Compression vidéo FGS à bas débit basée sur une décomposition ondelettes 2D+t

Jérôme Viéron, Christine Guillemot et Henri NICOLAS
IRISA/INRIA Rennes
Campus universitaire de Beaulieu
35042 Rennes Cedex
Firstname.Lastname@irisa.fr

Résumé

Ce papier décrit un algorithme de compression vidéo finement scalable et à bas débit permettant une adaptation aisée des flux aux caractéristiques non stationnaires des réseaux de transmission tels que l'Internet. La solution proposée ici repose sur une décomposition par ondelettes spatio-temporelles compensées en mouvement. L'algorithme EBCOT [1] est ensuite appliqué à des groupes de sous-bandes spatio-temporelles permettant une optimisation globale débit-distorsion par groupe d'images. Les performances sont comparées aux standards MPEG-4 et H.264.

Mots Clef

Vidéo, ondelettes 3D, finement scalable, bas débit

1 Introduction

Le domaine de la compression vidéo a connu, ces dernières années, de fortes évolutions menant à l'émergence d'un nombre important de standards internationaux (H.26X, MPEG-X). Malgré le nombre important de solutions, la compression reste un domaine de recherche ouvert notamment dans le cadre de transmision audiovisuelle sur différents types de canaux filaires ou non. L'arrivée de ce type d'infrastructures a également été à l'origine de nombreux travaux visant à optimiser la qualité de service de bout-en-bout. Ces travaux concernent la compression bas débit mais également la résistance aux pertes et aux erreurs et, plus généralement, la flexibilité d'adaptation des flux compressés aux caractéristiques non stationnaires du réseau. En particulier, le concept de scalabilité à grain fin (FGS) a été introduit afin de permettre, notamment, l'adaptation de flux pré-encodés dans des scénarios de streaming. Les techniques de compression vidéo basées sur des décompositions spatio-temporelles sont des solutions privilégiées pour répondre à cet objectif. L'un des problèmes posé dans la conception de telles approches est le choix de la transformation temporelle et de son couplage avec les modèles de mouvement. Les critères de choix sont conditionnés par un compromis entre une faible énergie résiduelle (bonne exploitation de la redondance temporelle), la fiabilité du modèle de mouvement et son coût de codage. Ainsi, la première utilisation d'une transformée en ondelettes-3D [2] a été rafinée au cours de la denière décennie par la prise en compte de ces critères. Dans [3] et [4] les auteurs opèrent tout d'abord une compensation de mouvement globale afin d'aligner spatialement les images sur une grille de plus forte résolution avant d'appliquer le filtrage temporel. Des compensations de mouvement locales sont considérées dans [5], [6]. Les principales limitations dont souffrent ces approches sont inhérentes aux modèles de mouvement et résident dans la gestion des pixels non connectés, ceux-ci ayant un impact important sur le processus de décorrélation temporelle. Des schémas de lifting compensés en mouvement ont alternativement été proposés dans [7] et [8] afin de contourner ce problème. Un schéma de lifting peut être vu comme une implémentation efficace de la transformée temporelle. Chaque opération de lifting s'apparente à une prédiction compensée en mouvement bi-directionnelle. Une des principales limites de ces approches réside dans l'augmentation très significative du côut du mouvement.

Le second problème posé alors est le codage des sous-bandes spatio-temporelles générées. On trouve dans la littérature différentes approches reposant sur des extensions d'outils de compression d'images fixes [9, 10, 11]. Différentes propositions de solutions, faisant suite notamment aux travaux [6] et [11], sont actuellement à l'étude au sein d'un groupe de travail MPEG. Il faut noter, toutefois, que l'essentiel des approches proposées dans la littérature font l'hypothèse de codage haut voire très haut débit.

L'algorithme de compression vidéo que nous proposons ici se place dans le cadre de la compression bas débit. Il met d'abord en oeuvre une décomposition par ondelettes de Haar temporelles compensées en mouvement sur un groupe d'images (GOF), suivie d'une décomposition spatiale basée sur une technique de lifting 9-7. L'algorithme EBCOT [1] est ensuite utilisé afin de coder les groupes de sous-bandes spatio-temporelles. Une prédiction temporelle inter-GOF permet, de plus, d'exploiter la corrélation temporelle résiduelle. Les résultats d'expérimentation montrent un gain important par rapport

à MPEG-4 part 2 et proches de ceux obtenus avec H.264 (TML8 version 8.4 sans optimisation globale R-D).

2 Structure générale

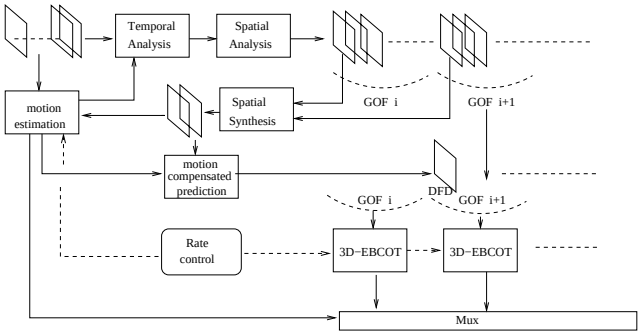

FIG. 1 – *Architecture globale.*

L'architecture du codeur proposé ici est donné sur la figure 1. Les images sont traitées par groupe appelé GOF. Après une phase d'estimation de mouvement, les GOF passent par une phase de transformation temporelle compensée en mouvement. Chaque sous-bande temporelle est ensuite décomposée par un filtre d'ondelettes bi-orthogonal 9-7 dans les directions horizontale et verticale, à l'aide d'un schéma de lifting. Les sous-bandes spatio-temporelles sont ensuite quantifiées uniformément. Enfin, celles-ci sont encodées avec le codeur arithmétique basé contexte EBCOT permettant d'obtenir un train binaire scalable à grain fin.

3 Mouvement : modèle et codage

Afin de réguler finement le débit alloué aux champs de mouvement, l'estimation de mouvement basée bloc utilise une structure d'arbre contraint en débit (i.e. *Quadtree*). La taille des blocs est ensuite adaptée aux caractéristiques des mouvements locaux au sens débit-distorsion. Le débit fait reférence ici au budget (en bits) alloué à l'encodage des vecteurs mouvement et la distorsion fait reférence à l'EQM résultante. Afin d'accélerer l'estimation de mouvement, nous utilisons une estimation hiérarchique. Les vecteurs mouvement obtenus dans une première étape (à faible résolution) sont ensuite raffinés pour les résolutions supérieures. De plus, les estimations des blocs de tailles importantes, situés dans les premiers niveaux du quadtree, servent de base pour l'estimation de mouvement des blocs de tailles inférieures. Enfin, afin d'obtenir un champ de mouvement plus lissé, limitant les pixels connectés, nous utilisons ici une méthode d'estimation de mouvement avec recouvrement de blocs. Les tailles de blocs varient entre 64x64 et 8x8 et l'estimation est réalisée avec une précision pixelique. Après élagage du quadtree, les vecteurs mouvement sont codés de manière prédictive. Le prédicteur utilisé est la valeur médiane des vecteurs associés aux blocs voisins. L'erreur de prédiction est alors codée par des codes de Huffman.

4 Filtrage Spatio-temporel compensé en mouvement

4.1 Filtrage temporel

Les techniques basées sur une estimation de mouvement bi-directionnelle sont souvent avancées comme permettant l'utilisation de filtre temporel long. La longueur des filtres peut, en effet, contribuer à améliorer l'exploitation de la redondance temporelle, induisant ainsi une diminution de l'énergie résiduelle présente dans les hautes fréquences. Cependant, les longueurs de filtres sont en étroite corrélation avec la continuité du mouvement dans la dimension temporelle. Ainsi, afin de pleinement bénéficier de la longueur des filtres, un pixel doit se trouver sur une unique trajectoire de mouvement. En pratique, et particulièrement lorsque l'estimation de mouvement est basée bloc, c'est très rarement le cas. De ce fait, ce type d'approches ne répondent pas vraiment au problème. Aucune continuité de mouvement entre les images successives ne peut, en effet, être garanti. Il a été montré, de plus, dans [12], que le compromis entre une faible énergie résiduelle et le coût de codage de l'information de mouvement, fourni par ces approches, s'avère non satisfaisant à bas débit. C'est pourquoi, nous avons choisi d'utiliser ici un filtre temporel court (le filtre de Haar) appliqué sur 3 niveaux, permettant d'obtenir un filtre équivalent plus long.

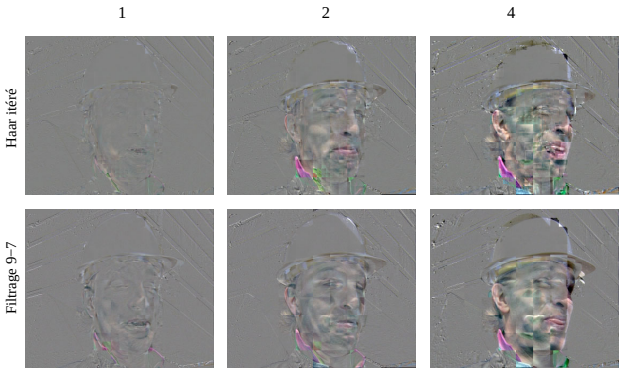

FIG. 2 – *Energie résiduelle présente dans des sous-bandes temporelles haute fréquence lorsque le débit du mouvement est contraint à environ 40 kbit/s pour une approche utilisant un filtre 9-7 versus un filtrage de Haar.*

La figure 2 illustre la répartition de l'énergie résiduelle dans différentes sous-bandes basse fréquence, lorsque le coût du mouvement est contraint à 35% du débit global fixé à $140kbit/s$, pour la méthode basée sur un filtrage de Haar (i.e. de longueur 2) et la méthode utilisée dans [7] basée sur un filtre 9-7. La taille de GOF est fixé à 8. Les sous-bandes illustrées représentent les différents types de sous-bandes haute fréquence obtenues lors des 3 différents niveaux de décomposition. On peut voir que l'énergie résiduelle est plus faible avec la technique utilisant le filtre de Haar.

La technique de gestion des pixels connectés et non connectés retenues est celle proposée dans [6]. Chaque pixel de l'image de référence t et son correspondant dans l'image $t+1$, s'il existe, définissent une paire de pixels *connectés* (s'il en existe plusieurs, le premier dans l'ordre lexicographique est choisi). Soit $p' = p + d$ et d le vecteur déplacement associé au pixel p, alors le filtrage de ces pixels est donné par

$$\begin{cases} L_t(p') & = & \frac{1}{\sqrt{2}}(I_{t+1}(p) + I_t(p')) \\ H_{t+1}(p) & = & \frac{1}{\sqrt{2}}(I_{t+1}(p) - I_t(p')). \end{cases}$$

Les autres pixels sont dits *non connectés*. Lorsqu'un tel pixel p est présent dans l'image t, une basse fréquence temporelle est produite par

$$L_t(p) = \frac{2}{\sqrt{2}} I_t(p). \quad (1)$$

Les pixels *non connectés* de l'image $t+1$ produisent, quant à eux, une haute fréquence donnée par

$$H_{t+1}(p) = \frac{1}{\sqrt{2}}(I_{t+1}(p) - I_t(p')). \quad (2)$$

4.2 Filtrage spatial

Nous utilisons ici une implémentation lifting d'un filtre de Daubechies 9-7 pour la transformation ondelettes spatiale. Afin d'améliorer les performances, des niveaux de décomposition spatiale différents selon les sous-bandes temporelles sont utilisés. Ainsi, sur la basse fréquence temporelle 3 niveaux de décomposition sont utilsés alors que sur les sous-bandes hautes fréquences seulement 2 niveaux sont appliqués. En effet, il n'est pas nécessaire de décomposer plus car la quantité d'information à décorréler dans les hautes fréquences temporelles est moins importante. Par conséquent, 2 niveaux de décomposition sont également utilisés pour l'ondelette spatiale appliquée à la DFD utilisée dans la prédiction inter-GOF.

5 Prédiction inter-Gof

Une prédiction temporelle inter-GOF a également été rajoutée au système de codage. On distinguera alors deux types de GOF : Intra et Inter. Ce mécanisme de prédiction temporelle est réalisé en boucle fermée et necessite un champ de mouvement supplémentaire. La prédiction en boucle fermée peut-être réalisée en prenant comme information de reférence une image (sous-bande) décodée à un débit plus faible, comme dans les couches basses d'une représentation scalable classique.

Après le codage d'un GOF Intra, la basse fréquence temporelle est reconstruite à l'encodeur par une phase de décodage et de synthése temporelle. Cette sous-bande reconstruite est ensuite utilisée comme information de reférence dans la compensation de mouvement de la sous-bande basse fréquence temporelle du GOF suivant (Inter).

6 Codage des sous-bandes

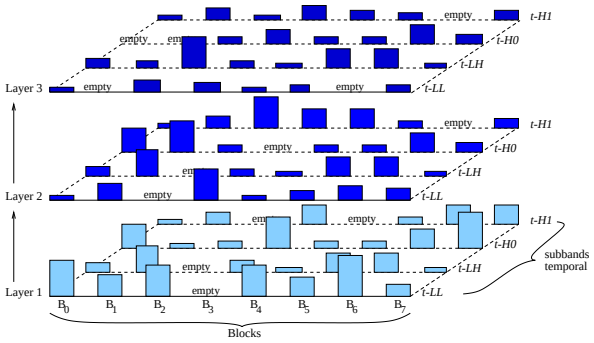

FIG. 3 – *Couches de qualité EBCOT-3D.*

Les coefficients quantifiés des sous-bandes sont codés en utilisant un algorithme de type EBCOT étendu à la dimension temporelle (cf figure 3). Le principe de base consiste à décomposer chacune des sous-bandes en blocs (typiquement 64x64 coefficients). Chaque bloc est ensuite compressé indépendamment à l'aide d'un codeur arithmétique contextuel. Les trains binaires obtenus sont troncables en un nombre multiple de points. Etant donné un débit alloué à une couche donnée, le train binaire de chacun des blocs est tronqué de façon à minimiser la distorsion globale (algorithme PCRD : *Post Compression Rate-Distortion*) associée au décodage de la couche considérée. Cette optimisation, menant à la formation de couches de qualité (cf figure (3)), est particulièrement bien adaptée à une régulation fine de l'information de texture, et permet d'obtenir une scalibilité à grain fin. De plus, la flexibilité permise dans l'agencement des paquets EBCOT permet d'obtenir un train binaire hautement scalable.

7 Résultats expérimentaux

La figure 4 compare les PSNR obtenus avec différentes solutions à 140 kbit/s : (a) l'approche proposée avec que des GOF Intra, (b) l'approche incluant les GOF Inter, (c) MPEG-4 part2 à 165kbit/s et (d) H.264 à 110kbit/s(TML8 version 8.4 sans optimisation R-D globale).

La séquence de test utilisée est *foreman* à 15 Hz. Les résultats obtenus avec notre technique de codage 2D+t ont une précision de mouvement pixelique alors que MPEG-4 a une précision au $1/2$ pixel et H.264 au $1/4$ pixel. L'implémentation MPEG-4 met en oeuvre, de plus, une compensation de mouvement avec recouvrement de bloc (OBMC). La complexité de l'algorithme décrit ici est inférieure à celle des références MPEG-4 et H.264. Il faut noter également que ces derniers ne supportent aucun mécanisme de régulation de débit et encodent à qualité constante. Ceci explique pourquoi, leurs PSNRs sont plus stables. Toutefois, cette stabilité mène à des variations de débit importantes sur les segments vidéo comportants des forts mouvements. La figure 5 permet des observations similaires pour un débit de 200 kbit/s.

On peut voir, tout d'abord, que l'approche utilisant des

GOF Inter (un sur deux) permet d'améliorer les résultats par rapport à l'approche purement Intra. Le schéma proposé offre également des performances significativement supérieures à MPEG-4. De plus, à bas débit les performances obtenues sont assez proches de celles de H.264. Pour une complexité nettement moindre, la solution proposée est finement scalable. Les résultats sont, de plus, finement contrôlés en débit sur des fenêtres de 8 images, et sont ainsi directement utilisables pour la transmission sur réseau, ce qui n'est pas le cas des implémentations MPEG-4 et H.264 considérées ici.

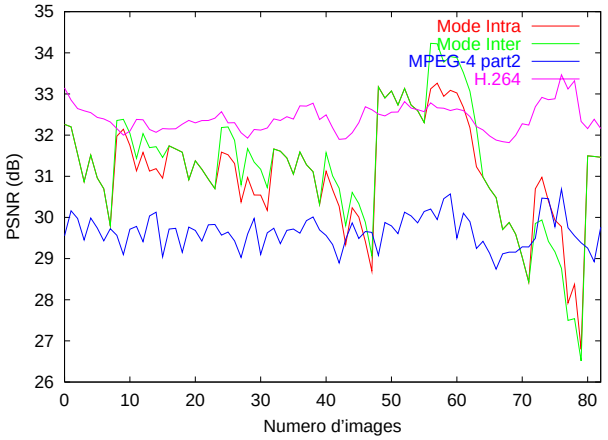

FIG. 4 – *Courbes de PSNR pour (a) Mode Intra (b) Mode Inter à 140 Kbit/s (c) MPEG-4 part 2 à 165 kbits/s (d) H.264 (TML8 version 8.4) à 110kbit/s*

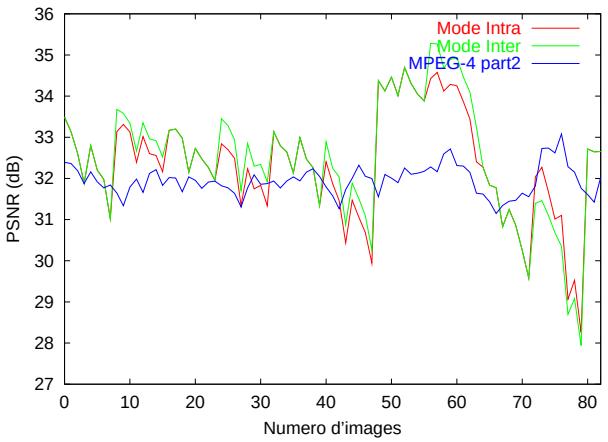

FIG. 5 – *Courbes de PSNR pour (a) Mode Intra (b) Mode Inter à 200 Kbit/s (c) MPEG-4 à 225 kbits/s*

8 Conclusion et perspectives

Dans ce papier nous avons décrit un approche de codage vidéo bas débit et finement scalable basée sur une décomposition ondelettes 2D+t. Les performances obtenues sont nettement supérieures à celles de MPEG-4 et sont assez proches de celles de H.264. pour une complexité moindre. L'algorithme proposé est, en contradiction avec MPEG-4 et H.264, hautement scalable et, par conséquent, particulièrement bien adapté à la transmision sur réseau de type Internet. Plusieurs pistes sont envisagées afin d'améliorer les performances du schéma obtenu. Ainsi, l'utilisation d'un codeur arithmétique basé contexte dans l'esprit de CABAC utilisé dans H.264, est envisagée. De plus, l'utilisation de GOF de taille variable semble être une bonne alternative. Enfin, l'ajout de contextes prenant mieux en compte la dimension temporelle dans EBCOT est également à l'étude.

Références

[1] D. Taubman. High performance scalable image compression with EBCOT. *IEEE Trans. on Image Proc.*, 9(7):1158–1170, Jul. 2000.

[2] G. Karlsson and M. Vetterli. Three-dimensional subband coding of video. In *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pages 1100–1103, April 1988.

[3] D. Taubmann and A. Zakhor. Multirate-3d subband coding of video. *IEEE Trans. on Image Processing*, 3(5):572–588, September 1994.

[4] A. Wang, Z. Xiong, P.A. Chou, and S. Mehrotra. Three-dimensional wavelet coding of video with global motion compensation. In *Proc. IEEE Intl. Conf. on Data Compression*, pages 404–413, March 1999.

[5] J-R. Ohm. Three-dimensional subband coding with motion compensation. *IEEE Trans. on Image Processing*, 3(5), November 1994.

[6] S-J.Choi and J.Woods. Motion-Compensated 3-D Subband Coding of Video. *IEEE Trans. on Image Processing*, 8(2):155–167, February 1999.

[7] A. Secker and D. Taubman. Motion-compensated highly scalable video compression using adaptive 3d wavelet transform based on lifting. In *Proc. IEEE Intl. Conf. on Image Processing*, Oct. 2001.

[8] L. Luo, J. Li, S. Li, Z Zhuang, and Y-Q. Zhang. Motion compensated lifting wavelet and its application in video codi ng. In *Proc. IEEE Intl. Conf. on Image Processing*, Oct. 2001.

[9] K.Z. Xiong and W.A. Pearlman. Low bit-rate scalable video coding with 3D set partitioning in hierachical trees (3d-spiht). *IEEE Trans. on Circuits and Systems for Video Technology*, 10(8):1374–1387, December 2000.

[10] V. Bottreau, M. Bénetière, B. Felts, and B. Pesquet-Popescu. A fully scalable 3D subband video codec. In *Proc. of IEEE Int. Conf. on Image Proc., ICIP 01*, Oct. 7-10 2001.

[11] Shih-Ta Hsiang and John W. Woods. Embedded video coding using motion compensated 3-d subband/wavelet filter bank. In *Proc. of Packet Video Workshop, PV 00*, May 2000.

[12] J. Viéron, C. Guillemot, and S. Pateux. Motion compensated 2d+t wavelet analysis for low rate fgs video compression. In *Proc. of the International Thyrrhenian workshop on digital communications, IWDC 02*, September 2002.