# Suivi d'objets multiples représentés par leurs frontières pour l'interpolation temporelle dans une séquence d'images

**Laurent Bonnaud et Claude Labit**

IRISA/INRIA-Rennes, Campus de Beaulieu, 35042 Rennes Cédex,
E-mail: <nom>@irisa.fr

RÉSUMÉ

Dans cet article, nous présentons une représentation d'une segmentation basée mouvement par les frontières de ses régions. À cette structure est associé un algorithme de suivi temporel capable de prendre en compte plusieurs objets, en présence d'occlusions partielles. Cette représentation de segmentation est composée des frontières entre tous les couples de régions adjacentes, approximées polygonalement. Nous l'enrichissons avec une étiquette sur chaque frontière indiquant à quelle région celle-ci appartient du point de vue du mouvement. Cette information est utilisée pour un suivi prédictif de cette structure au cours du temps. De plus, nous montrons que l'information structurelle contenue dans cette représentation est très utile pour des régions en occlusion. Cette représentation peut être utilisée à de nombreuses fin à la fois en analyse de séquences d'images (par exemple la fusion de régions, leur mise en couches) et dans le domaine du codage et des applications multimédia (par exemple la compression par compensation de mouvement, l'interpolation temporelle dans une séquence d'images, la manipulation d'objets). Parmi de nombreuses applications, nous nous intéressons en particulier à l'interpolation temporelle faite du côté récepteur d'une chaîne de codage/décodage.

ABSTRACT

In this paper we present a boundary-based representation of a motion-based image sequence segmentation. A tracking algorithm is associated to this structure which is able to handle several objects simultaneously, in the presence of partial occlusions. This segmentation representation consists in the boundaries between all couples of adjacent regions approximated by a polygonal line. We extend it with a label on each boundary indicating to which region it is belonging from a motion point of view. This information is useful for an efficient predictive tracking of this representation over time. Moreover, we show that the structural information enclosed in our representation is very useful when regions are occluding. This representation can be used for multiple purposes in image sequence analysis (for example region fusion, layering) as well as coding and multimedia applications (for instance motion-compensated compression, temporal interpolation, object manipulation). Among many applications, we particularly consider a temporal image sequence interpolation which can be done at the receiver end of a coder/decoder chain.

## 1 Introduction

Dans le contexte de MPEG4 [8], les problèmes de manipulation d'objets, de base de données multimedia font l'objet d'un intérêt croissant. La plupart des algorithmes dans ce domaine ont besoin d'une segmentation basée mouvement de bonne qualité. Par exemple, notre application à l'interpolation nécessite une segmentation stable dans le temps et dont les frontières sont précisément situées sur les contours des objets.

La représentation la plus simple d'une segmentation mosaïque est une carte des étiquettes de régions. Une telle représentation peut être suivie par une prédiction compensée en mouvement suivie d'un ajustement utilisant un modèle markovien et une minimisation d'une fonction énergétique globale [1, 3]. Les inconvénients de cette phase d'ajustement résident dans le fait que la stabilité ne peut être garantie (des régions peuvent disparaître ou changer d'étiquette) et que les contours ne sont pas très précis.

Des travaux [7] ont étendu cette représentation pour prendre en compte les frontières entre régions. Cet article et les suivants des mêmes auteurs montrent des segmentations utilisant des opérateurs morphologiques qui donnent des contours plus précis. Cependant, le suivi est toujours fait au niveau des régions décrites comme un ensemble de pixels (voir la figure 1). L'ordre de superposition des régions est défini sur une structure auxiliaire : le graphe d'adjacence des régions.

Des représentations de segmentation par frontières de régions ont été utilisées dans le domaine des contours actifs. Mais dans ces travaux demeure une limitation importante sur le nombre des objets suivis ; en général, le cas des objets se recouvrant ou se découvrant n'est pas bien traité. Une solution est donnée dans [10]. Une segmentation est représentée par un ensemble de régions décrites par un contour fermé polygonal. Le suivi se décompose en 2 étapes : une prédiction du contour par application du mouvement de la texture de la région, un raffinement par polygones ajustables, et un traitement *a posteriori* des intersections entre régions et des zones découvertes.

La représentation de segmentation décrite dans cet article repose entièrement sur les frontières entre régions. Elle peut être suivie par des modèles de contours actifs donnant une bonne localisation des contours, tout en étant plus efficace que les autres représentations. Elle se prête aussi mieux à l'introduction d'information structurelle permettant de résoudre en amont le problèmes des occlusions, grâce à une étiquette supplémentaire sur chaque frontière indiquant son appartenance à l'une des 2 régions limitrophes.

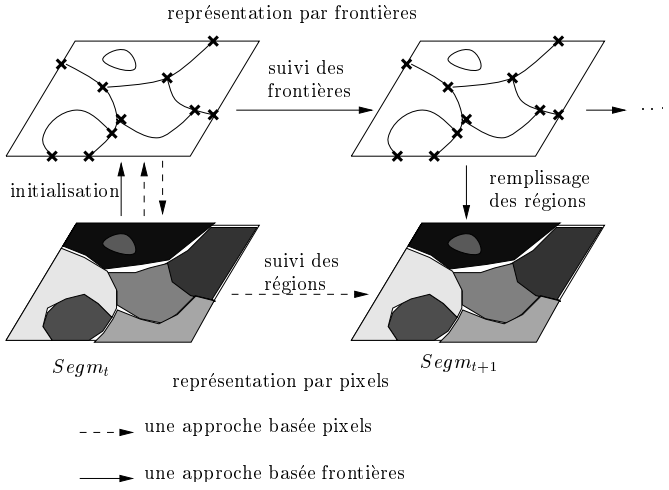


FIG. 1 — Représentations par pixels et par frontières.

## 2 Une représentation par frontières

Cette représentation de segmentation est constituée de 3 listes. Une 1ère liste contient les "points multiples" de la segmentation. Un point multiple est un point dans le voisinage duquel se trouvent 3 régions (ou plus). Trois frontières de régions (ou plus) aboutissent à un point multiple. Une 2ème liste contient les frontières séparant les couples de régions adjacentes, approximées par une ligne polygonale (voir la figure 3). On distingue 2 types de frontières. Les frontières ouvertes sont celles qui relient 2 points multiples ; elles sont stockées sans leurs extrémités pour que la structure ne soit pas redondante. Les frontières fermées, ou boucles, sont celles qui ne contiennent pas de point multiple. Une 3ème liste contient les régions. Chaque région est définie par son contour extérieur et ses éventuels contours intérieurs, pour une région à trous. Chaque contour est lui même une liste d'indices dans la liste des frontières. Contrairement à [10], cette représentation est non redondante car chaque frontière n'est stockée qu'une seule fois, bien qu'elle fasse partie du contour de 2 régions. Comme nous allons le voir, cela permet un algorithme de suivi plus efficace.

Sa construction est faite ainsi. On part d'une segmentation initiale basée pixels. Dans l'exemple de la figure 3, il s'agit d'une segmentation spatiale obtenue grâce à une modélisation en régions de niveau de gris constant et optimisée selon le critère du MDL [4]. Les points multiples sont extraits en premier. Ensuite les chaînes de contours de Freeman les reliant sont extraites, suivies par les boucles. Les frontières sont ensuite approximées par des lignes polygonales.

## 3 Suivi de la structure

Cette structure représente une segmentation basée mouvement où le mouvement apparent de la texture de chaque région est homogène et décrit par un modèle affine. La notation $\Theta^{\pm}_{\mathcal{R},,t_1\to t_2}$ désigne un mouvement de l'image $I_{t_1}$ vers l'image $I_{t_2}$ pour la région $\mathcal{R}$. L'exposant est un $^+$ si $t_1 < t_2$ ou un $^-$ if $t_2 < t_1$. Comme dans [10], le suivi est réalisé par un algorithme de prédiction/ajustement, mais nous suivons les frontières ouvertes au lieu des contours fermés.

**Étape de prédiction** On suppose connue la segmentation $S_{t-1}$ de l'image $I_{t-1}$ et l'on cherche $S_t$. On prédit le mouvement $\Theta^{+}_{\mathcal{R}_i,,t-1\to t}$ pour toutes les régions $\mathcal{R}_i$ grâce à un filtre de Kalman sur les paramètres comme dans [3]. Mais contrairement à cette référence, c'est directement $\Theta^{+}_{\mathcal{R}_i,,t-1\to t}$ qui nous intéresse, au lieu de $\Theta^{-}_{\mathcal{R}_i,,t\to t-1}$ car cet algorithme n'est plus utilisé pour un codage par compensation de mouvement. Ensuite ce mouvement prédit est utilisé comme initialisation pour un algorithme d'estimation de mouvement robuste multirésolution [6] qui cherche à minimiser l'EQM de compensation de mouvement de la texture de la région. Enfin, chaque frontière, y compris ses points multiples, est prédite avec le mouvement estimé de la région $\mathcal{R}_i$ à laquelle elle appartient.

**Étape d'ajustement** Elle se décompose en 3 phases : un ajustement affine [3], la recréation des points multiples, et un ajustement par déformation libre. Soit $F$ une frontière polygonale définie par ses sommets $(x_k)_{k\in F}$ (y compris les points multiples) et $\Theta$ un mouvement affine. On définit l'énergie $E^d_F(\Theta)$ comme l'intégrale de $-\|\nabla I_t\|$ le long des segments déplacés $\Theta([x_k, x_{k+1}])$. Pour la 1ère phase, les frontières appartenant à une même région $\mathcal{R}_i$ sont regroupées dans un ensemble $\mathcal{F}_i$. On cherche alors le mouvement $\Theta$ qui minimise la somme des énergies $E^d_F(\Theta)$ pour $F \in \mathcal{F}_i$ avec une descente de gradient. Ce regroupement de frontières donne plus de robustesse à la méthode que si elles étaient ajustées séparément ou que si toutes les frontières d'un contour de région étaient prises en compte, à cause des occlusions.

À ce stade, les points multiples n'ont plus d'existence puisque certaines frontières ont été déconnectées entre elles. Pour obtenir de nouveau une structure complète pour la phase suivante, il est nécessaire de recréer les points multiples. Les intersections entre frontières sont d'abord détectées et donnent la localisation des nouveaux points multiples. Les frontières sont coupées en conséquence. Ensuite, les frontières toujours déconnectées sont prolongées par un segment les reliant à leur voisin le plus proche. Nous avons préféré cet algorithme purement géométrique qui n'utilise pas d'information image car il pourra être utilisé tel quel pour l'interpolation. De même, il permet un fonctionnement synchrone du codeur et du décodeur, sans transmission d'informations autres que les mouvements.

La 3ème phase est une déformation libre. On fait en sorte qu'elle soit suffisamment petite pour ne pas changer la topologie de la structure de sorte qu'il est possible d'en tirer partie. Pour chaque frontière $F$ de sommets $x_k$ (y compris les points multiples), le raffinement est un vecteur $D$ contenant les déplacements individuels $d_k$. Comme précédemment, on définit l'énergie d'attache aux données $E^d_F(D)$ pour la frontière $F$ déplacée de $D$. Puis on définit une énergie favorisant de petits déplacements par

$$E^s_F(D) = \alpha \sum_{k\in F} \|d_k\|^2$$

et une énergie de régularisation par

$$E^r_F(D) = \beta \sum_{k\in F} \|d_k - d_{k+1}\|^2$$

pour favoriser la continuité des déplacements [5]. Chaque point multiple a un unique déplacement même s'il contribue à 3 frontières, ce qui contribue à régulariser la solution de façon structurelle. Enfin, on minimise la somme sur toutes les frontières des 3 termes énergétiques. Il s'agit d'une minimisation dans un domaine continu car à ce stade les $x_k$ ont des coordonnées réelles, et elle est réalisée par descente de gradient. Pour quantifier ces coordonnées en nombre entiers, on fait une dernière minimisation avec un algorithme HCF. Jusqu'à maintenant toutes les minimisations étaient locales à une frontière et pouvaient donc être traitées en parallèle. Si on choisit d'effectuer cette dernière phase, il n'y a plus de localité, mais il exite des algorithmes parallèles de relaxation.

Dans [10], chaque frontière est ajustée 2 fois car elle appartient à 2 contours de 2 régions. De plus, le post-traitement des zones découvertes doit être effectué à chaque image. Dans notre algorithme, chaque frontière est ajustée une fois seulement. Le traitement des zones découvertes et des occlusions est implicite grâce à l'attribution de chaque frontière à une région, utilisée comme une information structurelle supplémentaire. Cela fonctionne tant qu'une région découverte $\mathcal{R}_d$ reste en contact avec la région $\mathcal{R}_r$ qui la recouvre partiellement. Mais si après plusieurs images ces régions se séparent, cette méthode n'est plus valable. Pour détecter cette situation, on vérifie l'EQM calculée par l'estimation de mouvement de la texture de $\mathcal{R}_d$. Si elle augmente suffisamment, on recalcule une segmentation spatiale de $\mathcal{R}_d$ pour détecter les régions nouvellement apparues dans l'espace entre $\mathcal{R}_d$ et $\mathcal{R}_r$. On réalise tout de même un gain de complexité puisque cette situation ne se produit que dans une image alors que les cas où le suivi marche se produisent dans toutes les autres images.

# 4 Affectation des frontières

Nous avons vu que dans l'étape de prédiction du suivi, on prédit chaque frontière avec le mouvement de l'une des régions qu'elle délimite. Ceci est fait selon l'hypothèse qu'une frontière a le même mouvement apparent que le mouvement de la texture de la région à laquelle est appartient. Cette région est aussi celle qui est située au dessus de sa voisine.

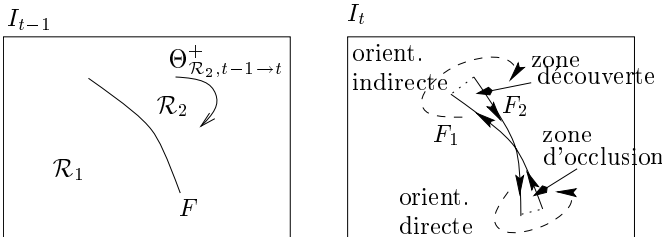


FIG. 2 — Un critère basé mouvement pour déterminer l'appartenance d'une frontière. Comme seul le mouvement relatif des 2 régions est important, on peut supposer que $\mathcal{R}_1$ a un mouvement nul.

Soit $F$ une frontière entre 2 régions $\mathcal{R}_1$ et $\mathcal{R}_2$ à l'instant $t-1$. Pour déterminer son appartenance à l'instant $t$, on considère les 2 prédictions $F_i = \Theta^{+}_{\mathcal{R}_i,,t-1\to t}(F)$, $i = 1, 2$. La zone d'occlusion ou de découvrement pour $F$ est le polygône $P$ défini par $F_1$ et $F_2$ et les 2 segments reliant les points multiples (les lignes pointillées dans la figure 2). Le critère est différent pour chacun des 2 types de zone donc il est nécessaire de déterminer le type de chaque zone. La figure 2 montre un exemple complexe où les 2 types sont présents simultanément. $F_i$ est orienté de sorte que $\mathcal{R}_i$ se trouve à sa gauche. Avec cette convention, une zone d'occlusion $Ocl$ est contenue dans un polygône orienté dans le sens trigonométrique direct, et une zone découverte $Dec$ dans un polygône orienté dans le sens indirect. Ensuite on calcule les EQM suivantes :

$$EQM_i^{Ocl} = \sum_{p\in Ocl} [I_t(p) - I_{t-1}(\Theta^{-}_{\mathcal{R}_i,,t\to t-1}(p))]^2 \text{ et}$$

$$EQM_i^{Dec} = \sum_{p\in Dec} [I_t(p) - I_{t+1}(\Theta^{+}_{\mathcal{R}_i,,t\to t+1}(p))]^2.$$

Finalement $F$ appartient à la région $\mathcal{R}_i$ qui minimise $EQM_i^{Ocl} + EQM_i^{Dec}$.

Si la surface de $P$ est trop petite, ce critère n'est pas fiable. Sauf dans le cas peu probable où la frontière est une ligne droite et que le mouvement relatif entre les 2 régions lui est parallèle, cela signifie que le mouvement relatif est petit. Dans ce cas, n'importe quel choix n'influencera pas beaucoup la prédiction donc on l'effectue avec la moyenne des 2 mouvements et la frontière est étiquetée comme "ambiguë".

Les bords de l'image sont traités à part : une région spéciale est créée en dehors de l'image. Elle a un mouvement nul et est située au dessus de toutes les autres régions, de sorte que dans l'algorithme de suivi, les bords de l'image restent statiques.

# 5 Interpolation

Des systèmes d'interpolation existent déjà dans les convertisseurs numériques entre standards de TV de fréquences différentes [9]. Ils se trouvent en amont du codeur si bien qu'ils peuvent utiliser des algorithmes performants (basés sur l'estimation d'un champ dense de mouvement avec discontinuités) mais tournant sur du matériel complexe. Un défaut de ces systèmes est cependant la transmission d'images redondantes si la fréquence en sortie est supérieure à la fréquence en entrée. Dans [2], nous avons proposé un algorithme de faible complexité pouvant être utilisé dans un simple décodeur car il n'utilise que les informations transmises à ce décodeur (segmentation, descripteurs de mouvement) sans avoir besoin de les recalculer. Un champ dense de mouvement peut permettre une meilleure interpolation qu'un algorithme basé régions et modèles de mouvement. Mais l'avantage de notre approche est qu'une unique transmission à une fréquence fixée peut être décodée sur différents types de décodeurs fonctionnant à des fréquences différentes (*multicast*). Elle est aussi utile dans le contexte du codage progressif ou robuste aux erreurs pour interpoler des images manquantes causées par une réduction du débit, des erreurs sur le canal ou des paquets perdus.

L'algorithme de suivi peut être adapté pour fournir une implantation du mode interpolatif que nous avions appelé "prédiction bidirectionnelle de segmentation" [3]. Il s'agit d'un mode dans lequel, pour les images interpolées, le codeur ne transmet que les descripteurs de mouvement sans la segmentation. Mais celle ci est toujours transmise pour les images de

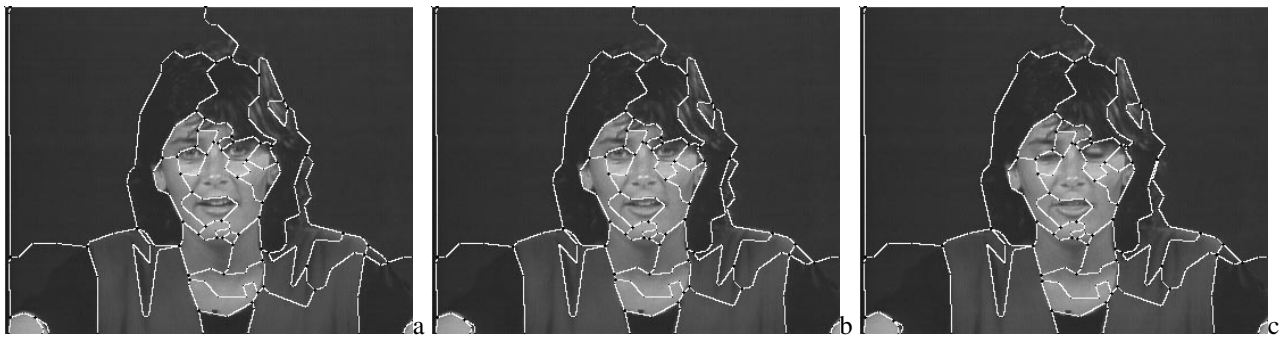

FIG. 3 — Segmentation des 3 premières images. Les frontières sont en blanc et les points multiples sont représentés par des croix noires. Les sommets des polygones sont marqués par des points noirs.

référence. Le décodeur peut appliquer les mouvements transmis aux frontières de la segmentation d'une image de référence. Pour reconstruire la segmentation manquante, il peut les considérer comme des frontières après ajustement affine et appliquer la phase de l'algorithme où les points multiples sont reconstruits.

L'étiquette d'ambiguïté qui est mise sur certaines frontières est aussi très utile pour l'interpolation. Quand un objet de mouvement complexe est segmenté en plusieurs petites régions de mouvement homogène, les frontières internes de cet objet sont probablement de ce type à cause de la continuité du mouvement global. Dans le processus d'interpolation, on teste si la position compensée en mouvement d'un pixel à interpoler appartient à la même région. Dans ce cas, il est raisonnable d'autoriser une interpolation au delà des frontières, donc, à titre de perspective, nous proposerions une modification de ce test pour inclure les régions adjacentes séparées par une frontière ambiguë.

## 6 Résultats, conclusion

La figure 3a montre la première image de la séquence. Pour obtenir une bonne initialisation, la segmentation pixellique initiale basée sur le critère MDL a été retouchée. Les régions sont d'abord re-étiquettées pour les rendre connexes. On fait ensuite subir à la carte d'étiquettes un filtrage majoritaire, de sorte à lisser les aspérités des contours. À chaque étape, les régions trop petites sont éliminées par fusion avec la région adjacente de moyenne la plus proche. Finalement, une fois la représentation par frontières extraite, les points multiples trop proches sont fusionnés pour éliminer les frontières de trop petite taille.

On constate sur les images 3b et 3c que le suivi assure un bon positionnement des frontières sur les contours des objets, tout en maintenant une cohérence temporelle très forte. Il est toutefois à noter que la gestion d'une telle structure de représentation est très lourde et délicate, de nombreux cas particuliers géométriques pouvant se présenter. Pour ces raisons, ce genre de méthode ne peut marcher que pour des segmentations ne comportant pas trop de petites régions et de frontières de faible longueur.

## Références


[1] Black (M.). – Combining intensity and motion for incremental segmentation and tracking over long image sequences. *In : Proceedings of ECCV'92*, pp. 485–493. – Santa Margherita Liguere, Italy, Mai 1992.

[2] Bonnaud (L.) et Labit (C.). – Codage interpolatif de séquences d'images utilisant un suivi temporel de segmentation spatio-temporelle. *In : Actes du GRETSI 95*. – Juan les pins, Sept. 1995.

[3] Garcia-Garduño (V.), Labit (C.) et Bonnaud (L.). – Temporal linking of motion-based segmentation for object-oriented image sequence coding. *In : Proc. of EUSIPCO 94*. – University of Edinburgh, Scotland, UK, Sept. 1994.

[4] Leclerc (Y. G.). – Constructing simple stable descriptions for image partitioning. *International Journal of Computer Vision*, vol. 3, 1989, pp. 73–102.

[5] Mardia (K.), Hainsworth (T.) et Haddon (J.). – Deformable templates in image sequences. *In : Proceedings of ICPR*, pp. 132–135. – La Haye, The Netherlands, Sept. 1992.

[6] Odobez (J.) et Bouthemy (P.). – Robust multiresolution estimation of parametric motion models. *Journal of Visual Communication and Image Representation*, vol. 6, n4, Déc. 1995, pp. 348–365.

[7] Pardas (M.), Salembier (P.) et Gonzalez (B.). – Motion and region overlapping estimation for segmentation-based video coding. *In : Proceedings of ICIP'94*, pp. 428–432. – Austin, Texas, Nov. 1994.

[8] Pereira (F.). – MPEG4 : A new challenge for the representation of audio-visual information. *In : PCS'96*. – Melbourne, Australia, 1996.

[9] Tziritas (G.) et Labit (C.). – *Motion analysis for image sequence coding — Motion-compensated image interpolation*, chap. 7, pp. 269–285. – Elsevier, 1994.

[10] Wu (L.), Benois (J.) et Barba (D.). – Spatio-temporal segmentation of image sequences for object-oriented low bit-rate coding. *In : Proceedings of ICIP'95*, pp. 406–409. – Washington, D.C., USA, Oct. 1995.