Codage vidéo par schéma lifting avec gestion des occlusions

Thomas ANDRÉ, Marc ANTONINI, Michel BARLAUD

Laboratoire I3S, CNRS - UMR 6070, Université de Nice-Sophia Antipolis
Bât. Algorithmes / Euclide B, BP 121, 2000 route des Lucioles - 06903 Sophia Antipolis Cedex, France
{andret,am,barlaud}@i3s.unice.fr

Résumé – Le schéma lifting compensé en mouvement est utilisé dans la plupart des codeurs vidéo basés ondelettes. Cependant, l'estimation et la compensation de mouvement à l'aide de blocs entraîne l'apparition d'artéfacts visibles autour des objets en mouvement et du bord des images. Dans ce papier, nous proposons une nouvelle méthode de filtrage temporel qui fait appel à une segmentation et une estimation de mouvement conjointes. Le principe consiste à attribuer un mouvement à des régions de forme adaptable au lieu d'utiliser des blocs. Nous présentons d'une part l'algorithme de filtrage «Puzzle» et étudions les conditions de son inversibilité. D'autre part, nous proposons une méthode d'extraction des régions d'occlusion à partir des informations de segmentation et de mouvement ; ces régions sont ensuite utilisées pour gérer les occlusions. Les premiers résultats expérimentaux confirme la diminution des effets de blocs ; la bonne gestion des occlusions permet une baisse significative de l'entropie des sous-bandes temporelles.

Abstract – Motion-compensated lifting schemes have become a reference for the temporal filtering of video data. However, block-based motion estimation and compensation produce annoying blocking artifacts around the moving objects and near the borders of the images. In this paper, we propose a new lifted temporal filtering method, based on joint segmentation and motion estimation. This method consists in attributing locally the motion information to content-adapted regions instead of blocks. We first present our "Puzzle filtering" algorithm and we state the conditions for its invertibility. Then, we propose a method to extract regions of occlusion from the motion and segmentation information. The obtained regions are finally exploited within the proposed Puzzle filtering. First experimental results show that the blocking artifacts are completely removed and that the occlusions are successfully managed, which results in an important subband entropy decrease.

1 Introduction

La compression vidéo connaît un essor important depuis quelques années. Des normes très efficaces ont été établies, aboutissant par exemple aux codeurs hybrides MPEG-4 et H.264/AVC. Les codeurs basés ondelettes [1, 2], qui font intervenir un schéma lifting $t + 2D$ compensé en mouvement [3], permettent un meilleur support de la scalabilité [4] et atteignent presque les performances des codeurs hybrides [5]. Toutefois, de nombreux travaux de recherches visent encore à en améliorer les performances, et notamment celles du filtrage temporel en ondelettes. En particulier, la plupart des codeurs vidéo cités ci-dessus font appel à une estimation et une compensation du mouvement à base de «block-matching», qui ne gère pas les occlusions ni le chevauchement d'objets. En conséquence, systématiquement, certains blocs sont situés à cheval sur deux régions de mouvements différents, ce qui provoque l'apparition d'effets de blocs. Une meilleure gestion de ce problème devrait permettre d'améliorer grandement les performances de codage.

Nous proposons une méthode de filtrage temporel fondée sur le schéma lifting compensé en mouvement [5], qui utilise les informations d'estimation de mouvement et de segmentation, afin de mieux prendre en compte les objets en mouvement. Nous supposons qu'il n'y a dans l'image que deux régions différentes, un «objet» et un «fond» ; cette hypothèse, réaliste si locale, permet de simplifier le problème, d'autant plus que la segmentation ne doit pas être nécessairement sémantique pour que l'algorithme fonctionne.

Dans ce papier, nous décrivons tout d'abord une méthode de filtrage temporel compensé en mouvement fondée sur le schéma lifting, que nous appelons «Puzzle», capable de filtrer différentes régions avec un mouvement différent (section 2). Ensuite, nous décrivons un algorithme capable d'extraire les régions utiles à la gestion des occlusions (section 3), à partir des informations fournies par un algorithme conjoint de segmentation et d'estimation de mouvement. Enfin, nous présentons les premiers résultats expérimentaux, qui montrent l'efficacité de la méthode pour gérer les occlusions et donc diminuer fortement l'entropie des sous-bandes temporelles.

2 Schéma lifting orienté régions

2.1 Diviser un macrobloc en régions

La compensation de mouvement a permis une amélioration significative des performances des codeurs vidéos, même avec un modèle très simple de mouvement de translation uniforme par blocs. Cependant, cette technique entraîne l'apparition d'artéfacts (effet de blocs) dans les images reconstruites. Plusieurs méthodes ont été proposées pour réduire ces effets de blocs. Par exemple, les derniers standards MPEG et H.264, ainsi que de nombreux codeurs basés ondelettes, font appel à des blocs de taille variable. D'autres alternatives, comme les maillages déformables [6, 7], ont également été envisagées.

Nous proposons une méthode flexible qui utilise des régions de forme arbitraire. La figure 1 montre en effet qu'il est souvent plus précis de diviser un macrobloc en deux régions plutôt qu'en blocs plus petits. Nous avons choisi de modéliser la limite des régions par quatre points de contrôle interpolés par

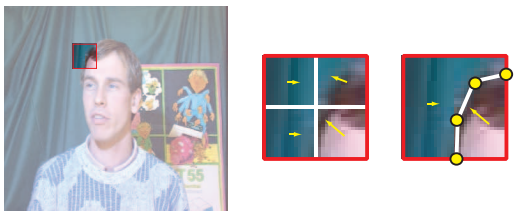


FIG. 1 – Division d'un macrobloc en deux régions adaptatives au lieu de quatre blocs.

une spline, qui permet une représentation précise de la courbe tout en limitant son coût de codage. Il est à noter que le surcoût d'information à transmettre est comparable dans les deux cas. En effet, 4 vecteurs se codent sur $4 \times 2 \times N_v$ bits, alors que 2 vecteurs et 4 points se codent sur $2 \times 2 \times N_v + 4 \times 2 \times N_p$ bits, où $N_v \in [5;8]$ est la précision (en bits) des coordonnées des vecteurs, et $N_p \in [4;5]$ celle des coordonnées des points, fonctions des paramètres d'estimation du mouvement et de segmentation. Par ailleurs, nous montrons en section 3 que cette représentation en régions permet une gestion précise des occlusions.

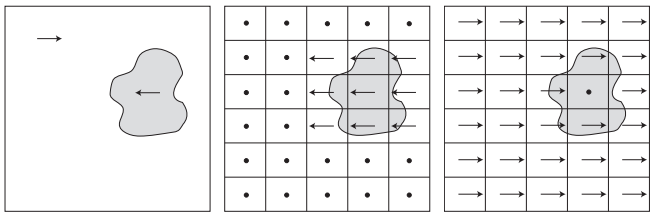
La méthode proposée nécessite au préalable de connaître la segmentation de chaque image en régions de mouvements différents (Fig. 2). Cette information peut être obtenue manuellement (par exemple en post-production cinématographique), ou bien à l'aide d'un algorithme conjoint de segmentation et d'estimation du mouvement [8, 9].

2.2 Filtrage orienté régions

Dans cette partie, nous supposons connue la segmentation de chaque image de la séquence en deux régions : un «objet» arbitraire, et un «fond» arbitraire. Cette segmentation n'est pas nécessairement sémantique, et le rôle de l'objet et du fond est totalement symétrique. Nous supposons également que nous connaissons le mouvement respectif de ces deux régions. Ces informations sont données par un algorithme conjoint d'estimation de mouvement et de segmentation, ou bien par une segmentation manuelle suivie d'une estimation de mouvement classique, telle que celle utilisée dans MPEG.

L'objectif est de filtrer temporellement l'objet et le fond de façon indépendante, en utilisant leur mouvement respectif, à l'aide par exemple de la transformée en ondelettes (2,2) sous sa forme de schéma lifting. La figure 2a montre l'exemple d'une séquence simple dans laquelle un objet rigide se déplace sur un fond également en mouvement. L'estimateur de mouvement découpe chaque image en macroblocs, puis détermine dans chaque macrobloc un «objet» et un «fond», d'après les informations de segmentation. Ensuite, il détermine séparément le mouvement de ces deux régions, à l'aide d'un algorithme classique de mise en correspondance. Finalement, on obtient deux régions par macrobloc, et un vecteur mouvement par région et par bloc, comme indiqué sur les figures 2b et 2c.

Notons (I_i) les images de la séquence originale. Pour chaque image i, notons M_i^{obj} le masque de l'objet : pour chaque pixel p de l'image, on aura $M_i^{obj}(p) = 1$ si p est situé sur l'objet, et $M_i^{obj}(p) = 0$ sinon. Le masque du fond, noté M_i^{fond}, est



(a) Segmentation (b) Mouvement : objet (c) Mouvement : fond

FIG. 2 – Exemple d'une séquence simple où un objet rigide (gris) se déplace sur un fond en mouvement. Il y a un vecteur mouvement par bloc pour l'objet, et un par bloc pour le fond, représentés par des flèches. Un point représente un vecteur nul.

défini de la même façon, et on a $M_i^{fond} = \overline{M_i^{obj}}$.

Notons également $v_{i\to j}^{obj}(b)$ le vecteur représentant le mouvement de l'objet du bloc b, de l'image i vers l'image j. Le mouvement du fond, $v_{i\to j}^{fond}(b)$, est défini de la même façon. Ces vecteurs permettent de compenser en mouvement des images ou des masques ; par exemple, on notera $I_{2i-1}\left(v_{2i-1\to 2i}^{fond}\right)$ l'image I_{2i-1} compensée en mouvement à l'aide des vecteurs mouvement du fond $v_{2i-1\to 2i}^{fond}(b)$, donc recalée temporellement sur l'image I_{2i}.

2.2.1 Analyse

La séquence originale (I_i) peut être filtrée temporellement avec les filtres de la transformée (2,2) compensée en mouvement. Selon si le mouvement du fond ou celui de l'objet est utilisé, on obtient la sous-bande haute-fréquence H_i^{fond} ou H_i^{obj}, calculées comme suit :

$$H_i^{fond} = I_{2i} - \frac{1}{2}\left[I_{2i-1}\left(v_{2i-1\to 2i}^{fond}\right) + I_{2i+1}\left(v_{2i+1\to 2i}^{fond}\right)\right]$$

$$H_i^{obj} = I_{2i} - \frac{1}{2}\left[I_{2i-1}\left(v_{2i-1\to 2i}^{obj}\right) + I_{2i+1}\left(v_{2i+1\to 2i}^{obj}\right)\right]$$

Puisque l'on souhaite filtrer l'objet avec le mouvement de l'objet, et le fond avec le mouvement du fond, il suffit d'exprimer la sous-bande haute-fréquence finale H_i sous la forme :

$$H_i = H_i^{fond} * M_{2i}^{fond} + H_i^{obj} * M_{2i}^{obj}$$

où l'opérateur $*$ représente la multiplication terme à terme de deux matrices. Finalement, H_i est localement égale à H_i^{fond} ou bien à H_i^{obj}, les deux masques M_i^{obj} et M_i^{fond} étant complémentaires dans l'image I_i.

La sous-bande basse-fréquence est calculée de façon similaire.

2.2.2 Synthèse

La réversibilité du schéma proposé peut être établie rapidement à partir des équations précédentes. En inversant le schéma lifting de façon classique, on obtient :

$$I_{2i-1} * \left[M_{2i-1}^{fond} + M_{2i-1}^{obj}\right] = L_i + f_L(H_{i-1}, H_i)$$

$$I_{2i} * \left[M_{2i}^{fond} + M_{2i}^{obj}\right] = H_i + f_H(I_{2i-1}, I_{2i+1})$$

où $f_L(H_{i-1}, H_i)$ et $f_H(I_{2i-1}, I_{2i+1})$, obtenues après un calcul simple, sont des expressions indépendantes de l'image que l'on cherche à reconstruire. Puisque la réunion des masques M_i^{obj} et M_i^{fond} couvre la totalité de la surface de l'image, ceci établit la réversibilité de la transformée.

2.2.3 Généralisation à N régions

Par la suite, nous verrons que la gestion des occlusions nécessite de distinguer plus de 2 régions, qui seront traitées différemment, avec des mouvements ou des filtres différents. Supposons que chaque macrobloc est maintenant composé de N régions. Notons M_i^n le masque qui caractérise la région n de l'image i, et $v^n_{i\to j}$ le mouvement de cette région n entre l'image i et l'image j.

En nous appuyant sur les résultats précédents, nous pouvons écrire par exemple la sous-bande HF comme suit :

$$H_i = \sum_{n=1..N} H_i^n * M_{2i}^n$$

avec, dans le cas d'un filtrage (2,2) :

$$H_i^n = I_{2i} - \frac{1}{2}\left[I_{2i-1}(v^n_{2i-1\to 2i}) + I_{2i+1}(v^n_{2i+1\to 2i})\right]$$

La réversibilité est alors établie de la même façon que précedemment, en inversant le schéma lifting. Par exemple, en inversant le pas de prédiction, on obtient l'équation de reconstruction suivante :

$$I_{2i} * \left[\sum_{n=1..N} M_{2i}^n\right] \quad = \quad H_i + f_H(I_{2i-1}, I_{2i+1})$$

En d'autres termes, le schéma proposé est réversible si les masques $(M_i^n)_{n=1..N}$ sont complémentaires dans l'image i.

Les calculs qui concernent le pas de mise à jour sont similaires.

3 Gestion des occlusions

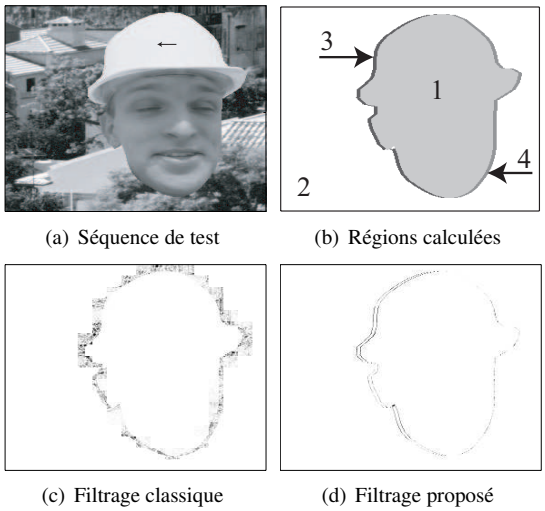
Le schéma de filtrage proposé en section 2, dans sa version généralisée, peut être utilisé pour gérer les occlusions lors du filtrage temporel compensé en mouvement. Il suffit, en effet, de subdiviser les deux régions «objet» et «fond» en sous-régions occultées ou non, et d'appliquer le bon filtre à chaque sous-région.

3.1 Régions et filtrage

Ainsi, par exemple, le fond de l'image I_i peut être subdivisé en trois régions : le fond qui est apparu par rapport à l'image précédente I_{i-1}, le fond qui est apparu par rapport à l'image suivante I_{i+1}, et le fond qui est visible partout sur les trois images successives. Cette dernière région est filtrée à l'aide d'une transformée symétrique, comme la transformée (2,2). Les deux autres régions le sont par une transformée causale ou anti-causale selon la direction souhaitée, comme l'ondelette de Haar. De cette façon, les régions occultées n'interviennent pas dans le calcul des sous-bandes.

L'objet est traité de la même manière que le fond ; au total, six sous-régions doivent être déterminées afin de traiter les occlusions.

La figure 3b montre un exemple simple d'une telle séparation en sous-régions. Dans cet exemple, l'objet n'est pas occulté : il est donc entièrement prédictible en bidirectionnel et sera filtré par transformée (2,2). En revanche, le fond doit être divisé en 3 sous-régions afin de tenir compte des occlusions ; les sous-régions occultées seront filtrées à l'aide de l'ondelette de Haar dans la direction appropriée. Les sous-bandes HF et BF finales seront obtenues à l'aide des équations généralisées données en section 2.2.3.



(a) Séquence de test (b) Régions calculées

(c) Filtrage classique (d) Filtrage proposé

FIG. 3 – Séquence de test utilisée (a) ; régions extraites par l'algorithme proposé (1 = objet prédictible dans les deux directions, 2 = fond prédictible dans les deux directions, 3 = fond prédictible à partir de l'image suivante, 4 = fond prédictible à partir de l'image précédente) ; exemple de sous-bandes haute-fréquence (HF) obtenues avec un filtrage classique (c) et avec l'algorithme proposé (d).

3.2 Extraction des régions

Il est maintenant nécessaire de déterminer ces 6 sous-régions, en utilisant uniquement des informations disponibles au codage et au décodage. Nous proposons donc un algorithme qui permet de les extraire à partir des informations sur la segmentation simple en «objet» et «fond», et sur le mouvement de ces deux régions. Par exemple, déterminons la région du fond de I_i qui est prédictible à partir de l'image suivante I_{i+1} seulement ; cette région est définie par son masque que nous noterons $M_i^{fond,suiv}$. Il s'agit de la zone du fond qui est visible dans l'image I_i mais pas dans l'image I_{i-1} ; en d'autres termes, elle correspond au fond de l'image I_i, privé du fond de l'image I_{i-1} que l'on a recalé au préalable sur l'image I_i :

$$M_i^{fond,suiv} = M_i^{fond} - \left[M_i^{fond} \bigcap M_{i-1}^{fond}\left(v_{i-1\to i}^{fond}\right)\right]$$

Le masque du fond prédictible à partir de l'image précédente I_{i-1} seulement, noté $M_i^{fond,prc}$, est calculé de la même manière. On s'assure alors que les deux masques sont complémentaires. Dans le cas contraire, on leur retranche leur partie commune, qui correspond à la zone du fond qui n'est prédictible dans aucune direction. Enfin, la zone du fond prédictible dans les deux directions est calculée en retranchant à la zone du fond les deux zones calculées précédemment.

Le traitement du «fond» et de «l'objet» est réalisé de manière symétrique.

4 Premiers résultats

La figure 3 montre les premiers résultats expérimentaux obtenus avec une séquence de test synthétique, dans laquelle un

(a) Filtrage classique (b) Filtrage par Puzzle

FIG. 4 – Filtrage HF de la séquence «Erik» par les deux méthodes : image 2. Les pixels les plus foncés représentent les plus grandes valeurs absolues.

objet rigide texturé (la tête du personnage) bouge sur un fond texturé. La segmentation de l'objet a été ici réalisée à la main, et contient une erreur d'un pixel sur la position du contour. Le mouvement a été estimé à l'aide d'un algorithme de «block-matching» classique, pour une taille de blocs de 16×16 pixels.

L'algorithme d'extraction de régions présenté en section 3 fournit 4 régions (b), et non 6, puisque l'objet est rigide et ne disparaît pas d'une image à l'autre. Un filtrage par schéma lifting classique compensé en mouvement fournit des sous-bandes haute-fréquence (HF) telles que celle montrée en (c). L'algorithme proposé fournit la sous-bande (d), qui ne contient plus d'effets de blocs mais seulement des «contours» d'erreur dus à l'imprécision de la segmentation. L'étude de l'énergie et de l'entropie des sous-bandes (tableau 1) confirme que ces dernières peuvent être codées bien plus efficacement par la méthode proposée que par la méthode classique.

Filtres		$(2,2)$		$(2,0)$	
Filtrage		B	P	B	P
Puissance	HF $(.10^{-3})$	9.4	1.0	9.4	1.0
	BF $(.10^{4})$	1.54	1.53	1.53	
Entropie (bpp)	HF	1.36	0.17	1.36	0.17
	BF	8,52	7.75	7.60	

TAB. 1 – Comparaison de la méthode classique de filtrage par blocs (B) et de la méthode proposée (P), pour deux ensembles de filtres temporels [10] : puissance et entropie de la sous-bande haute-fréquence (HF) et basse-fréquence (BF), moyenne sur les 8 premières images.

Nous avons également testé la méthode proposée sur la séquence réelle «Erik» ; un exemple des sous-bandes HF obtenues est présenté sur la figure 4. Nous avons comparé les sous-bandes obtenues en subdivisant des macroblocs de 32×32 en blocs de 16×16 (méthode classique) ou en régions (méthode proposée) là où c'était nécessaire, c'est-à-dire dans les zones d'occlusion. Si les sous-bandes produites par les deux méthodes avaient une énergie et une entropie comparables, il apparaît que la méthode proposée permet néanmoins d'obtenir des sous-bandes plus propres, avec beaucoup moins d'effets de blocs.

5 Conclusion

Nous avons présenté une nouvelle méthode de filtrage temporel par schéma lifting qui utilise des informations de segmentation et d'estimation de mouvement conjointes afin de gérer les occlusions. Cette méthode consiste à appliquer un schéma lifting temporel adapté à des régions, de forme quelconque, qui prennent en compte les occlusions. Ces régions sont calculées à partir des seules informations de segmentation et de mouvement. Les premiers résultats montrent une amélioration de la qualité des sous-bandes et une diminution des effets de blocs autour des objets en mouvement. Les travaux futurs concerneront l'intégration de cette méthode dans un codeur vidéo complet.

Références

[1] S.J. Choi and J.W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. on Image Processing*, vol. 8, no. 2, pp. 155–167, Feb. 1999.

[2] S. Cho and W.A. Pearlman, "A full-featured, error-resilient, scalable wavelet video codec based on the Set Partitioning in Hierarchical Trees (SPIHT) algorithm," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 12, pp. 171–170, Mar. 2002.

[3] J. Viéron, C. Guillemot, and S. Pateux, "Motion compensated 2D+t wavelet analysis for low rate fgs video compression," in *Proc. of Tyrrhenian Intern. Workshop on Digital Comm.*, Capri, Italy, Sept. 2002.

[4] G. Pau, C. Tillier, B. Pesquet-Popescu, and H. Heijmans, "Motion compensation and scalability in lifting-based video coding," *EURASIP Signal Processing : Image Communication, special issue on Wavelet Video Coding*, pp. 577–600, Aug. 2004.

[5] M. Cagnazzo, T. André, M. Antonini, and M. Barlaud, "A model-based motion compensated video coder with JPEG2000 compatibility," in *IEEE Intern. Conf. on Image Processing*, Singapore, Oct. 2004, pp. 2255–2258.

[6] A. Secker and D. Taubman, "Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression," *IEEE Transaction on Image Processing*, vol. 12, no. 12, pp. 1530–1542, Dec. 2003.

[7] N. Cammas and S. Pateux, "Fine grain scalable video coding using 3d wavelets and active meshes," in *SPIE Visual Communications and Image Processing, VCIP 2003*, January 2003.

[8] E. Debreuve, M. Gastaud, M. Barlaud, and G. Aubert, "A region-based joint motion computation and segmentation on a set of frames," in *Proceedings of European Workshop on Image Analysis for Multimedia International Services (WIAMIS)*, Montreux, Swiss, Apr. 2005.

[9] S. Boltz, E. Debreuve, and M. Barlaud, "A joint motion segmentation algorithm for video coding," in *Proceedings of EUSIPCO*, Antalya, Turkey, 2005 (to appear).

[10] T. André, M. Cagnazzo, M. Antonini, M. Barlaud, N. Božinović, and J. Konrad, "(N,0) motion-compensated lifting-based wavelet transform," in *Proc. IEEE Intern. Conf. on Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004.