En effet, il existe une hiérarchie entre les images issues du codage en GOP et également une hiérarchie entre les sous-bandes d'ondelette. Nous considérons dans cet article un système MIMO 4 × 4 (quatre antennes à l'émission et à la réception). Nous disposons alors de quatre voies pour transmettre notre flux V-WTSOM. La stratégie d'allocation consiste à décomposer notre flux en couches de qualité hiérarchisées. Si le canal est bon, nous aurons donc accès à toute l'information et donc à la qualité maximale, alors que si le canal est dégradé, nous favoriserons une qualité de base assurée par la transmission d'une seule couche. Le schéma ci-dessous illustre la décomposition du flux à transmettre et leur allocation sur les sous-canaux SISO (fig. 6) :

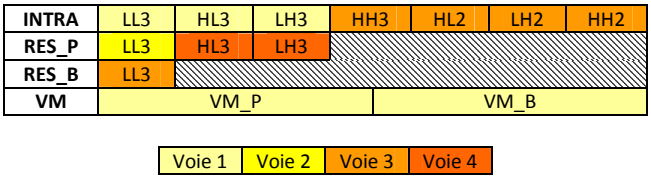


Figure 6 - *Allocation du flux V-WTSOM sur les différentes voies virtuelles SISO*

Ce découpage a été réalisé pour obtenir quatre sous-flux de taille équivalente correspondant à des couches de qualité successives. Pour décoder la couche *n*, il est nécessaire de disposer de toutes les couches inférieures à *n*. Ainsi la couche 1 constitue la couche de base indépendante des autres. Toutefois, les caractéristiques propres à V-WTSOM imposent d'obtenir les informations des voies 1 et 2 pour obtenir une vidéo reconstruite d'une fluidité optimale. Les voies 3 et 4 représentent ainsi des couches de qualité supérieure exploitables lorsque le canal est jugé bon.

4 Résultat de simulation

4.1 Chaîne de simulation

Les simulations ont été réalisées par le biais de la chaîne de transmission présentée sur la figure 7. Elle est initialement utilisée avec le codeur H.264/SVC et détaillée dans [8]. Le principal changement par rapport à la chaîne originale concerne la suppression des CCE. La vidéo originale est codée suivant l'algorithme V-WTSOM avec les paramètres correspondant au tableau 1. Ce codage fournit un flux hiérarchisé et découpé en 4 sous-flux comme indiqué sur la figure 6. Le précodeur utilisé est WF (Water-Filling). Pour la phase de précodage, nous supposons une connaissance parfaite de l'état du canal (Tx-CSI (fig. 7) pour Transmitter Channel State Information). La modulation MAQ-16 est couplée à la technique de modulation multi-porteuse OFDM dont les paramètres, séquences pilotes et d'apprentissage correspondent à la norme 802.11n [9]. Du coté du récepteur, le décodage est assuré par un décodeur à Maximum de Vraisemblance (MV) avant démodulation OFDM et MAQ-16. Le canal de transmission est un canal sans mémoire dont les coefficients sont tirés suivant une loi de Rayleigh. De plus la transmission est perturbée par un bruit blanc additif gaussien de moyenne nulle et de variance égale à 1.

4.2 Intérêt de l'allocation MIMO

La figure 8 permet de constater l'apport en qualité d'une allocation MIMO utilisant le précodeur WF en comparaison de la stratégie SISO classique. Ce précodeur permet de pondérer les différentes voies en maximisant la capacité du canal. Ces résultats ont été obtenus sans ajout de CCE (fig. 8) :

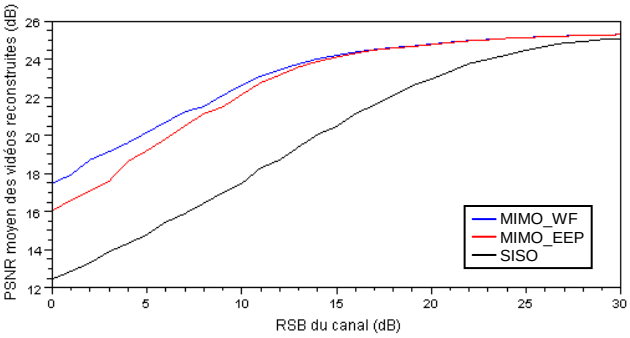


Figure 8 - *Evolution du PSNR moyen en fonction du SNR pour la vidéo Foreman*

Dans le cadre de la stratégie MIMO, nous utilisons quatre voies de données, le débit est donc quatre fois supérieur à la transmission SISO. L'information allouée aux différentes voies permet d'exploiter la hiérarchisation des sous-flux de données V-WTSOM en fonction de la qualité du canal de transmission. Ceci explique pourquoi les courbes MIMO donnent de meilleurs résultats que la courbe SISO. La stratégie permettant d'allouer la puissance sur les différentes voies permet d'augmenter la qualité visuelle par rapport à une transmission MIMO utilisant une pondération équivalente sur chaque voie (MIMO_EEP).

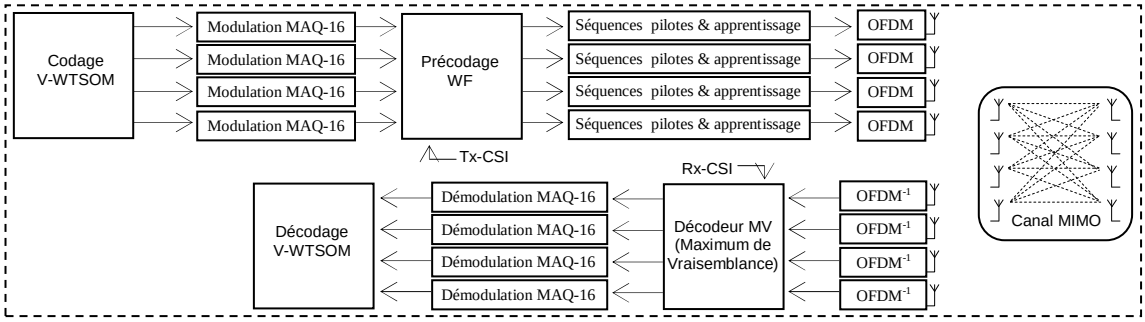


Figure 7 - *La chaîne de simulation*

Cette adéquation entre le contenu à transmettre et la technique de précodage permet d'assurer une qualité de service minimale dans les conditions les plus sévères du canal. Lorsque le canal est bon, les stratégies MIMO et SISO tendent vers une qualité visuelle équivalente.
L'accessibilité à une couche de donnée est définie par rapport au TEB. V-WTSOM faisant preuve d'une robustesse importante, le seuil d'accessibilité d'une couche a été fixé à 2.10^{-1} de TEB (pour cet exemple). La figure 9 permet de constater l'apport en qualité de ces différentes couches de données dans le cadre de la transmission MIMO_WF (fig. 9) :

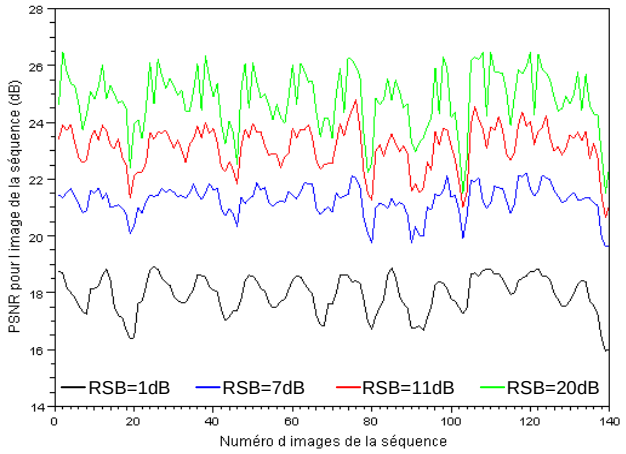


Figure 9 - *PSNR des images de la vidéo Foreman suivant les couches de qualité décodées*

On valide bien que la qualité de la vidéo augmente graduellement suivant la disponibilité des sous-flux à la réception. Une qualité de base est assurée même à bas débit. Puis, suivant la QoS désirée ou bien les ressources disponibles, nous pouvons envisager une qualité graduée des vidéos reconstruites. Ces résultats confirment également la robustesse dont fait preuve le codeur V-WTSOM. Malgré l'absence de CCE, nous constatons une augmentation de la qualité malgré un seuil d'acceptation des flux à 2.10^{-1} de TEB.

5 Conclusion

Le codeur V-WTSOM a pour but de transmettre des vidéos de type vidéoconférence de manière robuste sans ajout de CCE. Ce codeur s'appuie sur la compensation de mouvement par champs de vecteurs et sur des transformées en ondelette. Nous avons proposé une méthode robuste pour le codage des VM puis nous avons exploité le flux fortement hiérarchisé de V-WTSOM sur un canal MIMO en utilisant un précodeur autorisant une stratégie d'allocation par couches de qualité. Les résultats issus des simulations ont ainsi pu mettre en avant l'intérêt d'utiliser le MIMO en tenant compte du contenu de l'information à transmettre. L'utilisation modérée de CCE ne pourra qu'améliorer les performances de V-WTSOM.

Remerciement

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence ANR-08-VERS-002.

Références

[1] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "*Overview of the H.264-AVC Video Coding Standard*", ", in IEEE Trans. Circuits Syst. Video Technol., vol. 13, n° 7, pp. 560-576, July 2003

[2] N. Adami, "*State-of-the-art and trends in scalable video compression with wavelet-based approaches*", in IEEE Trans. Circuits Syst. Video Technol., vol. 17, n°9, pp. 1238-1255, September 2007

[3] M. Cagnazzo, T. André, M. Antonini, M. Barlaud, "*A model-based motion compensated video coder with JPEG2000 compatibility*", in Proc. IEEE Int. Conf. Image Process., pp. 2255-2258, Singapore, October 2004

[4] S. Benayoune, N. Achir, K. Bousseta, K. Chen, "A *study of H.264/AVC Robustness Over a Wireless Link*", Signal Processing and Its Applications, pp. 1-4, Feb. 2007

[5] S. Kumar, L. Xu, M. K. Mandal, and S. Panchanathan, "*Error resiliency schemes in H.264/AVC standard*", Journal of Visual Communication and Image Representation, vol. 17, n° 2, pp. 425-450, Apr 2006

[6] C. Chatellier, H. Boeglen, C. Perrine, C. Olivier, O. Haeberle, "*A robust joint source channel coding scheme for image transmission over the ionospheric channel*", Signal Proc. Image Communication, vol. 22, n° 6, pp. 543-556, July 2007

[7] A. Barjatya, "*Block Matching Algorithm for motion estimation*", in Technical Report, Utah State, 2004

[8] W. Hamidouche, C.Perrine, Y.Pousset, C.Olivier, "*Optimal solution for SVC-based video transmission over a realistic MIMO channel using precoder design*", IEEE ICASSP, March 2010

[9] IEEE Standard for Information Technology-Part 11 : *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) specifications Amendment : Enhancements for Higher Throughput (802.11n)*, 2009

Visualisation 3D d'un système de particules issu de capteurs de température

B. LANGE[1] N. RODRIGUEZ[1] W. PUECH[1] H. REY[2] X. VASQUES[2]

[1] LIRMM (Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier)

Université Montpellier 2, LIRMM - UMR 5506 - CC 477
161 rue Ada, 34095 Montpellier Cedex 5 – France
{benoit.lange, nancy.rodriguez, william.puech}@lirmm.fr

[2] IBM, Montpellier
Rue de la vieille Poste - BP 1021
34006 Montpellier
{REYHERVE, xavier.vasques}@fr.ibm.com

Résumé

Cet article traite de la visualisation 3D interactive de données issues de capteurs situés dans le centre de calcul IBM Montpellier (data center). Il est important de visualiser ces données dans le but d'analyser et par la suite réduire la consommation d'énergie totale dans des data centers. Un moteur de visualisation a été développé afin de pouvoir comprendre, en temps réel, le comportement des capteurs. Ce moteur se base sur un système de particules qui remplit et représente l'espace d'une salle de calcul. Pour cela, nous utilisons les diagrammes de Voronoï et la triangulation de Delaunay afin de pondérer l'influence de chacune des particules en fonction de leur distance aux capteurs. La visualisation d'une salle en 3D avec un grand nombre de données issues des capteurs est possible grâce à cette solution. La contrainte principale de notre visualisation 3D est l'interactivité du système qui doit garantir un meilleur suivi de l'évolution des mesures prises par les capteurs.

Mots clefs

Capteurs, diagramme de Voronoï, triangulation de Delaunay, visualisation temps-réel, système de particules.

1 Introduction

Dans cet article, nous présentons une nouvelle méthode que nous avons développée pour modéliser et visualiser en 3D des données issues de capteurs de température d'un data center. De manière plus générale, ces capteurs peuvent être de différents types et prendre des mesures telles que la température, la pression, le bruit, l'humidité et même le poids des personnes. La modélisation et la visualisation de cet ensemble de capteurs doit permettre de mieux appréhender le comportement général du data center. Dans ce contexte, nous avons identifié plusieurs points clés pour notre système : fiabilité, rapidité d'exécution, qualité de rendu et fluidité de l'interaction. Dans leurs travaux, Madougou *et al.* [1] parlent de temps réel pour les applications ayant un nombre d'images par seconde de l'ordre de 30 FPS (Frames Per Second). Mais pour d'autres applications, une définition plus stricte du temps réel est souvent nécessaire. Par exemple, les contrôleurs de jeux ou caméras peuvent avoir un rafraichissement bien plus important. La caméra de la PlayStation 3 produit une acquisition de l'ordre de 60 FPS tandis que la caméra de la wiimote a 100 FPS.

Un des problèmes récurrents de la visualisation de données issues de capteurs est l'interpolation de celles-ci afin de prendre en compte tout l'espace et non seulement le champ d'action du capteur. Dans cet article nous présentons deux solutions que nous avons explorées pour résoudre ce problème.

Nous présentons dans la section 2 des travaux en rapport avec notre problématique. En section 3, nous détaillons notre proposition. En section 4, nous présentons les premiers résultats obtenus à partir de données réelles. Enfin en section 5, nous parlons de visualisation interactive et en section 6, nous présentons nos conclusions et perspectives.

2 État de l'art

Dans cette section, nous présentons différents travaux traitant de la visualisation de données et des solutions mathématiques dont nous nous sommes servis dans notre approche.

2.1 Solutions de visualisation

Dans cette partie, nous présentons des solutions qui exploitent des données de grande taille. Les premiers travaux que nous avons étudiés utilisent des données issues de capteurs qui analysent les mouvements du manteau terrestre. Le flux de données est si important qu'il est très complexe

de le visualiser correctement [2, 3]. Ainsi les auteurs souhaitent réaliser une application interactive qui affiche ces données en 3D. Pour la visualisation, ils disposent d'une salle équipée de plusieurs écrans (10 au total). Pour traiter le flux de données, les auteurs ont besoin de deux calculateurs haute performance (High-performance computing : HPC) : un pour les données et un pour le rendu. Le traitement se veut donc en temps réel. L'apport principal de cet article est le traitement des données sur des architectures parallèles. Les auteurs mettent en évidence le fait de traiter des données et de les stocker en mémoire en fonction de leur usage. Par exemple chaque processeur du HPC possède un fichier qui lui est propre, celui-ci lui permet de savoir quelles données il doit traiter. Le résultat n'est toujours pas en temps réel mais des optimisations sont prévues.
Certaines applications utilisent des techniques basées sur un système de particules pour la visualisation [4]. Kapferer *et al.*, expliquent la solution qu'ils ont utilisée afin de visualiser les résultats de simulations astronomiques. Leur contrainte principale est la visualisation en temps réel, mais la qualité du rendu fait aussi partie des points clés. En effet, certaines visualisations ou algorithmes suppriment des caractéristiques importantes qui auraient dû être visualisées. Le problème posé ici peut être résolu avec un calcul par GPU (Graphics Processing Unit). Dans cet article, les auteurs se servent de la puissance de calcul des GPU afin de traiter des données massives au lieu d'utiliser un HPC. Les auteurs font des tests sur différents frameworks qui ne leur donnent pas entière satisfaction d'un point de vue visuel. Afin de dessiner les étoiles, les auteurs se servent des points à la place des sphères habituelles. Ceci permet un affichage plus dense par rapport aux solutions existantes. Tous les calculs sont effectués dans le GPU afin d'approximer les trajectoires de chaque particule. Ensuite, le système de visualisation sur écran large leur donne la possibilité de voir les données de façon plus détaillée que sur un ordinateur classique. Différents algorithmes sont utilisés pour la visualisation comme des algorithmes de simplification par cellule ou par extraction d'iso surface. La solution par GPU semble dans leur cas la plus adaptée pour les simulations d'objets en déplacement.

2.2 Solutions mathématiques

Dans cette section, nous étudions des solutions mathématiques de partitionnement et des solutions d'inclusion de points dans une forme. Dans un premier temps, nous présentons brièvement l'extraction de diagramme de Voronoï et la triangulation de Delaunay [5]. Nous étudierons ensuite l'inclusion de points dans une forme simple.
Les diagrammes de Voronoï et la triangulation de Delaunay font partie des outils retenus pour notre solution. Différentes librairies fournissent ce type d'extraction mathématique dont la librairie QHULL. Barber *et al.* présentent la librairie QHULL, une solution qui sert à produire des enveloppes convexes pour un maillage 3D [6]. Les premiers résultats montrent que le coût de calcul est moins important que lors de l'utilisation d'autres algorithmes. Il est possible d'utiliser les méthodes de la librairie QHULL avec différentes approches mathématiques comme la triangulation de Delaunay ou l'extraction de diagramme de Voronoï. Le coût en temps de calcul est l'un des plus faibles des librairies existantes. Une autre librairie étudiée est décrite par Rycroft [7] extrait des diagrammes de Voronoï sur des objets de grande taille.
Dans cette partie, nous décrivons l'inclusion d'un point dans un maillage. Cette technique permet de connaitre l'emplacement d'un point par rapport à une forme géométrique. L'article de Li *et al.* [8] présente un benchmark entre divers algorithmes de test d'inclusion. Il définit un certain nombre de méthodes tels que grid, octree et bsp-tree. Au final, les résultats présentent l'algorithme de Feito et Torres [9] comme le plus stable et robuste pour la détection de l'inclusion d'un point dans un polygone. Il est inspiré de la triangulation de Delaunay et introduit la notion d'arêtes visibles. Il démontre aussi comment calculer l'inclusion d'un point dans un polygone.
John M. Snyder introduit en 1987 le principe du Ray tracing. Il présente les aspects mathématiques de cette solution dans [10]. Il détermine qu'un modèle peut être vue sous une forme hiérarchique et qu'il est possible de lui appliquer récursivement les rayons de lumières sur chacune de ces facettes. Il introduit diverses équations afin de reproduire les effets de la lumière sur un objet de grande taille. Il présente aussi différentes implémentations à travers une liste ou une grille et présente les résultats. Ces résultats ont largement démontré leur intérêt, par exemple, il montre une forte corrélation entre le rendu du Ray tracing et le nombre de polygones d'un maillage.

3 Proposition

Dans cette section, nous développons les différentes propositions que nous avons faites pour la visualisation de données issues de capteurs de température.
Notre problématique est de visualiser l'évolution thermique dans une salle de calcul. Pour ce faire nous avons utilisé une modélisation de l'espace de notre salle de calcul avec les différents capteurs. Après nous avons utilisé des méthodes mathématiques afin d'interpoler les données issues de ces capteurs.

3.1 Modélisation

Ainsi, afin de modéliser une salle, nous avons utilisé la géométrie du data center. Elle consiste dans un pavé de taille définie par ses trois dimensions : Longueur ($L \in \mathbb{N}$), Largeur ($l \in \mathbb{N}$) et Hauteur ($H \in \mathbb{N}$) illustré en figure 1. Le data center étudié est composé de diverses couches de capteurs ($C \in \mathbb{N}$). Dans notre modèle étudié, les couches sont aux nombre de trois. Enfin sur ces couches, on dépose un nombre M de capteurs. Ceux-ci sont placés dans l'espace grâce au triplet $\{X_i, \mathrm{Y}_i, \mathrm{C}_i\}$, où $i \in \mathbb{N}$ est le numéro

de capteur, leur localisation dépend de leurs coordonnées réelles dans la pièce. Enfin pour finir la modélisation, on utilise un sytème à particules afin de modéliser l'espace de la pièce. Dans notre approche une particules est un objet déposé régulièrement $N \in \mathbb{N}$ représente le nombre de particules du système. Ces particules sont espacées de $X \in \mathbb{R}$. Leur nombre peut être calculé de la manière suivante :

$$N = \frac{((L+1)*(l+1)*(H+1))}{X^3}. \quad (1)$$

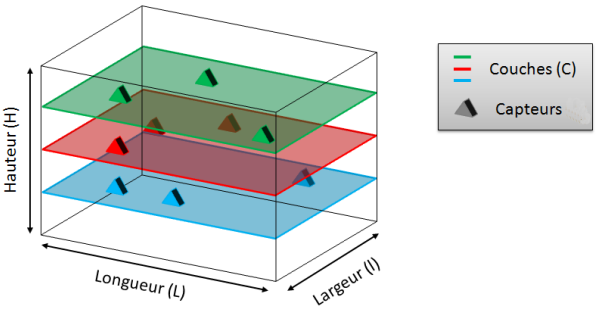


Figure 1 – *Modélisation de la salle de calcul et représentation des différentes couches de capteurs.*

3.2 Partitionement

Nos particules ne sont pas localisées par rapport aux capteurs. Pour cela, on utilise différentes méthodes afin de leur attribuer le coefficient d'influence des capteurs qui les entourent. On utilise donc deux méthodes de partitionnement : la triangulation de Delaunay et l'extraction de cellule de Voronoï.

La triangulation de Delaunay est le dual du partitionnement de Voronoï. Celle-ci va nous permettre de mieux calculer l'influence des particules en localisant entre quels capteurs elles se situent. Cette méthode est utilisée afin de pouvoir localiser les particules dans le maillage des capteurs. Le fonctionnement de Delaunay est simple, la fonction recherche les sommets les plus proches autour d'un point. La structure obtenue est un maillage triangulé. En 2D, la triangulation de Delaunay produit une série de triangles. Dans un monde 3D, il ne s'agit plus de triangles mais de tétraèdres comme le montre la figure 2.
Suite à la triangulation de Delaunay, il est nécessaire de faire une analyse des particules par rapport aux tétraèdres. Cette étape a pour but d'identifier à quel tétraèdre appartient une particule, c'est à dire si elle est située à l'intérieur des différents plans du tétraèdre. Cette méthode est issue de la solution introduite par le lancer de rayons [7]. Nous proposons de tester pour chaque face du tétraèdre si une particule est incluse dans le tétraèdre. Pour cela, la solution calcule les normales de chaque face, puis elle les projette à partir de la particule. Si au moins 3 rayons coupent 3 faces alors le point appartient au tétraèdre. Les particules

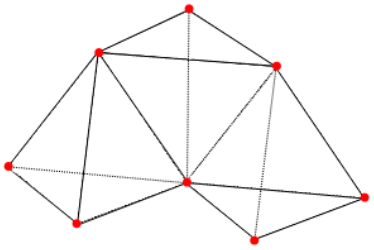


Figure 2 – *Maillage 3D.*

n'appartenant à aucun tétraèdre utiliseront le diagramme de Voronoï pour être prises en compte.
La complexité de cette algorithme est également très élevée, de l'ordre de O (N*M*4), avec N$\in \mathbb{N}$ le nombre de particules et M$\in \mathbb{N}$ le nombre de capteurs.

L'autre méthode utilisée pour les particules externes au maillage de capteurs est l'extraction des cellules de Voronoï. Cette méthode expliqué dans [5] est un système de partitionnement de l'espace. Ce diagramme est surtout utilisé en 2D, mais des solutions existent aussi δD ($\delta \in [0, +\infty[$. Dans le cadre d'un système à particules, cette technique peut être assimilée à une collision entre une sphère et un point. De plus la méthode est discrète contrairement aux méthodes de l'état de l'art. L'algorithme 1 présente cette solution. Afin de pondérer les particules,

Algorithm 1 Extraction de cellule de Voronoï

```
for all particules do
  for all capteurs do
    calculer distance particule-capteur
    if capteur dans sphère d'influence then
      Ajouter capteur à particule
    end if
  end for
end for
```

nous avons utilisé la formule suivante :

$$Mesure\,(p) = \frac{\sum_{i=1}^{M} Mesure\,(i)}{M}, \quad (2)$$

où p représente une particule.
Malheureusement, la complexité de cet algorithme reste très élevé, de l'ordre de O(N*M). La figure 3 illustre les artefactes qui apparaissent lors de l'utilisation de cette solution. Le partionement reste très cubique.

4 Résultats

Dans cette section, nous illustrons la visualisation des données issues d'un centre de calcul IBM. D'abord, il est indispensable de réaliser le modèle géométrique de la salle

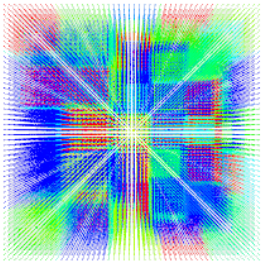


Figure 3 – *Visualisation des cellules de Voronoï en 3D.*

pour la visualiser. La première intuition est de réaliser un maillage. Le problème avec cette solution est qu'elle ne permet pas de modéliser convenablement l'intérieur. Les parties vides ne pourront jamais être modélisées. Deux autres solutions d'affichage de modèle 3D existent : les voxels et les nuages de points. Ces deux manières de modéliser un objet sont très proches. La figure 4 représente le nuage de points que nous avons utilisé pour modéliser l'intérieur d'une salle. Afin de modéliser les différences de températures, une échelle de couleur a été mise en place. Les couleurs chaudes sont représentées en rouge et les couleurs froides en bleu. Un dégradé, passant par le vert et le jaune, est appliqué aux températures médianes. Les pa-

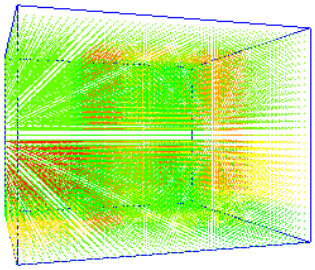
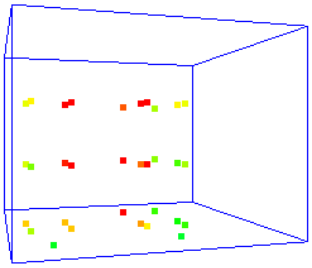


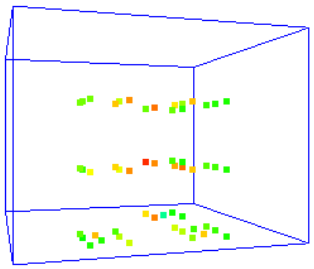
Figure 4 – *Visualisation des particules dans une salle.*

ramètres de notre modèle géométrique sont la longueur, la hauteur et la largeur. Les formes complexes que peut prendre une pièce ne sont pas encore modélisables. De plus l'intérieur de la pièce est considéré comme vide, sans mobilier. Dans notre prototype, nous avons modélisé deux salles de calcul de taille : 3 mètres * 4 mètres * 3 mètres illustrées figure 5.a et 5.b. Les capteurs sont placés dans la pièce sur trois couches. Les premiers sont installés au ras du sol, puis la seconde couche est à un mètre et enfin la dernière à deux mètres du sol. Les particules sont déposés tous les 10 centimètres de manière régulière. Ainsi 36000 particules par salle sont placées. Chaque salle possède son implantation unique des capteurs. Le seul paramètre stable est le fait d'avoir trois couches.

Pour la deuxième solution, les particules sont dans un



(a) Salle 1



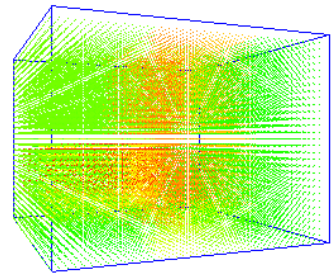
(b) Salle 2

Figure 5 – *Disposition des capteurs.*

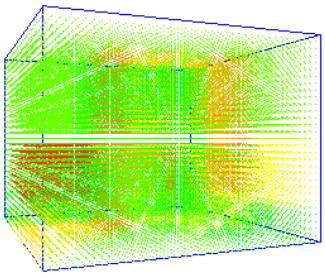
premier temps localisées et pondérées grâce à une triangulation de Delaunay. Pour produire cette triangulation, nous avons utilisé la méthode de QHULL : Qdelauney [6]. L'intersection des particules par rapport aux quatre plans de chaque tétraèdre produit est testée. Si la particule coupe au moins 3 plans, alors celle-ci est considérée comme intérieure au tétraèdre. Pour les particules restantes, nous utilisons le diagramme de Voronoï comme décrit dans la section 3.3. Les figures 6.a, 6.b et 6.c illustrent cette solution.

Afin de segmenter l'espace par Voronoï, nous avons dans un premier temps utilisé les librairies : QHULL [6] et Voro ++ [7]. Les premiers tests n'ont pas été satisfaisants, le système de particules s'adapte mal aux deux solutions présentées. Un coût supplémentaire qui consiste à identifier chaque particule dans les polygones issus de l'extraction est nécessaire.

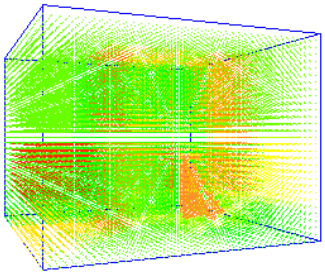
Nous avons donc décidé d'implémenter une nouvelle solution de construction de diagramme de Voronoï qui permet de travailler directement sur les particules. Cette solution est coûteuse en temps CPU. Toutes les particules sont testées et sont comparées aux capteurs. Cette méthode est insuffisante dans le cadre d'une simulation réaliste de température car il n'est pas trivial de pondérer efficacement les particules en fonction des capteurs. Seules les particules centrales à une jointure entre plusieurs zones se trouvent avec une pondération efficace. Les figures 7.a, 7.b et 7.c montrent la pondération des particules avec cette méthode. Le rendu du partitionnement devient alors très intéressant.



(a) temps = 0

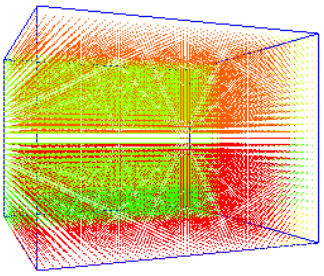


(b) temps = 1

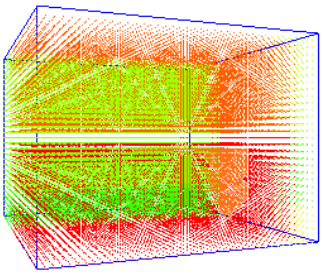


(c) temps = 100

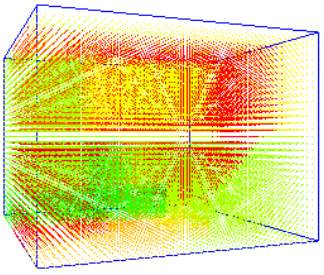
Figure 6 – *Visualisation avec des cellules de Voronoï.*



(a) temps = 0



(b) temps = 1



(c) temps = 100

Figure 7 – *Triangulation de Delaunay et cellule de Voronoï.*

On se retrouve avec des espaces plus fins et des particules pondérables de manière plus simple qu'avec la solution basée que sur les cellules de Voronoï. Les quatre sommets auxquels la particule appartient sont connus. Un simple calcul de norme totale (distance de la particule avec chaque sommet) est nécessaire pour déterminer le coefficient de pondération. Enfin, on applique l'inverse de la distance à chaque coefficient.

La qualité visuelle est améliorée par cette nouvelle approche. Cependant le coût en terme de calcul au lancement de l'application reste important. Cette proposition devra être améliorée. De plus de nombreuses autres possibilités sont à mettre en œuvre comme de véritables modèles physiques venant des études thermiques de bâtiments.

Enfin, au cours de notre implémentation, une mise en évidence de certaines zones de températures s'est avérée importante. Ainsi, un seuil réglable a été mis en place afin de visualiser des intervalles de températures par exemple les températures les plus faibles ou les plus élevées comme illustré figure 8.

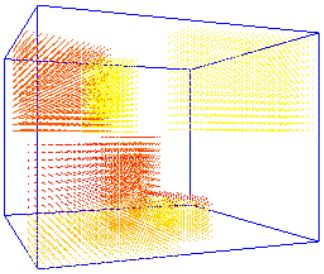


Figure 8 – *Visualisation des températures les plus élevées de la salle de calcul.*

Cette technique apporte une nouvelle approche en matière

de simulation de données issues de capteurs. De nombreuses approches étaient basées sur des systèmes de cartographie de couleur afin de simuler les variations de température. Ici, des modèles mathématiques sont utilisés basés sur la triangulation de Delaunay et les cellules de Voronoï.
Cette solution permet d'avoir un meilleur suivi visuel des incidents qui peuvent intervenir dans une salle de calculs. Les températures critiques peuvent ainsi être appréhendées plus facilement au travers de cette visualisation.

5 Visualisation Interactive

Dans le cadre d'une visualisation interactive des températures d'une salle de calcul, nous avons comme contrainte un travail sur le temps réel. Avec l'approche présentée dans cet article, le rafraichissement de l'affichage a un coût trop important et ne peut se faire rafraîchir qu'à un taux de 11 FPS. Ce coût est dû au trop grand nombre de particules présentes dans la scène. Des optimisations futures devraient permettre de gagner du temps de calcul.
La base de données que nous utilisons ne permet pas non plus de fournir des données en temps réel (de l'ordre de 40 ms). Les capteurs actuels disposés dans les salles de calcul IBM, nous renvoient seulement une information toutes les 5 minutes. En travaillant sur des données simulées, le taux de rafraîchissement minimum est de l'ordre de 300 ms. La complexité pour générer les nouvelles températures est trop élevée. Deux problèmes peuvent se poser : si le flux de données est trop important pour un rafraîchissement en temps réel ou si le flux de données est trop faible afin de donner un rendu intéressant. Le premier problème a été soulevé lors de simulations et l'autre lors de l'utilisation de données issues de capteurs.

6 Conclusions

Le système de particules présente dans cet article permet de modéliser cette approche une salle de calcul et d'interpoler les mesures de capteurs afin d'obtenir une visualisation 3D des températures. Cette approche permet une interpolation des températures de manière fidèle. L'utilisation de partitionnement mathématique fournit un premier aperçu des possibilités du système. L'extraction du diagramme de Voronoï n'est pas suffisamment adaptée à un rendu correct. Seules les particules entre les cellules ont une pondération efficace. Extraire le maillage des capteurs permet de produire une série de tétraèdres, il suffit d'extraire la position de la particule par rapport aux quatre capteurs les plus proches. Ensuite, pour les particules qui ne sont pas retenues grâce à cette solution, un diagramme de Voronoï est utilisé pour les prendre en compte.
Différentes améliorations sont à prévoir pour notre système. L'utilisation d'un véritable moteur physique afin de définir la température des particules est indispensable. Actuellement les particules n'ont pas d'interpolation dans le temps. Et les données sur lesquelles nous travaillons nous donnent des rafraîchissements toutes les cinq minutes. Ensuite, traiter les données de différents types est indispensable. Actuellement, les données sont totalement homogènes alors que dans le futur les données fournies seront de différents types. Une autre amélioration est de rendre plus fluide l'affichage. Du fait du grand nombre de données, nous avons 11 FPS. Ceci est loin des 24 FPS prévues pour obtenir une visualisation interactive et fluide. Enfin, le manque d'interaction ne permet pas de manipuler facilement l'environnement. Le clavier souris limite grandement celle-ci. Des contrôleurs comme une Wiimote sont à prévoir pour faciliter l'usage de notre visualisateur.

7 Remerciements

Nous tenons à remercier toute l'équipe d'IBM Montpellier, et surtout le PSSC (Products and Solutions Support Center) qui nous a fourni les données et le matériel pour les tests présents et futurs.

Références

[1] S. Madougou, V. Gouranton, et E. Melin. Vers une gestion efficace de donnÃ©es gÃ©oscientifiques complexes pour la visualisation sur grappe de pc. *AFRV*, 2006.

[2] M. Damon, M. Kameyama, M. Knox, D. Porter, D. Yuen, et E. Sevre. Interactive visualization of 3d mantle convection. *Visual Geosciences*, 2008.

[3] Kirk E. Jordan, David A. Yuen, David M. Reuteler, Shuxia Zhang, et Robert Haimes. Parallel interactive visualization of 3d mantle convection. *IEEE Comput. Sci. Eng.*, 3(4) :29–37, 1996.

[4] W Kapferer et T Riser. Visualization needs and techniques for astrophysical simulations. *New Journal of Physics*, 10(12) :125008 (15pp), 2008.

[5] A. Montanvert et JM. Chassery. *GÃ©omÃ©trie discrÃ¨te en analyse d'images*. HermÃ¨s, 1991. ISBN 978-2-7462-1643-3.

[6] C. Bradford Barber, David P. Dobkin, et Hannu Huhdanpaa. The quickhull algorithm for convex hulls. *ACM Trans. Math. Softw.*, 22(4) :469–483, 1996.

[7] C. H. Rycroft. Voro++ : a three-dimensional voronoi cell library in c++. *Chaos 19*, 2009. Lawrence Berkeley National Laboratory.

[8] Weishi Li, Eng Teo Ong, Shuhong Xu, et Terence Hung. A point inclusion test algorithm for simple polygons. 3480 :769–775, 2005.

[9] F. R. Feito et J. C. Torres. Inclusion test for general polyhedra. *Computers & Graphics*, 21(1) :23–30, 1997.

[10] John M. Snyder et Alan H. Barr. Ray tracing complex models containing surface tessellations. *SIGGRAPH Comput. Graph.*, 21(4) :119–128, 1987.

Utilisation de la géométrie de la scène pour l'analyse du trafic routier

M. Brulin[1,2], H. Nicolas[1]

C. Maillet[1,2]

[1] LaBRI (Laboratoire Bordelais de Recherche en Informatique)
Université de Bordeaux I,
351, Cours de la Libération,
33405 Talence Cedex – France

{brulin, nicolas}@labri.fr

[2] Adacis Sarl,
5, Ferreau Sud,
33320 Bayas – France

maillet@adacis.net

Résumé

Cet article présente un système d'analyse de trafic routier dans un contexte de vidéo-surveillance. La première étape de notre système consiste à extraire des informations sur la géométrie de la scène (position et caractérisation des voies). Les bordures des voies, l'estimation de la profondeur dans l'image et les informations de mouvement sont ensuite utilisées pour aider à la segmentation des objets et à leur suivi image après image. Nous nous plaçons dans le cas d'une seule caméra sans connaissance a priori sur les paramètres intrinsèques et extrinsèques la caractérisant.

Mots clefs

Analyse du trafic routier, modélisation de la scène, suivi d'objets.

1 Introduction

L'analyse du trafic routier par vision par ordinateur est un domaine de recherche en pleine expansion dont les applications sont essentielles pour la gestion, la sécurité et la compréhension du trafic. L'utilisation d'une ou plusieurs caméras est un choix économique qui offre les données nécessaires à de nombreuses fonctions, telles que le comptage et la détection d'anomalies ou d'accidents. Malgré les nombreuses recherches dans le domaine (ex. [1], [2]), l'élaboration d'un système robuste, efficace et automatique reste un challenge complet dû à la perte d'information causée par la projection perspective lors de la formation des images par le capteur. Plus récemment, des systèmes ont été proposées utilisant des caractéristiques de la route afin d'améliorer les résultats. Par exemple, Maduro et al. [3] utilisent les bordures des voies pour synthétiser une vue du dessus de la scène afin de segmenter les objets. Nous proposons ici de récolter des informations sur la scène afin d'aider à la segmentation.

La première étape de notre système consiste à estimer des informations sur la géométrie de la scène ainsi que le mouvement global. Les bordures des voies sont déterminées et l'information de profondeur est utilisée pour diviser les voies en sous-régions. La seconde étape effectue l'analyse du trafic, qui consiste en quatre phases : la soustraction d'arrière-plan, l'extraction des objets, leur suivi puis l'analyse de comportement. Les informations sur la géométrie de la scène sont exploitées dans l'extraction des objets et leur suivi en définissant à l'intérieur des objets des pixels dits *ambigus* et en découpant les objets temporairement à l'aide des délimitations des voies. Le processus de suivi prend en compte cette ambiguïté en utilisant une information temporaire.

Le système proposé n'utilise qu'une seule caméra placée au dessus des voies et nous ne connaissons aucune information a priori sur les paramètres intrinsèques ou extrinsèques de la caméra.

2 Modélisation de la scène

La modélisation se déroule en 4 étapes qui sont combinées afin d'obtenir une estimation de la géométrie (Figure 1). Dans un premier temps, une image d'arrière-plan est estimée et utilisée afin d'obtenir une carte de contours et en extraire les bordures des voies. Les résultats obtenus, combinés avec une détection de mouvement, permet de délimiter les zones correspondants aux voies. Le point de fuite de la scène est ensuite estimé et utilisé pour extraire les lignes de profondeur. Enfin, la dernière étape combine les zones obtenues, le mouvement et sa direction, pour

obtenir le model final des voies : chaque voie est découpée en sous-région en fonction des lignes de profondeur. Toutes ces informations sont utilisées dans la section 3 pour améliorer la segmentation et le suivi des objets et permettre le comptage.

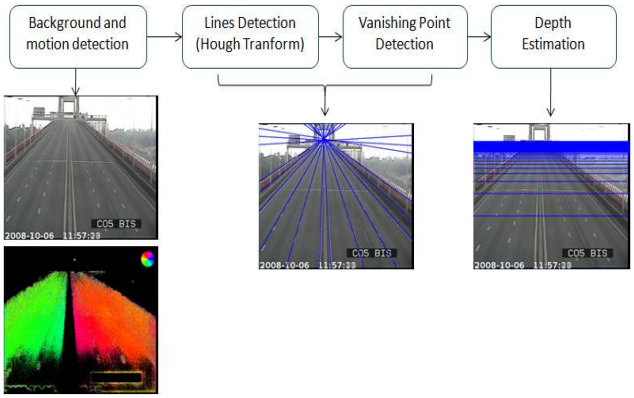


Figure. 1 – *Les quatre étapes dans la modélisation de la scène.*

Estimation de l'arrière-plan. L'information d'arrière-plan est obtenu en utilisant une modélisation par mixture de gaussiennes [4]. Cette modélisation consiste en une somme pondérée de lois gaussiennes, où les distributions couleur de chaque pixel peuvent être multi-modales (pour une étude approfondie, le lecteur intéressé pourra se référer à [5]). Le nombre de gaussiennes et leurs paramètres sont estimés et mis à jour par un algorithme Expectation-Maximization et seules les gaussiennes possédant un poids élevée et un faible écart-type sont considérées comme appartenant à l'arrière-plan (seuillage selon un critère sur le rapport poids - écart-type). Tous les pixels isolés sont supprimés et une détection d'ombre portée est appliquée en utilisant l'approche proposée par Horprasert et al. [6].

Détection de mouvement. Chaque nouvelle image de la vidéo est comparée à l'arrière-plan pour obtenir une première estimation des zones de l'avant-plan. Une étiquetage par analyse en composantes connexes [7] permet d'extraire les régions en mouvement. Pour chaque région, le flot optique est estimée en utilisant la méthode pyramidale proposée par Lucas Kanade [8]. L'ensemble des points d'intérêt [9] sont mis en correspondance dans l'image suivante, et les vecteurs de déplacement (norme et orientation) sont conservés.

Détection des voies. La détection de lignes appliquée ici est similaire à celle proposée dans [10]. Dans un premier temps et pour réduire le bruit, l'image d'arrière-plan est filtrée par un filtre médian (qui possède l'avantage de conserver les contours). L'information de contour est ensuite récupérée à l'aide du détecteur de Canny. Afin d'effectuer localement la détection, l'image contenant les contours est découpée en blocs de tailles égales. Pour chaque bloc, les lignes sont extraites en utilisant une transformée de Hough [11].

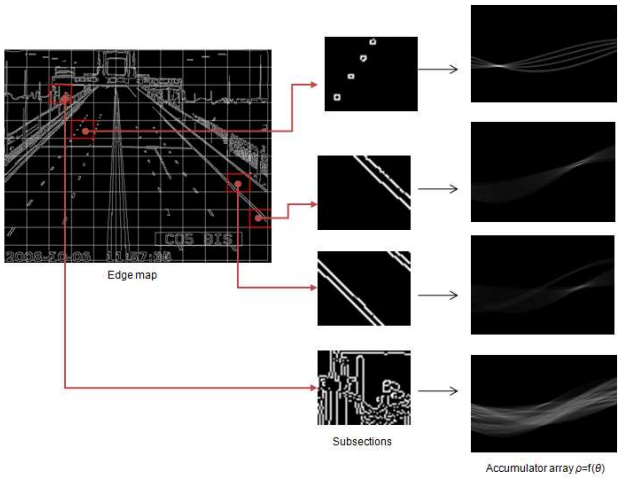


Figure 2 – *Détection locale des bordures des voies. La carte des contours est découpée en blocs homogènes. Pour chaque bloc, la matrice d'accumulation (espace de Hough) est construite (un point de l'image contour correspond à une sinusoïde dans l'espace de Hough). Par extraction des maximum locaux de l'accumulateur, on obtient les paramètres des lignes correspondantes aux bordures des voies.*

Estimation du point de fuite. En théorie, toutes les bordures de voies convergent vers un point de fuite. Il est donc intéressant d'estimer ce point pour valider et supprimer les lignes précédemment détectées. Le point de fuite est estimé en utilisant l'approche de Matessi et Lombardi [12] qui exploitent l'espace de Hough (représentation polaire). Leur méthode consiste à minimiser la fonction :

$$\min_{x_0,y_0} \sum_{i=1}^{n} W_i.(\rho_i - x_0.\cos(\theta_i) - y_0.\sin(\theta_i))^2 \quad ,$$

Où $W_i = v_i / V$, v_i est le nombre de votes dans l'espace des paramètres (lignes paramétrée par (ρ_i, θ_i)) et V est le nombre total de votes. Une fois le point estimé, les lignes qui sont éloignées de ce point (en utilisant un seuil) sont supprimées.

Estimation des lignes de profondeur. Le but de cette étape est d'estimer les lignes de profondeur dans la scène. Théoriquement, la profondeur d'un point physique (relative à la caméra) dépend de sa position verticale sur l'image [13] tel que:

$$z = h_{cam}.\tan\left(\tan^{-1}\left(\left(y_s - \frac{d}{2}\right)\Big/ f\right) + \alpha_{cam}\right)$$

où y_s représente la position du point sur le capteur, d est la longueur du capteur, f sa distance focale, z est la distance projetée du point sur le sol à la caméra, h_{cam} la hauteur de la caméra et α_{cam} son orientation. La première étape consiste à estimer α_{cam} en utilisant un algorithme des

moindres carrés. Puisque f et d sont inconnus, on suppose des valeurs standard et fixons f=35mm et d=24mm. La position de la caméra h_{cam} est seulement un coefficient qui relie la position au sol z et la position sur le capteur y_s, nous le fixons à 5m. Un exemple de lignes obtenues est affiché sur la Figure 1.

Une fois l'ensemble de ces informations collectées, elles sont combinées afin d'obtenir le modèle final des voies. Chaque voie est délimitée (par les lignes détectées) et découpée en sous-bandes à l'aide des lignes de profondeur. Les informations de mouvement (norme et orientation des vecteurs de déplacement) sont conservées pour chaque sous-bande (Figure 3). Si très peu de mouvement a été observé dans une zone (caractérisé par de faibles normes des vecteurs de déplacement), alors cette zone est ignorée (considérée à l'extérieur de notre zone d'intérêt).

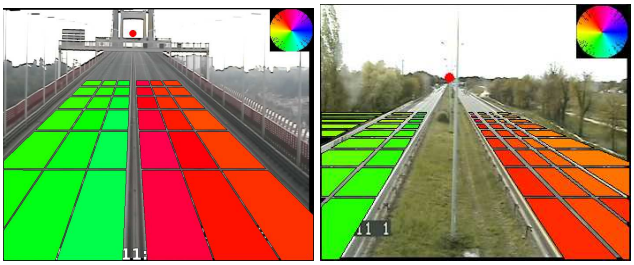


Figure 3 – *Modélisation des voies de la scène. Les sous-régions, déterminées à l'aide des lignes de profondeur et du point de fuite possèdent en théorie la même surface dans le monde réel. Les vecteurs de mouvement estimés (norme et orientation) ont été conservés et constituent une information utile pour l'étape d'analyse de comportement. La palette de couleur utilisée pour l'orientation des vecteurs mouvement est représentée en haut à droite de l'image.*

3 Analyse du trafic routier

L'analyse du trafic comprend les étapes suivantes (Figure 4) :

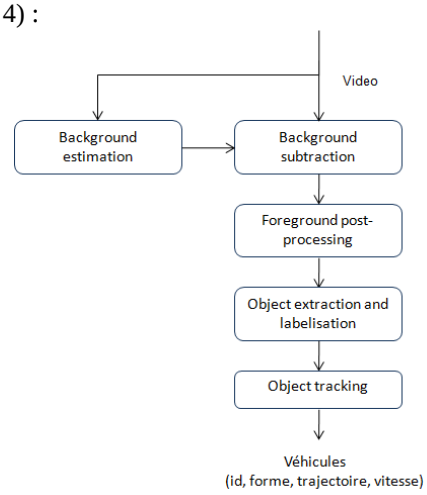


Figure 4 – *Etapes du système pour l'analyse du trafic.*

Tout d'abord, l'image d'avant-plan est obtenu par soustraction d'arrière-plan. Le résultat est ensuite filtré (suppression de l'ombre et des pixels isolés) et les objets en mouvement sont extraits en prenant en compte les bordures des voies. Le suivi des objets extraits permet ensuite de conserver l'identité des objets, ce qui permet d'en effectuer le comptage.

Extraction des objets. Une fois l'image d'avant-plan obtenue (cf. Section 2), les régions en mouvement sont extraits en utilisant un algorithme en composantes connexes [7]. La position au sol des objets est définie comme étant le centre du segment correspondant à la projection de la région sur la ligne perpendiculaire à la direction de la voie (sur laquelle se situe l'objet). Cette projection (le long de la ligne définie pour chaque point de la région et du point de fuite) nous donne un segment que nous appellerons PL (*projected line*). Pour des raisons de simplicité et pour atteindre un traitement en temps réel, chaque région est approximée par sa boite englobante, et seule sa couleur moyenne et son écart-type sont conservées. Pour de petits objets (véhicules) se déplaçant dans une voie (c.a.d. sans dépassements), la majorité des pixels sont localisés à l'intérieur de cette même voie sur l'image. Tandis que pour des objets plus volumineux (camions par exemple), un nombre plus important peuvent être sur plusieurs voies, ce qui pose problème pour la localisation de l'objet.

Figure 5 – *(en haut) Exemples d'arrière-plan obtenus en modélisant l'évolution des couleurs de chaque pixel par une mixture de gaussiennes. (en bas) Les zones estimées en mouvement sont affichées en rouge.*

Une première solution pour positionner l'objet correctement consiste à affecter l'objet à la voie dans laquelle la majorité des pixels se trouvent. Cependant, lorsque deux véhicules sont proches, ils sont souvent (par erreur) détectés comme un seul et même objet.

Pour contrer ce problème, chaque pixel des régions est classifié en deux classes : pixel *correct* ou *ambigu*, où les pixels *ambigus* sont ceux qui n'appartiennent pas au numéro de la voie qui a été associé à l'objet.

L'ambiguïté est levée en utilisant quatre critères :

1- Le rapport d'aire entre les pixels *ambigus* et la taille total de l'objet.
2- Le rapport entre la longueur du segment PL (*projected line*) et la largeur de la voie.
3- La différence euclidienne des distributions de couleur entre les pixels *ambigus* et *corrects* (un véhicule étant généralement globalement uniforme en couleur).
4- La différence entre les vitesses moyennes des pixels *ambigus* et *corrects*.

Pour chaque critère, les seuils ont été déterminés expérimentalement. Si au moins trois critères sont vérifiés, alors la région est découpée selon les bordures des voies (Figure 7).

Figure 6 – *Exemple d'extraction d'objets sur l'image d'avant-plan.*

Un véhicule qui change de voie est découpé en deux alors qu'il ne devrait pas l'être. On conserve donc l'état original du véhicule pendant le suivi afin d'être capable de revenir en arrière et corriger si nécessaire (durant l'étape de *suivi d'objets*).

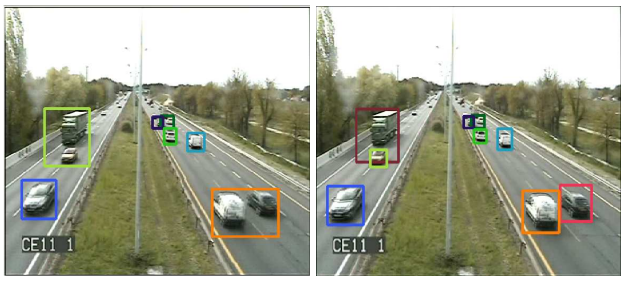


Figure 7 - *(à gauche) Exemple d'erreur de segmentation rencontrée lorsque deux véhicules sont proches. Les régions orange et verte contiennent deux objets, mais seul un seul objet est détecté. (à droite) Correction de la segmentation en utilisant les bordures des voies.*

Suivi d'objets. Le but de cette étape est de suivre un nombre variable de cibles et de maintenir leurs identités tout en prenant en compte les détections partielles et les occlusions. Le problème de suivi multi-cible est modélisé par un graphe pondéré dans lequel chaque nœud représente un objet détecté (Figure 8). Pour deux images successives, le graphe est réduit à un graphe biparti : Chaque nœud O_n^{t-1} du graphe G_{t-1} (à l'instant *t-1*) est relié à chaque nœud O_m^t du graphe G_t (à l'instant *t*). Un poids $w_{n,m}$ est associé à chaque arête du graphe et caractérise la similarité entre les objets.

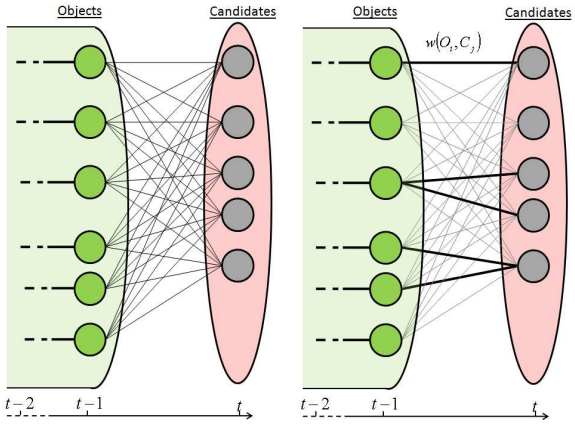


Figure 8 – *(à gauche) Modélisation du suivi d'objets par un graphe pondéré. Chaque nœud à l'instant t-1 est relié à chaque candidat potentiel à l'instant t. (à droite) Seules les arêtes dont le poids est supérieur à un seuil sont conservées.*

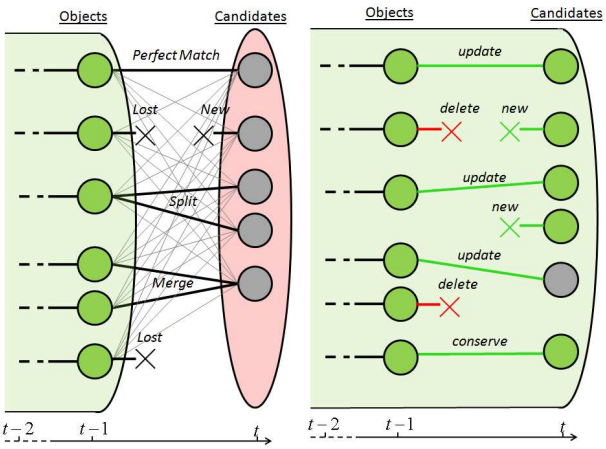


Figure 8 – *(à gauche) Hypothèses d'association parmi les cinq cas possibles. (à droite) Décision et mise à jour des objets selon les hypothèses d'association.*

Le processus d'association consiste à associer les objets suivis O_{t-1} (à l'instant *t-1*) avec les objets O_t détectés à l'instant *t*. Ce processus comporte quatre étapes :

1. La première étape est une étape de prédiction, dans laquelle la nouvelle position de l'objet est estimée à partir de sa vitesse.
2. La seconde étape est une mesure de similarité entre les objets permettant d'affecter un poids à chaque arête du graphe. Il s'agit d'une distance euclidienne entre les propriétés des objets (position, couleur, écart-type). Seules les arêtes dont le poids est supérieur à un seuil sont conservées.
3. La troisième étape concerne l'attribution d'hypothèses d'association pour chaque objet parmi les cinq cas possibles : *Correspondance Parfaite*, *Perdu*, *Nouveau*, *Split*, *Merge*.
4. La dernière étape met à jour les objets selon les hypothèses faites à l'étape 2.

Cette dernière étape (décision et mise à jour) se déroule de la façon suivante. Dans le cas d'une *Correspondance Parfaite*, l'objet est mis à jour avec le candidat correspondant. Un objet *Perdu* est supprimé s'il est *Perdu* suffisamment longtemps. Sinon, il est conservé et on continue de le suivre. Les *Nouveaux* objets sont directement ajoutés au graphe en temps que nouveaux nœud. Les cas de *Split* et de *Merge* sont résolus à l'aide d'informations temporelles, dans les deux cas, on conserve les objets originaux qui sont mis à jour avec les candidats ayant la meilleure similarité et les objets résultants sont créés et considérés comme *Nouveaux*, mais leur existence reste en suspend.

La décision finale est obtenue lorsque l'association de ces derniers ne change pas pendant un certain temps. Si deux objets sont séparés (*Split*) et se rejoignent de nouveau (*Merge*), seul le suivi de l'objet original est conservé. Deux objets qui ont subis un regroupement (*Merge*) ne sont validés que s'ils ont eu la même vitesse pendant la période de temps correspondant au déplacement d'un objet sur la moitié de la longueur de la voie (la partie supérieure de la route n'étant pas significative puisque les objets sont trop petits).

Comme mentionné dans la tâche d'*extraction d'objets*, lorsqu'un objet est découpé (par les bordures des voies), l'original est conservé et les objets résultants sont suivis indépendamment. Pour corriger les susceptibles erreurs de découpage, on utilise l'estimation de leur vitesse. S'ils ont conservés la même vitesse, alors ils sont de nouveaux regroupés. Le résultat de la détection d'objets et leurs trajectoires est montré sur la Figure 9.

Remarque. Les véhicules à l'arrêt ou en contre-sens sont détectés en utilisant les vitesses des objets qui sont comparées aux orientations globales des sous-voies. Si un véhicule est détecté comme étant à l'arrêt, la mise à jour de l'arrière plan est suspendue afin d'éviter l'incrustation de l'objet dans l'arrière-plan. Une classification grossière utilisant les tailles des objets (Figure 10) est faite pour chaque sous-voie en calculant quelques statistiques sur les tailles des objets qui ont été suivis. Ainsi chacune des voies construit son propre histogramme des tailles des objets. L'approche proposée est appliquée sur deux séquences de test dans lesquelles la vérité terrain a été obtenu manuellement. Le comptage de véhicules ainsi que leur classification sont effectués pour chaque voie. Les objets détectés et non-détectés sont comptés (un objet est considéré comme détecté s'il est correctement suivi durant la séquence, c.a.d durant la période de temps correspondant au déplacement d'un objet sur la moitié de la longueur de la voie), puis comparés à la vérité terrain, et on vérifie si les véhicules ont été détectés dans la bonne voie (Table I).

L'utilisation d'un découpage à l'aide des bordures des voies a permis d'améliorer la précision sur la détection (de 92% à 96% pour la première séquence et de 90 à 93% pour la seconde séquence), sur la localisation (de 90% à 93% pour la première séquence et de 93 à 96% pour la seconde séquence) et sur la classification (de 89% à 91% pour la première séquence et de 93 à 94% pour la seconde séquence).

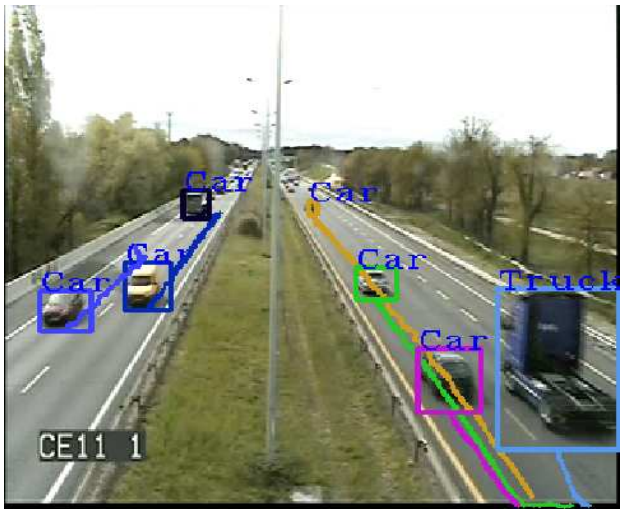


Figure 9 – *Résultat de la détection des objets, leur suivi (trajectoires) et de leur classification en fonction des histogrammes de tailles de chaque sous-voie.*

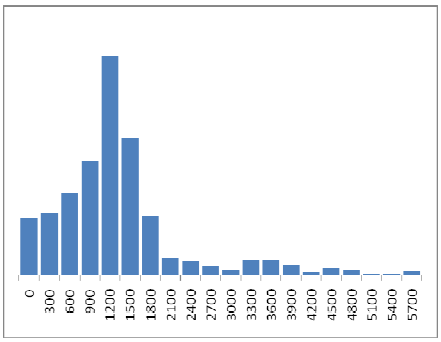


Figure 10 – *Exemple d'histogramme des tailles des objets détectés pour une sous-région de la scène. Chacune des sous-régions construit son propre histogramme utile à la classification.*

Séquence 1 (20 minutes)	Vérité Terrain	Précision Détection	Précision sur la position	Précision Classification
Sans découpage	224	92%	90%	89%
Approche proposée	224	96%	93%	91%

Séquence 2 (30 secondes)	Vérité Terrain	Précision Détection	Précision sur la position	Précision Classification
Sans découpage	30	90%	93%	93%
Approche proposée	30	93%	96%	94%

Table I – *Résultat du comptage de véhicules pour les deux séquences de test.*

4 Conclusion

Nous avons présenté un système d'analyse de trafic capable de compter et détecter les véhicules à l'arrêt ou en contre-sens. En définissant des pixels ambigus pour les objets détectés, le système utilise la modélisation de la scène pour améliorer la détection et le suivi des objets pendant le processus. Les résultats obtenus sont prometteurs et montrent la robustesse du système proposé à suivre plusieurs objets simultanément et à corriger les erreurs de segmentation. La simplicité de l'approche permet une implémentation facile et atteint un traitement en temps réel.

Références

[1] Z. Zhu and G. Xu, VISATRAM: A real-time vision system for automatic traffic monitoring. *Image Vis. Comput.*, vol. 18, no. 10, pp. 781-794, 2000.

[2] M. Yu and Y. D. Kim, Vision based vehicle detection and traffic parameter extraction. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, vol. E84A, no. 6, pp. 1461–1470, 2001.

[3] C. Maduro, K. Batista and J.Batista, Estimating Vehicle Velocity Using Image Profiles on Rectified Images. *IbPRIA*, pp. 64-71, 2009.

[4] C. Stauffer and W. Grimson, Adaptive background mixture models for real-time tracking. *Computer Vision and Pattern Recognition*, vol. 2, pp. 246–252, 1999.

[5] T. Bouwmans, F. El Baf and B. Vachon, Background Modeling using Mixture of Gaussians for Foreground Detection - A Survey. *Recent Patents on Computer Science* 1, 3, pp. 219-237, 2008.

[6] T. Horprasert, D. Harwood, and L. S. Davis, A statistical approach for real-time robust background subtraction and shadow detection. in *ICCV Frame-Rate WS*, pp. 1–19, 1999.

[7] S. Suzuki and K. Abe, Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics and Image Processing*, 30: pp. 32-46, 1985.

[8] C. Tomasi and T. Kanade, Detection and tracking of features points. *Carnegie Mellon University Technical Report* CMU-CS-91-132, April 1991.

[9] J. Shi and C. Tomasi, Good Features To Track. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR94)*, Seattle, June 1994.

[10] Y. Wang, E. Teoh, and D. Shen, Lane detection and tracking using b-snake. *Image and Vision Computing*, 22(4):269–280, 2004.

[11] J. Illingworth, and J. Kittler, A survey of the Hough transform. *Comput. Vision Graph. Image Process.* 44, 1, 87-116, Aug. 1988.

[12] A. Matessi and L. Lombardi, Vanishing Point Detection in the Hough Transform Space. *Proceedings of the 5th international Euro-Par Conference on Parallel Processing*, vol. 1685, pp. 987-994, 1999.

[13] C. Käs and H. Nicolas, Rough compressed domain camera pose estimation through object motion, *ICIP'09*, Cairo, Egypt, 2009.

Une méthode de compression d'images multi/hyperspectrales basée sur les ondelettes 3D anisotropes et son évaluation

J. Delcourt A. Mansouri T. Sliwa Y. Voisin

Le2i (Laboratoire Électronique, Informatique et Image) UMR-CNRS 5158

BP 16, Route des Plaines de l'Yonne
89010 AUXERRE Cedex – FRANCE

{jonathan.delcourt, alamin.mansouri, tadeusz.sliwa, yvon.voisin}@u-bourgogne.fr

Résumé

Dans cet article, nous explorons les stratégies les plus appropriées à la compression d'images multi/hyperspectrales. Pour ce faire, nous comparons la stratégie classique Multi-2D (ondelettes 2D + SPIHT 2D) et la stratégies Hybride (ondelettes 3D + SPIHT 2D) à une stratégie, que nous nommerons Full 3D (F3D), pour laquelle nous proposons une implémentation basée sur une décomposition en ondelettes 3D anisotropes suivie par un coder SPIHT 3D. Toutes ces stratégies sont combinées avec une étape de décorrélation par ACP pondérée afin d'optimiser leurs performances. La comparaison de la stratégie proposée avec les deux autres est réalisée suivant la variation du bitrate ainsi que les dimensions spatiales et spectrales des images. Pour une meilleure évaluation, nous proposons un cadre de comparaison plus large que le critère d'évaluation classique qu'est le PSNR, incluant neuf métriques divisées en quatre familles. Les résultats obtenus par la méthode proposée sont satisfaisants et la comparaison montre les points forts et les points faibles de chaque stratégie.

Mots clefs

Compression d'images multi/hyperspectrales, Ondelettes 3D anisotropes, SPIHT 3D, Cadre d'évaluation.

1 Introduction

Les images multi/hyperspectrales offrent plusieurs avantages par rapport à l'imagerie RVB conventionnelle et ont donc suscité un intérêt croissant au cours des dernières années et sont de plus en plus utilisées en géoscience, télédétection, contrôle qualité industriel, météorologie, et mesures de couleur exacte, etc. Les résolutions spatiales et spectrale augmentent suivant que de meilleurs capteurs hyperspectraux sont développés. Par ailleurs les limitations de vitesse de transmission et de capacité de stockage requièrent le développement de méthodes de compressions adaptées aux images multispectrales.

Généralement, une image multispectrale est représentée sous la forme d'un cube 3D comprenant une dimension spectrale et deux dimensions spatiales. Le fait qu'une image multispectrale soit constituée d'une série de bandes spectrales étroites et contiguës de la même scène produit une séquence d'images fortement corrélée. Cette particularité différencie les images multispectrales des images volumétriques qui possèdent trois dimensions spatiales isotropiques, ainsi que des vidéos qui ont une dimension temporelle et deux dimensions spatiales. Ainsi, les méthodes de compression conventionnelles ne sont pas optimales pour la compression d'image multispectrale.C'est pour cela que les algorithmes de compressions ont besoin d'être adaptés pour ce type d'image et, souvent, requièrent une étape de décorrélation spectrale.

L'une des meilleurs méthodes de compression d'images monochromes est le JPEG 2000 [1]. Son extension aux images multi/hyperspectrales est possible suivant différentes stratégies. Ces stratégies dépendent de la manière dont on considère le cube multi/hyperspectral après la phase de décorrélation (figure 1) :

- chaque bande spectrale de l'image multi/hyperspectrale est considérée séparément (ondelettes 2D + SPIHT 2D) : la stratégie Multi-2D (M2D).
- l'ensemble du cube est considéré comme paramètre d'entrée pour deux principales implémentation : la stratégie Hybride (ondelettes 3D + SPIHT 2D) et la stratégie Full 3D (F3D). Pour cette dernière nous proposons une décomposition en ondelettes 3D anisotrope (ondelettes 3D + SPIHT 3D).

Nous comparons ces trois différentes stratégies de compression en utilisant la même transformée en ondelettes (de type lifting scheme) que celui du standard JPEG 2000. De plus, pour une comparaison plus objective, nous proposons un cadre d'évaluation, composé de huit métriques en plus du classique PSNR. Ces métriques évaluent la qualité de reconstruction en termes de signal, réflèctance spectrale, aspects perceptifs et suivant une métrique basée sur la classification.

Dans la section suivante nous ferons d'abord un bref rappel de la manière dont nous utilisons l'algorithme de l'ACP

1. http ://www.jpeg.org

dans les trois stratégies de compression avant de les décrire. La troisième section présente le cadre d'évaluation en classant les différents critères en quatre catégories et donnant la formule explicite de chaque métrique. Nous expliquons les expérimentations ainsi que leurs résultats avant de les discuter dans la cinquième section. Les conclusions sont présentées dans la dernière section.

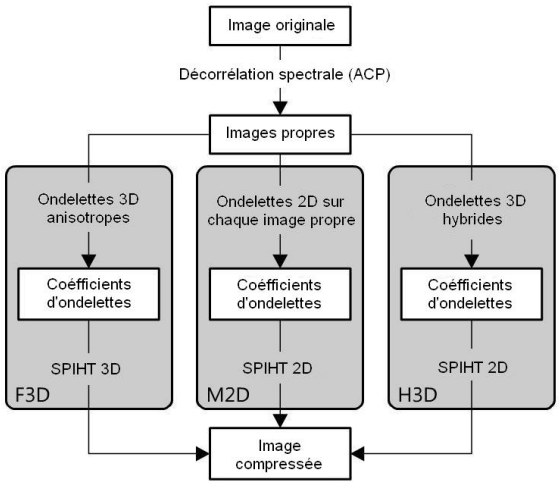


Figure 1 – *Schéma des trois stratégies de compression étudiés.*

2 Méthodes de compression

Comme indiqué précédemment, les images multispectrales ont une forte corrélation inter bandes. Pour obtenir le meilleur taux de compression, il est nécessaire de prendre en compte cette corrélation.

2.1 Décorrelation par ACP

Afin d'optimiser la compression des images multispectrales, une étape de décorrélation est souvent utilisée. Il a été montré que l'utilisation de l'ACP (KLT) est l'un des moyens les plus efficaces pour la décorrélation spectrale [1]. Dans [2, 3] l'utilisation de l'ACP comme décorrélateur pour les images multispectrales est considérée comme efficace. De notre côté, nous utilisons l'ACP en l'appliquant sur l'image multispectrale originale suivant la dimension spectrale. Comme résultat, nous obtenons une nouvelle image multi-bandes dans le domaine de la transformée dans lequel la corrélation des composants de la transformation est réduite. Les bandes dans le domaine de la transformée sont triées par valeurs décroissantes en fonction de la variance (ou en fonction des valeurs propres).

Après la décorrélation par ACP nous pouvons compresser l'image résultante totale en appliquant les trois différentes stratégies.

2.2 Stratégie Full 3D (F3D)

La stratégie F3D consiste à considérer l'ensemble du cube multispectral comme entrée de la transformée en ondelettes 3D. Dans notre cas, l'entrée est le résultat de l'ACP. Ensuite, une extension 3D du codeur SPIHT [4] est appliquée. Le SPIHT 3D de Kim *et al.* [5] est approprié au format en bloc de la décomposition en ondelettes 3D (figure 1).
De nombreux articles de la littérature ont étudié la compression d'images multi/hyperspectrales par transformée en ondelette 3D, mais ils n'utilisent que des transformées en ondelettes 3D isotropes (même type d'ondelettes suivant toutes les directions de l'image) [5, 6, 7, 8, 9, 10, 2, 11, 12, 13]. Cependant, la taille de la dimension spectrale est généralement plus faible que celles des dimensions spatiales, il est approprié d'utiliser un autre type de filtre d'ondelettes dans cette dimension. À cet effet, nous proposons d'utiliser une transformée en ondelettes 3D anisotropes réalisée avec des filtres CDF 9/7 suivant les dimensions spatiales et un filtre de type Haar suivant la dimension spectrale. Le filtre spectral a été choisi suivant les résultats obtenus par Mansouri *et al.* dans [14]. Ce résultat rejoint les conclusions de Kaarna et Parkkinen dans [15] où ils recommandent l'utilisation d'ondelettes à support court comme choix pour des ondelettes spectrales.
La transformée en ondelette utilisée est une extension de la classique transformée en ondelette 2D. Elle produit une transformée en ondelettes multidimensionnelle en appliquant un niveau de décomposition suivant chaque dimension. Cette étape étant réitérée sur le cube d'approximation obtenu jusqu'a obtenir le niveau de décomposition désiré (figure 2).

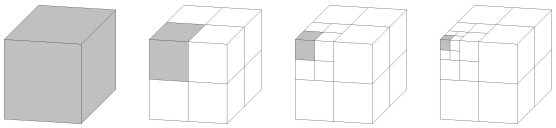


Figure 2 – *Représentation graphique de la décomposition en ondelette 3D avec trois niveaux de décomposition.*

2.3 Stratégie Multi-2D (M2D)

Cette stratégie consiste à appliquer sur chaque image propre de l'ACP la même décomposition en ondelette puis un codeur SPIHT 2D [2, 3, 6]. En raison de l'ACP, les images propres sont triées par énergies décroissantes. Afin de tenir compte de ce fait, il est préférable d'effectuer une pondération avant de coder chaque bande avec le SPIHT 2D. Comme pondération, nous définissons l'énergie E de chaque image propre suivant la formule :

$$E = \frac{\sqrt{\sum_{x,y} I_\lambda(x,y)^2}}{XY} \quad (1)$$

où I_λ est l'image propre à la longueur d'onde λ, X et Y sont ses dimensions, et x et y sont la position d'un pixel

dans l'image propre.

2.4 Stratégie Hybride (H3D)

La stratégie H3D consiste à appliquer une ondelettes 3D hybride rectangle/carré (figure 3) sur le résultats de l'ACP comme utilisé dans [12]. La décomposition en ondelettes est composée par des filtres CDF 9/7 suivant les directions spatiales et d'un filtre de type Haar suivant la direction spectrale. Le fait que cette transformée en ondelettes comporte deux étapes différenciées (transformée spatial suivie par la transformée spectrale) permet de considérer son résultats comme une association de plans 2D. Pour cette raison nous pouvons appliquer un codage par SPIHT 2D sur chaque bande résultante pour achever la compression, comme dans la stratégie de compressions M2D. Pour prendre en compte la différence d'énergie entre chaque bande, nous pondérons chaque bande par son énergie E comme indiqué dans l'équation (1).

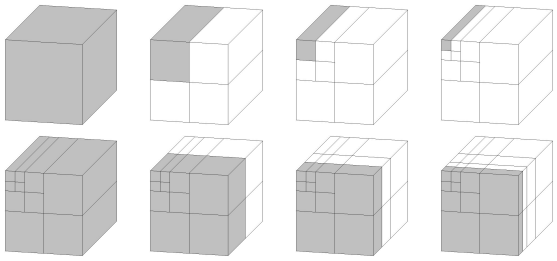


Figure 3 – *Représentation graphique de la décomposition en ondelette 3D hybride rectangulaire/carré avec trois niveaux de décomposition. Décomposition spatiale (haut) suivit par la decomposition spectrale (bas).*

3 Cadre d'évaluation de la compression

Quand la compression avec perte est utilisée, il est nécessaire d'évaluer et de décrire la nature et l'importance des dégradations dans l'image reconstruite (décompressée) [16]. Habituellement, dans le cadre d'images 2D classiques, les performances des méthodes sont évaluées à l'aide d'un critère qui privilégie un seul aspect (reflétant habituellement la perception de la vision humaine). Ce n'est pas le cas pour les images multi/hyperspectrales, qui sont utilisées dans des algorithmes de classification et de détection. À cet effet les métriques doivent correspondre aux applications. C'est pour cela qu'au lieu d'évaluer les performances de compression suivant une seule métrique ou un seul type de métrique, nous proposons l'utilisation de neuf métriques connues suivant quatre catégories. Nous les utilisons comme cadre d'évaluation de la compression.

Les métriques que nous proposons peuvent être divisées en quatre différentes familles : critères issues d'une extension isotrope de critères de traitement du signal (PSNR, RRMSE, MAE et MAD), métriques orientées multispectral (F_λ, MSA et GFC),un critère statistique évolué prenant en compte certains aspects perceptifs (UIQI) et une métrique orientée classification (K-means). Dans [17], Christophe *et al.* démontrent que l'utilisation d'un jeu de plusieurs métriques est plus pertinent que d'en utiliser un seul. Ainsi nous avions proposé dans [18] l'utilisation de quatre familles de métriques en plus du PSNR. L'utilisation de ce dernier résulte du fait qu'elle est la métrique la plus employée, ce qui facilite la comparaison avec les résultats d'autres méthodes.

4 Expériences et résultats

4.1 Données et expériences

Nous avons effectué nos expériences sur les images largement utilisées venant du capteur AVIRIS [2] (Cuprite, SanDiego, JasperRidge et MoffettField). Ces images représentent des paysages très différents, Cuprite et JasperRidge représentent des large zones uniformes alors que SanDiego représente un aéroport et MoffettField représente une zone urbaine contenant beaucoup de hautes fréquences.

La première expérience effectuée vise à comparer les performances de la stratégie F3D avec les stratégies M2D et H3D suivant différents bitrate et en faisant varier les dimensions spatiales de l'image. L'expérience est menée sur 32 bandes des image Cuprite, SanDiego, JasperRidge et MoffettField avec des dimensions spatiales de $64 * 64$ et $128 * 128$ pixels. Toutes les images sont codées en entier de 16 bits.

La deuxième expérience vise à évaluer la performance de la stratégie F3D avec les stratégies M2D et H3D suivant différents bitrate quand le nombre de bandes spectrales varie. Pour ce faire nous présentons les résultats relatifs à l'image SanDiego suivant plusieurs tailles spatiales en faisant varier chaque fois le nombre de bande spectrale de l'image (64, 128 and 192).

4.2 Résultats

Représenter les résultats des expériences suivant neuf paramètres est difficile. Un bon moyen d'y parvenir est d'utiliser un diagramme en étoile (radar), comme utilisé dans [19], qui donne dans ce cas une vision plus compacte et intuitive que les représentations x-y classique. Les neufs axes du diagramme correspondent aux neuf métriques. Les différents diagrammes en étoile ont tous la même échelle pour faciliter interprétation graphique ainsi que la comparaison. Les axes pour les RRMSE, MAD, MAE, MSA et K-means sont inversés, l'extrémité de l'axe correspond au minimum de dégradation et l'origine correspond au maximum de dégradation. Cette représentation permet une bonne lecture des résultats mais ne permet d'afficher, sur le même diagramme, qu'une seule valeur de bitrate. C'est pour cette raison que dans la figure 6 nous ne représentons les résultats que pour un birate de 1 bpp.

2. http ://aviris.jpl.nasa.gov

Les résultats de la première expérience (concernant les variations des dimensions spatiales) en terme de PSNR sont représentés dans la figure 4. Les résultats de compression pour les quatres différentes images montrent la même tendance. Cette tendance est caractérisée par le fait que la stratégie F3D surpasse les stratégies M2D et H3D pour les grandes valeurs de bitrate, alors que pour de faibles valeurs de bitrate c'est la stratégie M2D qui donne les meilleurs résultats. La stratégies H3D ne donne jamais les meilleurs résultats.
Pour la seconde expérience sur l'image SanDiego, les figures 5 et 6 montrent que la stratégie F3D surpasse toujours les deux autres stratégies de compression pour les fortes valeurs de bitrate. Quand le nombre de bandes spectrales augmente, la stratégies F3D surpasse les deux autres stratégies pour des valeurs de bitrate plus faibles. Les diagrammes étoiles (figures 6) montrent que toutes les métriques n'indiquent pas la même tendance. Pour les stratégies F3D et M2D toutes les métriques ont des résultats similaires excepté en terme de UIQI et pour la stratégies H3D, les résultats en termes de PSNR, GFC, MAD, MAE et UIQI sont relativement similaires, alors que les résultats en termes de RRMSE, fidélité spectrale F_λ et MSA indiquent une tendance inversée.

5 Discussion

Les deux expériences réalisées permettent de comparer la stratégie F3D aux stratégies M2D et H3D suivant les variations de dimensions spatiales et spectrale. Une tendance générale peut être observée : pour les fortes valeurs de bitrate la stratégie F3D donne les meilleurs résultats et pour les faibles valeurs de bitrate c'est la stratégies M2D qui donne les meilleurs résultats. Les résultats de la stratégie H3D sont compris entre les résultats des deux autres stratégies. Cette tendance peut être expliquée par deux points majeurs :

– Pour de faibles valeurs de bitrate la stratégies F3D donne de faibles résultats car le SPIHT 3D utilisé pour cette stratégie utilise des listes (listes de pixels significatifs et non-significatifs et liste des ensembles significatifs) qui grandissent très rapidement comparées aux listes du SPIHT 2D (chaque pixel a huit premiers descendants pour la version 3D contre seulement quatre pour la version 2D). Et pour les valeurs élevées de bitrate moins de coefficients sont ajoutés aux listes pour le SPIHT 3D que pour le SPIHT 2D. Cela peut expliquer le fait que la stratégies M2D donne de meilleurs résultats que la stratégie F3D seulement pour les faibles valeurs de bitrate.
– La stratégies H3D donne de mauvais résultats car c'est la combinaison d'éléments 2D et 3D. Ainsi l'utilisation du SPIHT 2D après une décomposition en ondelettes 3D ne semble pas être optimale.

6 Conclusion

Dans cet article, nous avons proposé une nouvelle stratégie de compression d'image multi/hyperspectrale basée sur une décomposition en ondelettes 3D anisotrope (F3D) et nous l'avons comparée à deux autres stratégies. : M2D et H3D. Toutes les stratégies sont combinées avec une étape de décorrélation spectrale par ACP. La comparaison de ces différentes stratégies est effectuée dans un cadre d'évaluation comprenant neuf métriques appartenant à quatre familles différentes : extension isotrope de critères de traitement du signal, métriques orientées multispectral, un critère statistique perceptif évolué et une métrique orientée classification basée sur les K-means. La comparaison des stratégies de compression suivant le cadre d'évaluation montre la même tendance suivant la majorité des métriques : la stratégie F3D est meilleure que les stratégies M2D et H3D pour des valeurs de bitrate élevées. Les résultats de la stratégies F3D sont meilleurs pour des images de grande dimensions spatiales et pour une grand nombre de bandes spectrales.

Références

[1] P. Ready et P. Wintz. Information extraction, SNR improvement, and data compression in multispectral imagery. *Communications, IEEE Transactions on [legacy, pre-1988]*, 21(10) :1123–1131, 1973.

[2] J. Mielikäinen et A. Kaarna. Improved back end for integer PCA and wavelet transforms for lossless compression of multispectral images. *Proceedings of 15th International Conference on Pattern Recognition*, 2 :257–260.

[3] Q. Du et J.E. Fowler. Hyperspectral image compression using JPEG2000 and principal component analysis. *IEEE Geoscience and Remote Sensing Letters*, 4(2) :201, 2007.

[4] A. Said et W.A. Pearlman. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Transactions on circuits and systems for video technology*, 6(3) :243–250, 1996.

[5] B.-J. Kim, Z. Xiong, et W.A. Pearlman. Low bitrate scalable video coding with 3-D set partitioning in hierarchical trees(3-D SPIHT). *IEEE Transactions on Circuits and Systems for Video Technology*, 10(8) :1374–1387, 2000.

[6] A. Kaarna, P. Toivanen, et P. Ker
"anen. Compression and classification methods for hyperspectral images. *Pattern Recognition and Image Analysis*, 16(3) :413–424, 2006.

[7] A. Kaarna, P. Zemcik, H. Kaelviainen, et J. Parkkinen. Multispectral image compression. Dans *INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION*, volume 14, pages 1264–1267. IEEE COMPUTER SOCIETY PRESS, 1998.

[8] A. Kaarna et J. Parkkinen. Wavelet filter selection in multispectral image compression. Dans *INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION*, volume 15, pages 242–245, 2000.

[9] A. Kaarna et J. Parkkinen. Comparison of compression methods for multispectral images. Dans *Proc. NORSIG–Nordic Signal Process. Symp*, pages 251–254.

[10] A. Kaarna. Integer pca and wavelet transforms for multipectral image compression. Dans *International Geoscience and Remote Sensing Symposium, IEEE IGARSS'2001*, volume 4, pages 1853–1855, 2001.

[11] S. Lim, K.H. Sohn, et C. Lee. Compression for hyperspectral images using three dimensional wavelet transform. Dans *International Geoscience and Remote Sensing Symposium, IEEE IGARSS'2001*, volume 1, pages 109–111, 2001.

[12] B. Penna, T. Tillo, E. Magli, et G. Olmo. Progressive 3-D coding of hyperspectral images based on JPEG 2000. *IEEE Geoscience and Remote Sensing Letters*, 3(1) :125–129, 2006.

[13] X. Tang, S. Cho, et W.A. Pearlman. 3D Set Partitioning Coding Methods in Hyperspectral Image Compression. Dans *Proceedings of IEEE International Conference on Image Processing (ICIP'03*, volume 2, pages 239–242.

[14] A. Mansouri, T. Sliwa, J.Y. Hardeberg, et Y. Voisin. Representation and estimation of spectral reflectances using projection on PCA and wavelet bases. *Color Research and Application*, 33(6) :485–493, 2008.

[15] A. Kaarna et J. Parkkinen. Wavelet compression of multispectral images. *Proceedings of the IASTED International Conference on Computer Graphics and Imaging*, pages 142–145, 1998.

[16] A.M. Eskicioglu et P.S. Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12) :2959–2965, 1995.

[17] E. Christophe, D. Léger, et C. Mailhes. Quality criteria benchmark for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 43(9) :2103, 2005.

[18] J. Delcourt, A. Mansouri, T. Sliwa, et Y. Voisin. A comparative study and an evaluation framework of multi/hyperspectral image compression. Dans *5th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS 2009)*, 2009.

[19] E. Christophe, D. Léger, et C. Mailhes. New quality representation for hyperspectral images. *The International Society for Photogrammetry and Remote Sensing (ISPRS)*, pages 315–320, 2008.

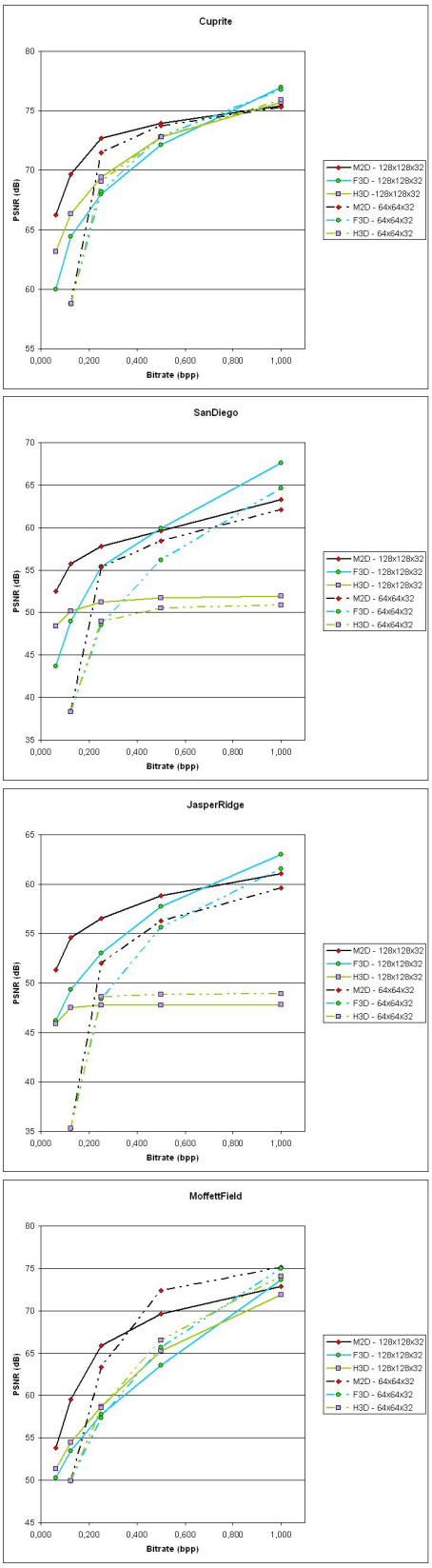
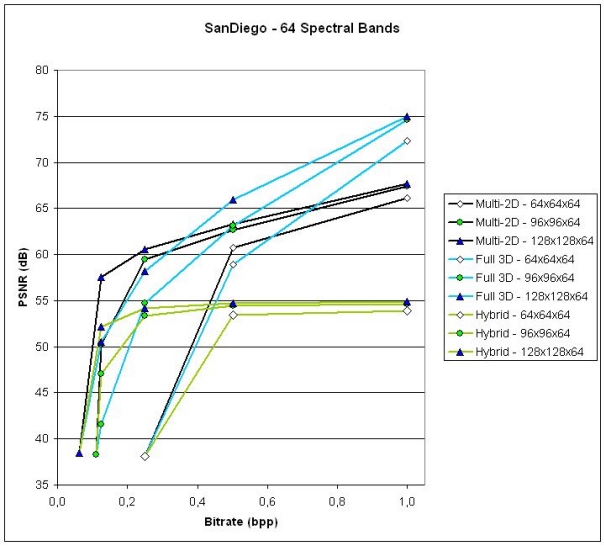
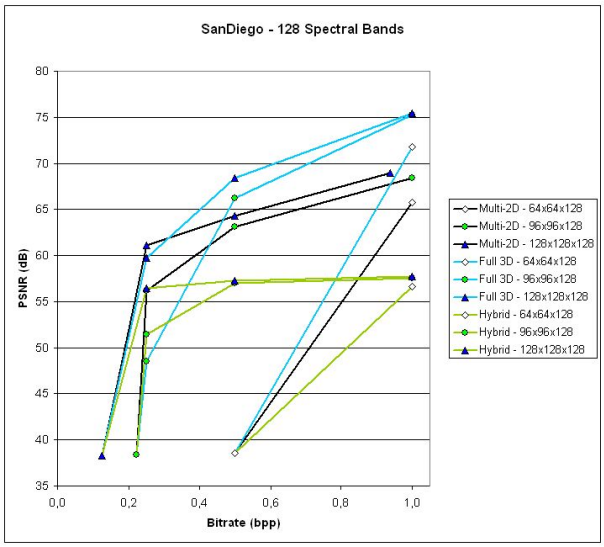


Figure 4 – *Résultats de compression en termes de PSNR pour les images Cuprite, SanDiego, JasperRidge et MoffettField.*





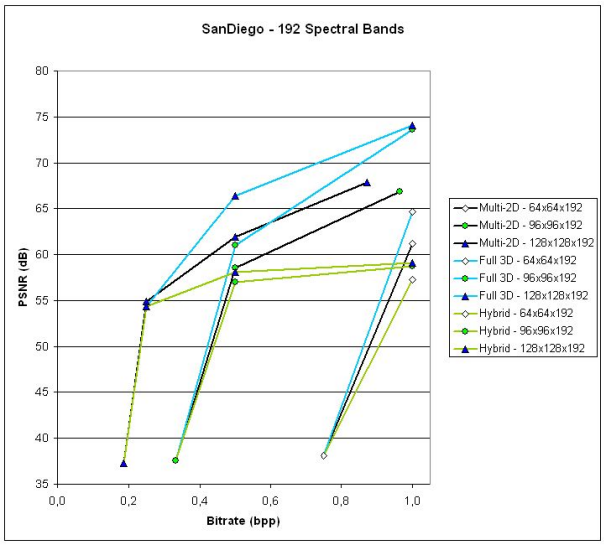


Figure 5 – *Résultats de compression en termes de PSNR pour 64, 128 et 192 bandes spectrales de l'image San-Diego.*

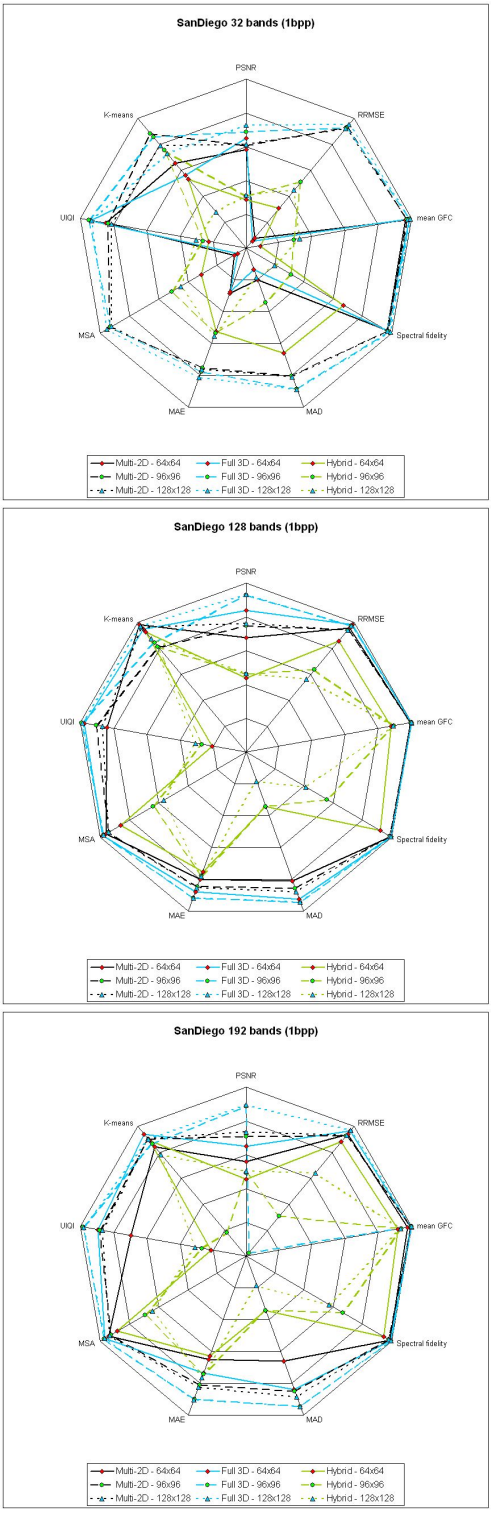


Figure 6 – *Résultats de compression pour 64, 128 et 192 bandes spectrales de l'image SanDiego poue un bitrate de 1 bpp.*

Reconnaissance de la sémantique émotionnelle portée par les images basée sur la théorie de l'évidence

N. Liu[1], E. Dellandrea[1], B. Tellez[2], L. Chen[1].

[1]Université de Lyon, CNRS, École Centrale, LIRIS, UMR5205, F-69134, France

[2]Université de Lyon, CNRS, Université Lyon1, LIRIS, UMR5205, F-69622, France

{ningning.liu, emmanuel.dellandrea,liming.chen}@ec-lyon.fr,
bruno.tellez@liris.cnrs.fr

Résumé

La reconnaissance de la sémantique émotionnelle d'une image prend une place de plus en plus importante dans la communauté de recherche. Elle offre en effet des perspectives nouvelles et motivantes pour retrouver et classer des images selon la charge émotionnelle qu'elles peuvent porter. Cependant, comme tout sujet émergeant, les contributions sur ce thème demeurent relativement rares et beaucoup de pistes doivent être étudiées. Dans cet article, nous nous proposons d'évaluer l'efficacité de différents types de descripteurs et de classificateurs pour la reconnaissance d'émotions visuelles dans les images. Dans un second temps, nous proposerons l'utilisation de la théorie de l'évidence de Dempster-Shafer qui permet la manipulation et la fusion de connaissances ambigües et incertaines telles que celles rencontrées dans le traitement des émotions. Les expérimentations menées sur la base d'images IAPS mettent en évidence l'efficacité de cette approche.

Mots-clefs

Emotion, image, classification, théorie de l'évidence, descripteurs visuels.

1 Introduction

Un des buts de l'informatique, et particulièrement de l'intelligence artificielle est d'élaborer des ordinateurs intelligents qui ont la capacité d'interagir avec des êtres humains de façon naturelle. Dès lors, une des questions essentielles est de permettre aux ordinateurs de reconnaitre, de comprendre et d'exprimer des émotions [1]. Plusieurs travaux ont été faits depuis plusieurs années sur ces aspects en informatique mais également en robotique. Quand il s'agit de reconnaitre des émotions (voir [2] pour un tour d'horizon très complet), les recherches portent principalement sur la reconnaissance d'affects dans des données audio (parole ou musique) et sur la reconnaissance visuelle d'expressions faciales. Très peu de contributions traitent de la reconnaissance de la sémantique émotionnelle portée globalement par les images que ce soit par ses couleurs, sa composition ou tout autre élément qui peut provoquer une émotion. Face à ce sujet de recherche émergeant, un grand nombre de questions doivent être abordées concernant principalement les trois problèmes suivants : la représentation des émotions, l'extraction de caractéristiques visuelles nécessaire à la reconnaissance des émotions et les modèles de classification pour traiter les différentes propriétés des émotions [3, 4, 5]. En effet, comme dans tous les autres problèmes de vision par ordinateur, la principale difficulté consiste à franchir le fossé sémantique qui existe entre les descripteurs bas-niveaux extraits des images et les concepts sémantiques de haut-niveau qui sont dans notre cas les émotions.

Dans cet article, nous nous proposons d'étudier l'efficacité de différents types de descripteurs visuels ainsi que les classificateurs nécessaires à la reconnaissance d'émotions dans les images. De plus, nous proposerons d'utiliser la théorie

de l'évidence de Dempster-Shafer [15,16], qui permet la manipulation de connaissances ambigües et incertaines comme celles relatives aux émotions.

Le reste de l'article est organisé de la façon suivante. Les différents modèles pour représenter les émotions sont décrits dans la partie 2. Les propriétés des images utilisées pour caractériser les émotions sont présentées dans la partie 3. Les classificateurs testés et choisis pour reconnaitre les émotions sont détaillés dans la partie 4. Les expérimentations sont présentées dans la partie 5. Enfin, nous ferons une synthèse de notre étude dans la partie 6.

2 Représentation des émotions

Plusieurs modèles ont été étudiés dans la littérature pour représenter les émotions [2]. Les deux principales approches sont le modèle discret et le modèle dimensionnel. Le premier modèle consiste à choisir des noms ou des adjectifs pour décrire les émotions, tels que le bonheur, la tristesse, la peur, la colère, le dégoût et la surprise. Le second modèle décrit les émotions selon une ou plusieurs dimensions où chacune représente une caractérisation de l'émotion, les plus utilisées étant l'appréciation, l'activité ou le contrôle. Ce deuxième modèle permet de représenter un plus large éventail d'émotions que le premier.

Le choix de la représentation émotionnelle est généralement guidé par l'application. Ainsi, les deux approches sont utiles et peuvent même être combinées, car elles peuvent apporter des informations complémentaires. Dans cet article, nous proposons une représentation hybride comme l'illustre la figure 1. Chaque image est ainsi représentée comme un point de l'espace constitué des deux dimensions que sont l'appréciation (variant de très déplaisante à très plaisante) et l'activité (variant de très calme à très dynamique). Cet espace est divisé en quatre quadrants permettant d'obtenir quatre types d'émotions distinctes afin de caractériser la charge émotionnelle de chaque image.

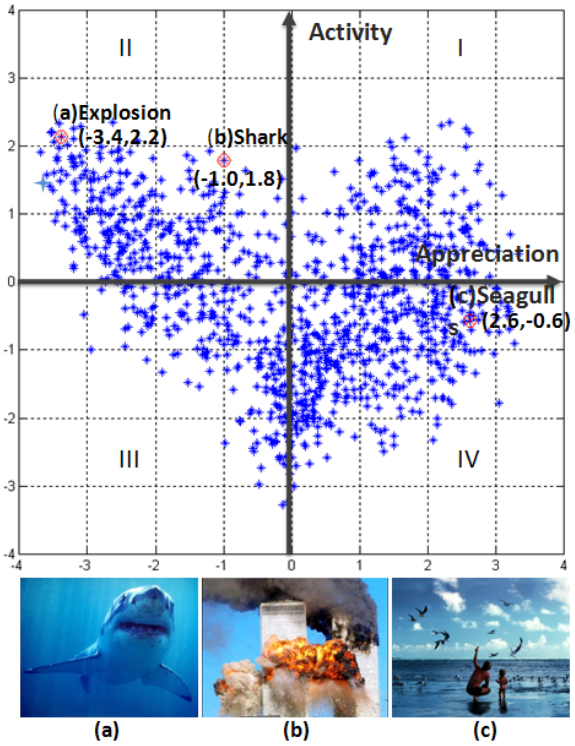


Figure 1 –*Représentation des images de la base IAPS selon des critères d'activité et d'appréciation [8].*

A terme, il serait même envisageable de réaliser un découpage plus fin de cet espace et d'associer à chaque région, un label spécifique représentant une émotion.

3 Descripteurs d'images pour la reconnaissance des émotions

L'extraction des caractéristiques propres d'une image est une question clé pour la reconnaissance de concepts dans des images, et en particulier, les émotions. Ces caractéristiques doivent porter les informations nécessaires pour permettre la reconnaissance des différents concepts. Comme la reconnaissance des émotions dans les images est un domaine de recherche émergent, très peu de travaux ont été réalisés pour identifier les caractéristiques de l'image qui sont les plus efficaces dans ce contexte.

3.1 Descripteurs d'images traditionnels

La plupart des travaux traitant de la reconnaissance des émotions utilisent les descripteurs qui le sont généralement également pour d'autres problèmes de vision par ordinateur. Les trois principales catégories de descripteurs d'images sont basées sur la couleur, la texture et la forme. En ce qui concerne la couleur, des études ont montré que l'espace HSV (Hue, Saturation, Value) est un espace de couleur qui est mieux adapté à la perception réelle des couleurs par l'homme que d'autres espaces tels que l'espace RGB traditionnel. Ainsi, sur la base de cet espace de couleur, plusieurs façons de décrire le contenu couleur des images peuvent être considérées tels que les moments de couleurs, les corrélogrammes et histogrammes de couleur ainsi que les histogrammes relatifs à la température de la couleur [10, 11].

En ce qui concerne la texture, la principale caractéristique demeure les matrices de cooccurrences [11,12]. Toutefois, les descripteurs de Tamura [12] peuvent également représenter une alternative intéressante. En effet, des descripteurs tels que la granularité, le contraste ou la directionnalité se sont avérés fortement corrélés avec la perception visuelle de l'homme.

Enfin, la description des formes peut être envisagée grâce à l'extraction des contours permettant l'obtention de l'histogramme d'orientation des lignes [6,12] ou encore les descripteurs de Haar [10,13].

3.2 Descripteurs sémantiques de l'image pour la reconnaissance des émotions

Certaines tentatives ont été faites pour identifier des descripteurs de plus haut-niveau liés aux émotions. En effet, les études sur les peintures ont mis en évidence la portée sémantique des couleurs et des lignes qui y apparaissent, comme cela est rappelé dans les travaux de [6] où sont proposés des descripteurs d'images plus corrélés aux émotions grâce à l'exploitation de ces informations. Ainsi, en utilisant la théorie des couleurs d'Itten, une signification émotionnelle des couleurs peut être dégagée. Tout d'abord, comme mentionné plus haut, les couleurs sont décrites en terme de teinte, de luminance et de saturation grâce à l'espace de couleur HSV, afin de se rapprocher de la perception humaine des couleurs. Ces couleurs sont ensuite projetées sur un cercle chromatique, appelé cercle d'Itten où les couleurs fortement contrastées ont des coordonnées opposées par rapport au centre du cercle. Itten a montré que les combinaisons de couleurs peuvent produire des effets tels qu'une harmonie, une disharmonie, du calme ou de l'excitation. Ainsi, l'harmonie sera détectée sur le cercle d'Itten si les positions des couleurs connectées entre elles constituent un polygone régulier comme montré dans la figure 2. Le descripteur correspondant à cette hypothèse est obtenu en mesurant la distance entre le centre du cercle d'Itten et le centre du polygone reliant les couleurs dominantes de l'image. Ces dernières sont préalablement obtenues par un algorithme basé sur les k-means.

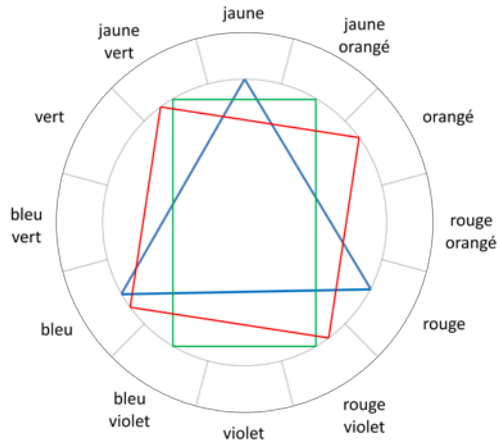


Figure 2. *Cercle d'Itten et exemples d'associations de couleurs harmonieuses (en rouge, vert ou bleu).*

Les lignes portent également une information sémantique importante sur les images. En effet, des lignes obliques suggèrent le dynamisme et l'action tandis que les lignes horizontales ou verticales communiquent plutôt le calme et la détente. Pour exprimer cela en terme de descripteurs d'images, les lignes sont d'abord extraites grâce à une transformée de Hough, puis le rapport entre le nombre de lignes obliques et

le nombre total de lignes dans une image est calculé.

4 Modèles de classification pour la reconnaissance des émotions

La plupart des travaux traitant de la classification des émotions dans les images reposent sur des approches traditionnelles de classification largement utilisées dans d'autres problèmes de vision par ordinateur. Malheureusement, elles ne sont pas toujours appropriées pour traiter de la spécificité des émotions. Parmi ces approches, on peut citer les réseaux de neurones [14], les machines à vecteurs supports (SVM) [9,10,12] ou les modèles par mélange de gaussiennes [10].

4.1 La théorie de l'évidence

Les émotions sont des concepts de haut-niveau sémantique qui sont, par nature, hautement subjectifs et ambigus. Ainsi, afin de s'acquitter efficacement de cette tâche de reconnaissance, il est nécessaire de traiter des informations qui peuvent être incertaines, incomplètes, équivoques et pouvant conduire à des conflits. C'est la raison pour laquelle nous proposons de faire usage de la théorie de l'évidence qui gère naturellement ces difficultés.

4.1.1 *Contexte de la théorie de l'évidence*

La théorie de l'évidence de Dempster-Shafer [15,16] propose un cadre permettant un raisonnement sur des connaissances qui peuvent être incertaines, incomplètes et conduisant à des conflits. Cette théorie s'appuie sur des fonctions de masse qui sont une généralisation des probabilités et des mesures de possibilité.
Soit $\Theta = \{\theta_1, \theta_2, ..., \theta_K\}$ un ensemble fini d'hypothèses possibles. Cet ensemble est nommé cadre de discernement, et l'ensemble puissance est dé signé 2^{Θ}. Les concepts de base de la théorie sont les suivants :
Fonction de masse de croyance élémentaire : la fonction de masse m, associée à une source d'information donné e (un type de descripteur dans notre cas), attribue une valeur comprise dans l'intervalle [0, 1] pour toute partie A de Θ et remplit les conditions suivantes:

$$m(\emptyset) = 0 \; and \sum_{\mathcal{A} \subseteq \Theta} m(\mathcal{A}) = 1 \qquad (1)$$

m (A) représente la confiance, ou croyance, que nous pouvons avoir dans la réalisation d'une hypothèse A.
Les éléments focaux sont des sous-ensembles A tels que m(A)> 0. Si m (Θ) = 1 alors la source est totalement incertaine alors que si m (θ_1) = 1 alors la source est parfaite pour l'hypothèse θ_1.
Règle de combinaison : l'une des propriétés les plus intéressantes de la théorie de la preuve réside dans sa capacité à combiner les fonctions de masse différentes issues de plusieurs sources d'information. Considérons $m_1(.)$ et $m_2(.)$ deux fonctions de masse provenant de deux sources d'information indépendantes S_1 et S_2 respectivement. Dès lors, $m_1(.)$ et $m_2(.)$ peuvent être combinées pour obtenir la masse de la croyance engagée sur $C \subseteq \Theta$, $C \neq \emptyset$; selon la formule de combinaison suivante (Shafer, 1976):

$$m(\mathcal{C}) = \frac{\sum_{\mathcal{B} \cap \mathcal{A} = \mathcal{C}} m_1(\mathcal{B}) \cdot m_2(\mathcal{A})}{1 - \sum_{\mathcal{B} \cap \mathcal{A} = \emptyset} m_1(\mathcal{B}) \cdot m_2(\mathcal{A})} \qquad (2)$$

Une fois que les fonctions de masse des différentes sources d'informations à notre disposition sont combinées en une seule fonction de masse, une décision finale peut être prise en considérant l'hypothèse qui est associée à la valeur la plus élevée.

4.1.2 *Construire l'évidence*

Une des principales difficultés rencontrées lors de l'élaboration d'une méthode de classification basée sur la théorie de l'évidence concerne la manière dont les fonctions de masse sont construites à partir des descripteurs d'images. Dans cet travail, nous avons utilisé l'approche proposée dans [7] qui estime les fonctions de masse à partir de classificateurs en minimisant

l'Erreur Quadratique Moyenne entre les résultats de la classification et les sorties attendues.

5 Expérimentations

Dans nos expérimentations, nous avons utilisé la base de données d'images IAPS qui est une base de référence en psychologie pour l'étude des émotions communiquées par les images [8]. Elle fournit une caractérisation des images selon trois critères en fonction de l'émotion produite : l'appréciation, l'activité et le contrôle. Cette base comporte 1192 images qui peuvent donc être représentées dans un espace dimensionnel des émotions, selon les axes d'appréciation et d'activité. Par commodité, cette représentation des émotions n'est pas utilisée directement, mais est utilisée pour définir 4 classes d'émotions correspondant aux 4 quadrants de la figure 1.
Le corpus IAPS est partitionné aléatoirement en un ensemble d'apprentissage (80% des données, 953 images) et un ensemble de test (20% des données, 239 images). Toutes les expériences sont répétées 10 fois pour obtenir un pourcentage moyen de classification correcte.
Pour évaluer la performance des différents classificateurs pour la reconnaissance des émotions dans les images, nous avons examiné quatre classificateurs représentatifs : machines à vecteurs supports (SVM), réseaux de neurones (Feed-Forward Neural Networks), Adaboost et K-plus proches voisins. Le schéma de classification que nous avons retenu consiste à utiliser deux classificateurs binaires. Le premier est entrainé pour identifier l'activité, et le second sert à identifier l'appréciation. Les résultats sont ensuite combinés pour identifier l'une des 4 classes d'émotion.

Les caractéristiques d'entrée sont générées en utilisant les techniques décrites dans la partie 3 et alignées en un seul vecteur, ce qui correspond à une fusion précoce. Les résultats de classification sont donnés dans le tableau 1. Nous pouvons observer que les classificateurs SVM avec un pourcentage moyen de classification correcte de 62,6% réalisent la meilleure performance parmi les quatre types de classificateurs, même si Adaboost atteint des performances très proches.

	NN (%)	*SVM (%)*	*Adaboost(%)*	*Knn(%)*
I	57.21	61.55	**65.02**	51.33
II	63.42	60.34	62.53	**64.42**
III	58.21	**62.61**	61.31	51.52
IV	61.72	**65.75**	64.30	61.71

Table 2 – *Pourcentages moyens de classification correcte pour 4 classes d'émotion obtenus par les 4 classificateurs.*

Un autre aspect intéressant consiste à comparer la capacité des différents types de descripteurs d'images à porter l'information relative aux émotions. Ainsi, le système de classification basé sur SVM décrit précédemment a été appliqué indépendamment pour chaque type de descripteurs. Les résultats sont donnés dans la figure 3. Cette figure présente également le pourcentage de bonne classification obtenu avec la fusion de tous les descripteurs en s'appuyant sur l'approche fondée sur la théorie de l'évidence présentée à la section 4.2. La première remarque est que la performance entre les différents descripteurs est très similaire, variant de 53,3% pour les corrélogrammes jusqu'à 58,2% pour LBP. Toutefois, parmi les différents descripteurs, il semble que la texture (LBP et la matrice de cooccurrences) soit le type de descripteurs le plus efficace. En outre, les descripteurs de plus haut-niveau (dynamisme et harmonie) même s'ils peuvent paraître moins performants au premier abord ne reposent que sur une seule valeur et donc, leur efficacité est tout à fait remarquable. Enfin, il faut mentionner que l'approche proposée pour la fusion de l'ensemble des descripteurs basée sur la théorie de l'évidence, et dont la matrice de confusion est donnée dans le tableau 2, donne les meilleurs résultats avec un pourcentage moyen de classification correcte de 64,6%. Cette valeur montre la capacité de la théorie de l'évidence à combiner différentes sources d'information et à exploiter leurs complémentarités.

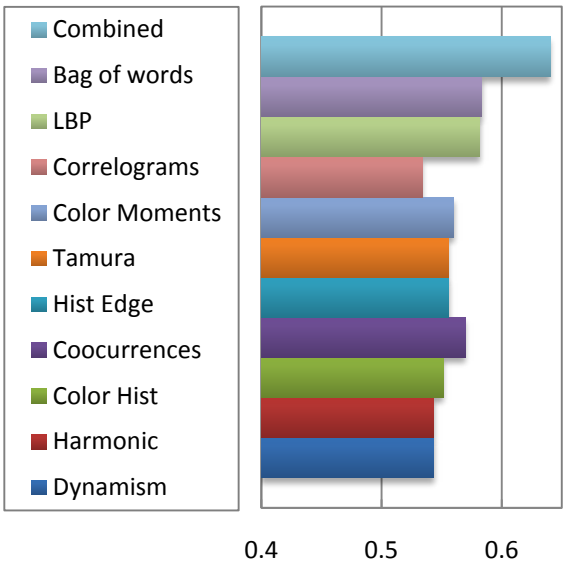


Figure 3 – *Taux de reconnaissance moyen obtenus pour chaque type de descripteurs et par fusion (combined).*

Prédit / Réel	I	II	III	IV
I	**63.32**	12.25	11.23	11.15
II	11.05	**61.42**	12.27	11.82
III	16.21	12.53	**66.19**	10.52
IV	10.42	13.80	10.31	**67.51**
Total	100	100	100	**100**

Table 2 – *Matrice de confusion pour les 4 classes d'émotion en utilisant la théorie de l'évidence.*

6 Conclusion

Dans cet article, nous avons étudié l'efficacité des différents types de caractéristiques et de classificateurs pour la reconnaissance des émotions dans des images. En outre, nous avons proposé une méthode de classification basée sur la théorie de l'évidence, qui présente la capacité de traiter les connaissances ambiguës et incertaines, comme celles qui peuvent caractériser les émotions. Les expériences sur la base de données IAPS ont mis en évidence que, parmi les classificateurs traditionnels, SVM obtient les meilleurs résultats, et que la texture ainsi que les descripteurs de dynamisme et d'harmonie portent des informations importantes liées aux émotions. Enfin, grâce à notre approche basée sur la théorie de l'évidence, nous avons pu atteindre un taux de reconnaissance globale de 64,6% que nous considérons comme encourageant dans un contexte aussi peu exploré que celui de l'identification de la charge émotionnelle portées par les images.

Références

[1] R.W.Picard. Affective Computing. MIT Press, Cambridge, 1997.

[2] Z. Zeng et al. A survey of affect recognition methods: audio, visual and spontaneous expressions. IEEE Transactions on PAMI, 31(1):39-58, 2009.

[3] S. Wang, X. Wang. Emotion semantics image retrieval: a brief overview. ACII, pp. 490-497, 2005.

[4] W. Wei-ning, Y. Ying-lin, and J. Sheng-ming. Image retrieval by emotional semantics: A study of emotional space and feature extraction. IEEE ICSMC, 4, 2006.

[5] W. Wang and Q. He. A survey on emotional semantic image retrieval. ICIP, pp. 117-120, 2008.

[6] C. Colombo, A. Del Bimbo, P. Pala. Semantics in visual information retrieval. IEEE Multimedia, 6(3):38-53, 1999.

[7] Al-Ani, M .Deriche. A new technique for combing multiple classifier using the Dempster Shafer theory of evidence, J. Artif. Intell. Res. 17 (2002), pp. 333-361

[8] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, the IAPS: Technical manual and affective ratings, Tech. Rep., GCR in Psychophysiology, 1999

[9] V.Yanulevskaya, J .C.Van Gemert. et al. Emotional valence categorization using holistic image features. IEEE, ICIP, pp. 101-104, 2008

[10] P. Dunker, S. Nowak, A. Begau, C. Lanz. Content-based mood classification for photos and music. ACM MIR, pp. 97-104, 2008.
[11] C.-T. Li, M.-K. Shan. Emotion-based impressionism slideshow with automatic music accompaniment. ACM Multimedia, pp. 839-842, 2007.
[12] Q. Wu, C. Zhou, C. Wang. Content-based Affective Image classification and retrieval using support vector machines. ACII, pp. 239-257, 2005.
[13] S.-B. Cho, J.-Y. Lee. A human-oriented retrieval system using interactive genetic algorithm. IEEE Transactions on systems, man and cybernetics, 32(3):452-458, 2002.
[14] K. Kuroda, M. Hagiwara. An image retrieval system by impression words and specific object names IRIS. eurocomputing, 43:259-276, 2002.
[15] A.P. Dempster. A Generalization of Bayesian Inference. J. Royal Statistical Soc. Series B, vol. 30, 1968.
[16] G. Shafer. A Mathematical Theory of Evidence. Princeton University Press, 1976.

Une analyse multirésolution adaptative pour la compression d'images multispectrales

J. Delcourt A. Mansouri T. Sliwa Y. Voisin

Le2i (Laboratoire Électronique, Informatique et Image) UMR-CNRS 5158

BP 16, Route des Plaines de l'Yonne
89010 AUXERRE Cedex – FRANCE

{jonathan.delcourt, alamin.mansouri, tadeusz.sliwa, yvon.voisin}@u-bourgogne.fr

Résumé

Dans cet article, nous nous intéressons à la compression d'images multispectrales. En particulier, nous proposons de substituer la transformée en ondelettes utilisée dans le JPEG 2000 par une analyse multirésolution adéquate que nous élaborons dans le cadre du Lifting-Scheme. Nous comparons la méthode proposée avec la méthode utilisant une transformée en ondelettes classique selon les stratégies de compression Multi-2D et Full 3D. Les deux stratégies sont combinées avec une étape de décorrélation spectrale par ACP pour optimiser leurs performances. Pour une évaluation objective, nous utilisons un cadre d'évaluation rassemblant quatre familles de métriques incluant le PSNR. De bons résultats ont été obtenus, montrant la pertinence de l'approche proposée, en particulier pour les images de grandes dimensions.

Mots clefs

Analyse multirésolution adaptative, Compression d'images multispectrales, SPIHT, SPIHT 3D.

1 Introduction

Les images multispectrales sont largement utilisées en géosciences et télédétection. Elles sont aussi de plus en plus utilisées dans d'autres champs d'application, comme l'imagerie médicale, le contrôle de qualité en industrie, la météorologie ou la mesure de couleur exacte. Une image multispectrale est générée en collectant des dizaines d'images spectrales, ou chacune d'elles est une image monochromatique centrée sur une longueur d'onde particulière du spectre électromagnétique. En conséquence, les images multispectrales sont de taille importante, avec une seule image pouvant occuper des centaines de mégaoctets. Leur compression est donc nécessaire afin de faciliter leur stockage et leur transmission.

L'une des méthodes de compression les plus efficaces pour les images monochromatiques est le JPEG 2000 [1] [1, 2]. Toutefois, son extension aux images multispectrales doit être adaptée, ce qui donne naissance à différentes stratégies. Ces stratégies reposent sur la manière dont on considère le cube multispectral :

- chaque bande spectrale de l'image est considérée séparément (ondelettes 2D + SPIHT 2D),
- le cube entier est considéré comme entrée (ondelettes 3D + SPIHT 3D).

La contribution développée dans cet article consiste en la substitution de la transformée en ondelettes utilisée par le JPEG 2000 par une analyse multirésolution adaptative que nous élaborons dans le cadre du Lifting-Scheme. Nous comparerons cette approche avec la méthode utilisant les ondelettes du JPEG 2000 dans le cadre des deux stratégies.

Dans les sections suivantes nous développerons la théorie de l'analyse multirésolution proposée. Nous décrirons ensuite les données utilisées, les expérimentations réalisées et les résultats obtenus avant de les discuter dans la troisième section. Nous conclurons dans la dernière section.

2 Description de l'analyse multirésolution proposée

Nous proposons de remplacer la transformée en ondelettes en construisant une analyse multi-résolution adaptative adéquate.

Nous nous plaçons dans le cadre de l'implémentation en Lifting-Scheme (LS) qui permet de produire aisément des analyses multi-résolution de seconde génération [3, 4]. En effet, un avantage majeur de ce cadre est que, quels que soient les filtres appliqués, la transformation inverse est déterminée explicitement et de manière exacte. Nous choisissons ici de nous limiter dans un premier temps à l'utilisation d'un schéma avec un filtre prédictif et un filtre de mise à jour. L'intuition première consiste à adapter le filtre prédictif à chaque étape de l'analyse multi-résolution, de manière à minimiser l'énergie de chaque signal de détail, et ce afin de réduire au maximum le nombre d'éléments non-nuls après quantification. Cependant, ce faisant, nous risquons de ne pas contrôler la propagation de l'erreur lors du passage par la transformée inverse. En effet, l'idéal en

1. http ://www.jpeg.org

compression est d'avoir des transformées parfaitement orthogonales de manière à transporter exactement l'énergie des signaux. Mais, lorsqu'on génère des filtres adaptatifs, rien ne garantit l'orthogonalité ni même le fait qu'on en soit proche. Pour minimiser ce problème, nous ajoutons comme contrainte le fait de limiter la non-orthogonalité par une optimisation sous contraintes.
Nous rappelons que nous nous situons dans le cadre d'un schéma avec un filtre prédictif f et un filtre de mise à jour g. Nous nous limitons ici au schéma séparable, ce qui implique des filtres directionnels à chaque échelle (horizontales, verticales, diagonales). L'opérateur linéaire correspondant à un changement d'échelle-direction et son inverse peuvent s'écrire ainsi [5, 6] :

$$\begin{pmatrix} A \\ D \end{pmatrix} = \begin{pmatrix} aI & 0 \\ 0 & bI \end{pmatrix} \begin{pmatrix} I & g \\ 0 & I \end{pmatrix} \begin{pmatrix} I & 0 \\ -f & I \end{pmatrix} \begin{pmatrix} O \\ E \end{pmatrix} \quad (1)$$

et

$$\begin{pmatrix} O \\ E \end{pmatrix} = \begin{pmatrix} I & 0 \\ f & I \end{pmatrix} \begin{pmatrix} I & -g \\ 0 & I \end{pmatrix} \begin{pmatrix} \frac{1}{a}I & 0 \\ 0 & \frac{1}{b}I \end{pmatrix} \begin{pmatrix} A \\ D \end{pmatrix} \quad (2)$$

Où A est le signal d'approximation, D le signal de détail, O et E les signaux pair et impair, a et b les coefficients de normalisation (qu'on peut imposer d'être > 0), I l'opérateur identité, et f et g sont assimilés ici à leur écriture sous forme de matrices de Toeplitz (au lieu des transformées en Z habituelles).
La non-orthogonalité d'un opérateur linéaire T peut par exemple être définie par l'une des deux écritures suivantes :

$$\varepsilon_1 = T^*T - I, \varepsilon_2 = T^* - T^{-1}, N(\varepsilon_k) \leq c \quad (3)$$

Où N est une norme matricielle et * représente l'adjonction (transposée du conjugué) ce qui revient simplement, lorsque l'on considère des filtres réels, à retourner les filtres (forme Toeplitz réelle). Comme nous sommes en dimension finie (car nombre d'échantillons fini), toutes les normes sont équivalentes. Notons aussi que les produits de filtres sont commutatifs par commutativité du produit de convolution ou, de manière équivalente, des sous-matrices de Toeplitz.
En premier lieu, (1) et (2) se simplifient en considérant classiquement ([5, 6]) :

$$b = 1/a \quad (4)$$

Par simplification du calcul direct dans (3), il apparaît qu'on annule les éléments non diagonaux d'ε_2 en posant :

$$g = f^*/a^2 \quad (5)$$

Ce qui nous amèneà simplifier la recherche de la minimisation de la non-orthogonalité parmi la famille de solutions suivantes :

$$b = \tfrac{1}{a}, g = \tfrac{1}{a^2}f^*, \begin{pmatrix} A \\ D \end{pmatrix} = \begin{pmatrix} aI - \frac{1}{a}ff^* & \frac{1}{a}f^* \\ -\frac{1}{a}f & \frac{1}{a}I \end{pmatrix} \begin{pmatrix} O \\ E \end{pmatrix} \quad (6)$$

Nous réinjectons alors (4) et (5) dans ε_1 de (3) et nous obtenons un résultat proportionnel à :

$$\begin{pmatrix} ff^*-(a^2-1)I & 0 \\ 0 & ff^*-(a^2-1)I \end{pmatrix} \begin{pmatrix} ff^*-a^2I & f^* \\ f & I \end{pmatrix} \quad (7)$$

En remplaçant T pat T^{-1}, nous obtenons cette fois un résultat proportionnel à :

$$\begin{pmatrix} ff^*-(a^2-1)I & 0 \\ 0 & ff^*-(a^2-1)I \end{pmatrix} \begin{pmatrix} I & f^* \\ f & ff^*-a^2I \end{pmatrix} \quad (8)$$

Le coefficient de proportionnalité étant une constante commune de valeur $1/a^2$. Cela n'intervient donc pas dans l'optimisation.
Nous prenons ensuite N comme la norme d'opérateur subordonnée à la norme vectorielle euclidienne. En d'autres termes, pour un opérateur A :

$$\|A\|_2 = \sup_{v \neq 0} \frac{\|Av\|_2}{\|v\|_2} = \sqrt{\rho(A^*A)} = \sqrt{\rho(AA^*)} \quad (9)$$

où $\rho(M)$ désigne le rayon spectral de la matrice M, c'est-à-dire la plus grande valeur propre de M en valeur absolue.
Les équations (7) et (8) montrent qu'il revient au même de raisonner sur T ou T^{-1}, ce qui veut dire intuitivement qu'ici les erreurs se propagent de la même manière dans le sens direct que dans le sens inverse.
En utilisant l'inégalité du produit pour la norme (9), (7) et (8) montrent aussi qu'il suffit d'évaluer indépendamment l'influence du premier et du deuxième terme du produit les constituant.
Pour le premier terme du produit, il est entièrement déterminé par :

$$ff^* - (a^2 - 1)I \quad (10)$$

qui est symétrique et permet donc de raisonner directement sur ses valeurs propres.
On calcule ensuite de manière directe le spectre de (10). Comme, pour un opérateur A et une constante α, les valeurs propres de $(A - \alpha I)$ ne sont rien d'autre que les valeurs propres de $A - \alpha$, on obtient que le spectre de (10) est le spectre de $(ff^*) - (a^2 - 1)$. Le spectre de ff^* prenant des valeurs entre $[\lambda_{min}, \lambda_{max}]$ (ff^* est symétrique réelle donc les valeurs propres sont réelles), celui de (10) prend ses valeurs dans $[\lambda_{min} - (a^2 - 1), \lambda_{max} - (a^2 - 1)]$ et on obtient donc la majoration pour (9) appliquée à (10) :

$$\max\big(|\max(Sp(ff^*))-(a^2-1)|, |\min(Sp(ff^*))-(a^2-1)|\big) \quad (11)$$

Précisons ici que, par rapport aux articles [7, 8], nous employons une notion de conditionnement plus simple, les résultats étant de toutes façons égaux à un difféomorphisme près.
Considérons maintenant le deuxième terme du produit. Tout d'abord, on constate qu'on a la relation suivante :

$$\begin{pmatrix} ff^*-(a^2+x)I & f^* \\ f & (1-x)I \end{pmatrix} \begin{pmatrix} (1-x)I & -f^* \\ -f & ff^*-(a^2+x)I \end{pmatrix}$$

$$=-x\begin{pmatrix} ff^*-\frac{1}{x}(x^2+(a^2-1)x-a^2)I & 0 \\ 0 & ff^*-\frac{1}{x}(x^2+(a^2-1)x-a^2)I \end{pmatrix} \quad (12)$$

On considère $x \neq 0$, ce qui ne pose pas de problème car ne constituant pas la plus grande valeur propre (à nouveau, les matrices sont symétriques et on raisonne directement sur les valeurs propres). Pour $x \neq 0$, nous avons :

$$\frac{1}{x}(x^2 + (a^2-1)x - a^2) = y \Leftrightarrow$$

$$x = \frac{1}{2}\left((y-(a^2-1)) \pm \sqrt{(y-(a^2-1))^2 + 4a^2}\right) \quad (13)$$

qui donne la relation entre les valeurs propres y de $ff*$ et les valeurs propres x recherchées. En posant z_{max} =(11), on obtient la majoration suivante pour le deuxième terme du produit :

$$\frac{z_{max} + \sqrt{z_{max}^2 + 4a^2}}{2} \quad (14)$$

Ce qui donne la majoration suivante pour le produit des deux termes :

$$z_{max}\frac{z_{max} + \sqrt{z_{max}^2 + 4a^2}}{2} \quad (15)$$

En identifiant cette majoration avec la contrainte c dans (3), nous obtenons :

$$z_{max} = \frac{c}{\sqrt{a^2+c}} = \frac{1}{a}\frac{c}{\sqrt{1+\frac{c}{a^2}}} < \frac{c}{a} \quad (16)$$

Le deuxième terme étant équivalent à $\frac{c}{a}$ par valeurs inférieures lorsque c tend vers 0^+. Pour a fixé, quitte à renommer c, nous pouvons remplacer c par $\frac{c}{a}$ dans (16) et nous obtenons :

$$z_{max} < c \quad (17)$$

Par conséquent, tout est entièrement déterminé par le spectre de $(ff^* - (a^2 - I))$.
Or, la norme rappelée dans (9) se réduit au rayon spectral lorsque la matrice est symétrique.
L'équivalence des normes en dimension finie nous permet alors, quitte à remplacer c par un nouveau c tendant vers 0 avec le précédent, de raisonner avec toute autre norme N à majorer (3).
A partir de là, nous prenons pour N la somme des valeurs absolues des coefficients de matrice, appliquée ici à $(ff^* - (a^2 - I))$. Comme il s'agit ici de matrices Toeplitz, il suffit de se restreindre à prendre la somme des coefficients du filtre correspondant.
Nous pouvons maintenant exprimer la minimisation sous contraintes, pour un filtre de support de longueur n paire centré sur le $\frac{n}{2}^{me}$ coefficient, sous la forme (dans [7, 8], les expressions sont données uniquement dans des cas simplifiés) :

$$\sum_{i=-\infty}^{\infty}\left(E_i - \sum_{k=-\frac{n}{2}+1}^{\frac{n}{2}} f_k\left(\tau_{-k}O\right)_i\right)^2$$

$$-\lambda\left(\sum_{i=-\frac{n}{2}+1}^{\frac{n}{2}}\left|f_i^2-(a^2-1)\right| + 2\sum_{k=1}^{\frac{n}{2}}\left|\sum_{i=-\frac{n}{2}+1}^{\frac{n}{2}-k} f_i f_{i+k}\right| - c\right) \quad (18)$$

Où τ représente l'opérateur de translation, i et k les indices aussi bien des signaux que des coefficients du filtre f, et λ un coefficient de Lagrange.
Le premier terme représente l'erreur de prédiction (énergie de la différence entre E et O filtré) et le deuxième terme la contrainte exprimant la majoration de la norme.
En pratique, pour la contrainte d'orthogonalité, nous avons déterminé empiriquement les réglages de paramètres suivants :

- Nous introduisons une borne $\mu_{max} = \frac{5}{n}$ sur la somme des valeurs absolues des coefficients du filtre pour s'assurer que le filtre ne fasse pas croître inconsidérément les amplitudes des signaux.
- $a = \sqrt{1+\mu_{max}}$ et $c = 1,25 * (1 - \frac{1}{n})$.

La taille du dictionnaire de filtres croît avec le nombre d'échelles et avec la longueur du support des filtres. Cela nous incite à ne tester que des filtres courts, en l'occurrence $n = 2$ et $n = 4$. Pour stocker les coefficients de chaque filtre adaptatif, nous lui soustrayons un filtre fixe moyen de longueur correspondante, de manière à réduire l'amplitude de ses coefficients. Nous nous basons par la suite sur leur amplitude signée maximale que nous répartissons sur 16 bits (première approche empirique).
La résolution de (18) s'effectue concrètement de la manière suivante :

- Nous testons toutes les combinaisons de signes possibles pour les valeurs absolues du deuxième terme de (18).
  Pour chaque combinaison, le système se résout de manière algébrique exacte. Ici, pour éviter de rajouter quelque erreur, tout ce qui peut être résolu en calcul symbolique l'est, y compris les calculs matriciels, que cela concerne les éléments purement symboliques ou les valeurs numériques.
  - Nous testons si la solution obtenue vérifie bien les hypothèses de signe de départ.
  - Nous testons de plus la condition sur μ_{max}.
- Nous comparons l'effet de chaque solution et nous sélectionnons celle qui aboutit à l'énergie la plus faible.

3 Expérimentations et résultats

3.1 Données et expérimentations

Nous avons réalisé nos expérimentations sur l'image multispectrale Cuprite, constituée de 32 bandes spectrales et codée en 16 bits entier, provenant du capteur AVIRIS [2]. Nous avons utilisé différentes dimensions spatiales de cette image ($64*64$, $128*128$ et $256*256$ pixels). Nous avons pour objectif de comparer les performances de l'approche adaptative proposée à la compression classique en faisant varier le bitrate de compression ainsi que les dimensions spatiales de l'image.

2. http ://aviris.jpl.nasa.gov

Pour optimiser la compression d'image multispectrale, une étape de décorrelation spectrale par ACP est appliquée. En conséquence de quoi, nous obtenons une nouvelle image multi-bandes dans le domaine de la transformée, pour laquelle la corrélation spectrale est réduite. Nous appliquons ensuite les stratégies de compression sur l'image transformée.

Deux stratégies de compression sont utilisées dans cet article, les stratégies Multi-2D et Full 3D. Pour l'approche 2D, chaque image propre issue de l'ACP est compressée séparément. Ensuite un codage par SPIHT 2D est appliqué sur chaque bande résultante de la transformée pour terminer la compression. En raison de l'ACP, les bandes de l'image résultante sont ordonnancées par énergie décroissance. Afin de prendre en compte ce fait, il est préférable de pondérer chaque bande. Comme pondération nous utilisons l'énergie définie par :

$$E = \sqrt{\sum_{x,y} I_\lambda(x,y)^2}/(XY) \tag{19}$$

ou I_λ est la bande spectrale de l'image à la longueur d'onde λ, X et Y sont ses dimensions, et x et y sont les positions du pixel dans la bande spectrale. En fonction de l'énergie de la bande spectrale, un nombre de bits est alloué pour la sortie de l'algorithme du SPIHT.

L'approche 3D consiste à considérer l'intégralité du cube multispectrale comme entrée pour la décomposition 3D. Pour achever la compression un SPIHT 3D [9] est ensuite appliqué.

3.2 Cadre d'évaluation de la compression

Quand la compression avec perte est utilisée, il est nécessaire d'évaluer et de décrire la nature et l'importance des dégradations dans l'image reconstruite (décompressée). Selon Eskicioglu [10], le principal problème dans l'évaluation des techniques de compression avec perte est la difficulté de décrire la nature et l'importance des dégradations sur l'image reconstruite. Dans le cas d'image 2D ordinaire, une métrique doit souvent tenir compte de la perception visuelle d'un observateur humain. Ce n'est pas le cas pour les images multispectrales, qui sont utilisées pour des classifications ou des reproduction de couleurs spectrales. C'est pourquoi au lieu d'évaluer les performance de la compression en fonction d'une seule métrique ou d'un seul type de métrique, nous utilisons neuf métriques que nous classons dans quatres familles pour évaluer les performance. Nous utilisons ceci en tant que cadre d'évaluation de la compression.

Les métriques que nous proposons peuvent être divisées en quatre différentes familles : critères issues d'une extension isotrope de critères de traitement du signal (PSNR), métriques orientées multispectral (fidélité spectrale F_λ), un critère statistique évolué prenant en compte certains aspects perceptifs (UIQI) et une métrique orientée classification (K-means). Dans [11], Christophe *et al.* démontrent que l'utilisation d'un jeu de plusieurs métriques est plus pertinent que de n'en utiliser un seul. Ainsi, nous avions proposé dans [12] l'utilisation de quatre familles de métriques en plus du PSNR. Nous utilisons ce dernier car c'est la métrique la plus employée, ce qui facilite la comparaison avec les résultats d'autres méthodes.

3.3 Résultats et discussion

Nous comparons la méthode multirésolution adaptative proposée à la méthode par ondelettes classique issue du JPEG 2000 selon deux stratégies de compression (Multi-2D et Full 3D). Nous obtenons ainsi quatres méthodes de compression : Multi-2D (M2D), Multi-2D multirésolution (MR-M2D), Full 3D (F3D) et Full 3D multirésolution (MR-F3D). Nous nous attachons à la comparaison des méthodes appartenant à la même stratégie suivant les différentes métriques. Les figures 1 et 2 représentent les résultats en termes de PSNR, fidélité spectrale, UIQI et K-means pour des tailles d'image de $128 * 128$ et $256 * 256$ pixels respectivement, alors que la figure 3 donne une représentation plus compacte comprenant les quatre métriques pour un bitrate de 1 bpp.

Pour la stratégie Multi-2D :

- En termes de PSNR et de fidélité spectrale, la méthode M2D donne de meilleurs résultats que la méthode MR-M2D excepté pour une taille d'image de $64 * 64$ où cette tendance et inversée.
- En termes de UIQI, la méthode M2D donne de moins bons résultats que la méthode MR-M2D. Cette métrique est plutôt basée sur la mesure des distorsions structurelles que sur la sensibilité aux erreurs.
- En termes de K-means, les méthode M2D et MR-M2D donnent des résultats fortement similaires. Une légère différence peut-être notée pour l'image de taille $128 * 128$ pixels où la méthode M2D est légèrement meilleure que la méthode MR-M2D pour des valeurs de bitrates supérieures à 0.25 bpp. Cette tendance est inversée pour l'image de taille $256 * 256$ pixels.

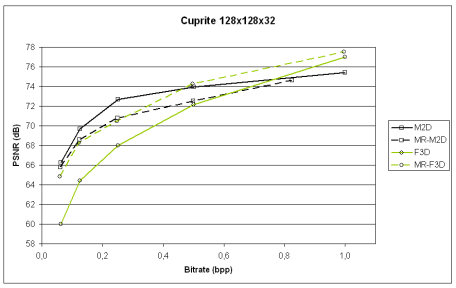
Pour la stratégie Full 3D :

- Pour toutes les tailles d'images et pour la majorité des métriques la méthode MR-F3D donne des résultats sensiblement meilleurs que la méthode F3D. Cette différence de résultats est particulièrement importante pour les faibles valeurs de bitrate et augmente avec les dimensions spatiales de l'image.
- La méthode F3D surpasse la méthode MR-F3D seulement en terme de K-means pour une taille d'image de $256 * 256$ pixels et pour une valeur de bitrate supérieure ou égale à 0.5 bpp.

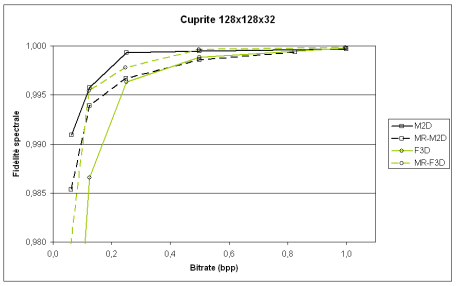
Pour la stratégie Full 3D, l'approche proposée surpasse l'approche classique dans quasiment tous les cas. Nous pouvons mettre en évidence le fait que la qualité des résultats, pour des valeurs faibles de bitrate, augmente proportionnellement au dimensions spatiales de l'image : les résultats de la méthode MR-F3D sont meilleurs que ceux de la méthode F3D d'au moins ≈ 0.5 dB et au plus de ≈ 8 dB. Pour de grandes valeurs de bitrates (≈ 1 bpp) les résultats

des deux méthodes deviennent très proches.

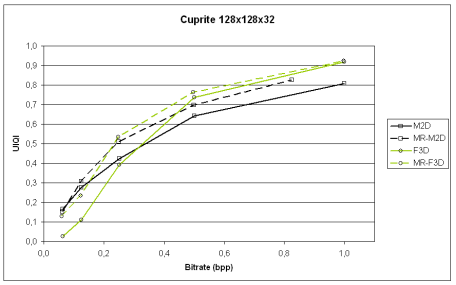
Pour la stratégie Multi-2D, la méthode M2D obtient des résultats légèrement meilleurs que la méthode MR-M2D. Nous pouvons expliquer cela par le fait que l'analyse multi-résolution crée un dictionnaire pour chaque bande spectrale de l'image, diminuant la place disponible pour enregistrer le résultats de SPIHT. Le seul cas où la méthode MR-M2D surpasse la méthode M2D est en termes de UIQI, ce qui nous permet d'en déduire que la méthode proposée créer moins de distortions structurelles que la méthode classique.



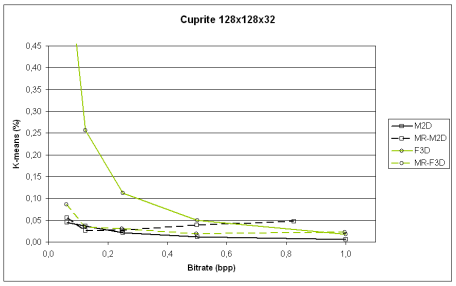
(1)



(2)

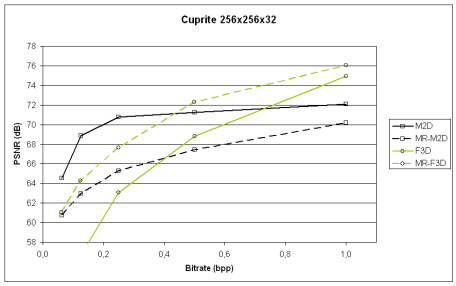


(3)

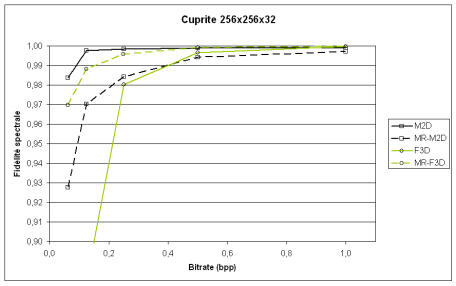


(4)

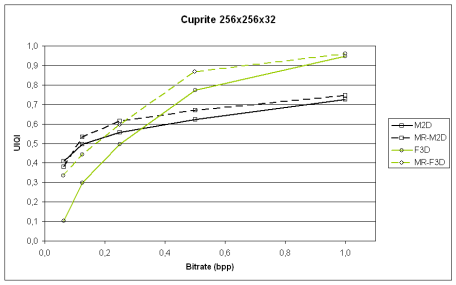
Figure 1 – *Résultats de compression sur l'image Cuprite de dimensions spatiales* $128 * 128$ *pixels en termes de PSNR (1), fidélité spectrale (2),UIQI () et K-means (4).*



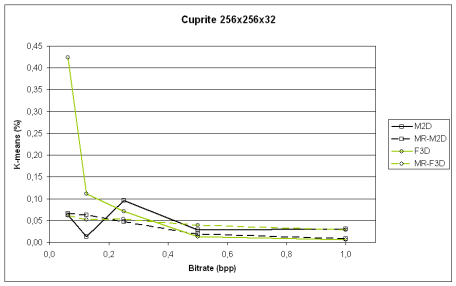
(1)



(2)

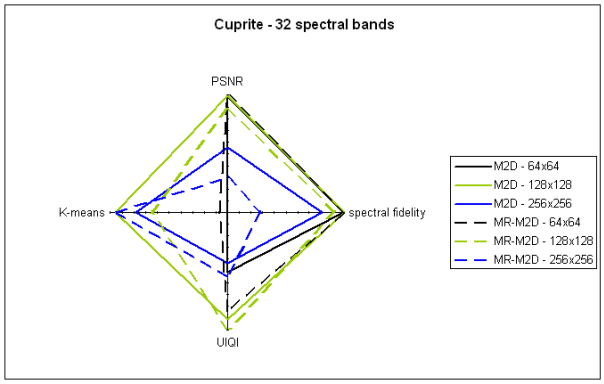


(3)

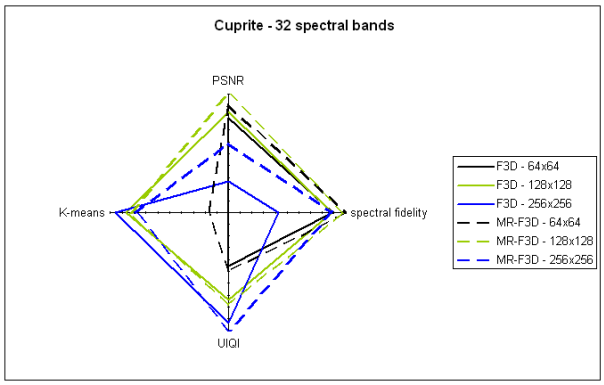


(4)

Figure 2 – *Résultats de compression sur l'image Cuprite de dimensions spatiales* $256 * 256$ *pixels en termes de PSNR (1), fidélité spectrale (2),UIQI () et K-means (4).*



(1)



(2)

Figure 3 – *Résultats des stratégies Multi-2D (1) et Full 3D (2) sur l'image Cuprite de dimensions spatiales* $64 * 64$, $128 * 128$ *et* $256 * 256$ *pixels en termes de PSNR, fidélité spectrale, UIQI et K-means.*

4 Conclusion

Dans cet article, nous avons introduit une nouvelle analyse multirésolution adaptative pour la compression d'images multispectrales. Cette analyse à été implémentée dans le cadre de deux stratégies de compression : les stratégies Multi-2D et Full 3D, et comparée à une implémentation utilisant la décomposition en ondelettes classique du JPEG 2000.

Les expérimentations réalisées et les résultats obtenus montrent que l'approche proposée est plus appropriée à la compression d'images multispectrales dans le cadre de la stratégie Full 3D et plus particulièrement pour des images de grandes dimensions spatiales.

Références

[1] C. Christopoulos, A. Skodras, et T. Ebrahimi. The JPEG 2000 still image coding system : An overview. *IEEE Transactions on Consumer Electronics*, 46(4) :1103–1127, 2000.

[2] D.S. Taubman, M.W. Marcellin, et M. Rabbani. JPEG2000 : Image compression fundamentals, standards and practice. *Journal of Electronic Imaging*, 11 :286, 2002.

[3] W. Sweldens et P. Schroder. Building your own wavelets at home. *ACM SIGGRAPH course notes*, pages 15–87, 1996.

[4] W. Sweldens. The lifting scheme : A construction of second generation wavelets. *Technical Report, Department of Mathematics, University of South Carolina*, 6, 1995.

[5] I. Daubechies et W. Sweldens. Factoring wavelet transforms into lifting steps. *Journal of Fourier Analysis and Applications*, 4(3) :247–269, 1998.

[6] M. Maslen et P. Abbott. Automation of the lifting factorisation of wavelet transforms. *Computer Physics Communications*, 127(2-3) :309–326, 2000.

[7] T. Sliwa, Y. Voisin, et A. Diou. Near-orthogonal and adaptive affine lifting scheme on vector-valued signals. Dans *Proceedings of SPIE*, 2003.

[8] T. Sliwa, Y. Voisin, et A. Diou. Adaptivity with near-orthogonality constraint for high compression rates in lifting scheme framework. Dans *Proceedings of SPIE*, volume 5208, page 107, 2004.

[9] L. Dragotti, G. Poggi, et A.R.P. Ragozini. Compression of multispectral images by three-dimensional SPIHT algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, 38(1) :416–428, 2000.

[10] A.M. Eskicioglu et P.S. Fisher. Image quality measures and their performance. *IEEE Transactions on Communications*, 43(12) :2959–2965, 1995.

[11] E. Christophe, D. Léger, et C. Mailhes. Quality criteria benchmark for hyperspectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 43(9) :2103, 2005.

[12] J. Delcourt, A. Mansouri, T. Sliwa, et Y. Voisin. A comparative study and an evaluation framework of multi/hyperspectral image compression. Dans *5th International Conference on Signal-Image Technology and Internet-Based Systems (SITIS 2009)*, 2009.

Index des auteurs