

TWAVIX : une solution basée-ondelettes (t+2D) pour le codage vidéo scalable

Guillaume Boisson¹

Édouard François¹

Christine Guillemot²

¹ THOMSON R&D France, 1 avenue de Bellefontaine
CS 17616 - 35576 Cesson-Sévigné, France

{guillaume.boisson, edouard.francois}@thomson.net

² IRISA, Campus Universitaire de Beaulieu, 35042 Rennes Cédex, France

christine.guillemot@irisa.fr

Résumé

Nous proposons ici une brève description du schéma de codage vidéo scalable basé-ondelettes TWAVIX (Thomson WAvelet-based Video scalable Coding Scheme). TWAVIX présente la particularité d'utiliser la norme JPEG-2000 pour le codage de la texture, bénéficiant ainsi d'importantes fonctionnalités de scalabilité. Pour tirer parti de la redondance temporelle, TWAVIX dispose d'une gamme étendue et flexible d'outils de filtrage compensé en mouvement basés-lifting, permettant de répondre à toute contrainte de délai en garantissant une compaction optimale de l'information sur l'axe temporel. TWAVIX a été proposé lors de la compétition MPEG sur la compression vidéo scalable, laquelle a permis de prendre conscience et d'adresser les nouveaux besoins de la chaîne de diffusion de l'image, en particulier en matière de scalabilité en résolution. Deux solutions sont ici présentées, l'une s'appuyant sur une description scalable du mouvement dans le cadre d'une analyse (t+2D), et l'autre reposant sur une approche multi-résolution emboîtée, respectant in-fine le principe d'échantillonnage critique.

Mots clefs

Compression vidéo, codage scalable, MPEG, ondelettes, JPEG-2000, filtrage temporel, scalabilité spatiale.

1 Introduction

On assiste à l'heure actuelle à un développement sans précédent des infrastructures de télécommunication et à l'explosion des applications des réseaux multimédias. Face aux nouvelles problématiques du monde de l'image transmise (communication point-à-point, diffusion vidéo sur des réseaux hétérogènes en débits, nécessité de partager une information visuelle sur un parc hétéroclite de terminaux – téléviseurs SD et HD, PDA, téléphones portables –, etc.), la compression vidéo scalable représente une solution séduisante et aujourd'hui mature.

Les technologies actuelles de compression vidéo scalable

sont pour une grande partie héritières des travaux réalisés autour de la transformée en ondelettes dans le domaine du codage image par Mallat, Daubechies et Barlaud [1, 2], puis dans la dimension temporelle par Taubman & Zakhor, Ohm, Woods et Pesquet-Popescu [3, 4, 5, 6, 7].

Dans cette même lignée, TWAVIX reprend le principe d'une analyse temporelle par filtrage compensé en mouvement, suivie d'une analyse spatiale par transformée en ondelettes et d'un encodage entropique à granularité élevée. Ces caractéristiques, explicitées en section 2, permettent à TWAVIX de délivrer un flux vidéo scalable en débit, en résolution et en cadence, tout en exhibant des performances de compression proches de l'état de l'art du codage non-scalable. Cette propriété a fait de TWAVIX un candidat potentiel à la compétition MPEG sur les technologies de compression vidéo scalables [8], compétition qui consistait à évaluer la qualité visuelle dans différentes configurations de décodage (de 6Mbps à 64Kbps, sur trois échelles de résolution) à partir de l'encodage d'un contenu HD. Des résultats comparatifs sont présentés en section 3 pour illustrer les performances de TWAVIX. Enfin dans la section 4 nous proposons au lecteur une nouvelle configuration du codec, inspirée pour une part des techniques concurrentes proposées au *Call for Proposals* de MPEG. Ce schéma de codage repose sur l'utilisation d'une approche multi-résolution pour l'analyse temporelle, mais ne présente pas de redondance spatiale. En effet le principe d'échantillonnage critique est respecté grâce à l'emboîtement de la pyramide laplacienne.

2 Architecture du schéma de codage

Nous récapitulons ici le fonctionnement du schéma de codage TWAVIX, déjà présenté dans [9, 10]. Il s'agit d'un codec basé-ondelettes (t+2D), ce qui signifie qu'à l'encodage, l'analyse en sous-bandes temporelles est réalisée avant l'analyse en sous-bandes spatiales (Cf. Fig. 1). En ce qui concerne la transformée spatiale en ondelettes, le codage entropique des sous-bandes de texture et l'ordonnement du train binaire (optimal au sens débit-distorsion),

nous utilisons les fonctionnalités de la librairie JPEG-2000 VM8.0 .

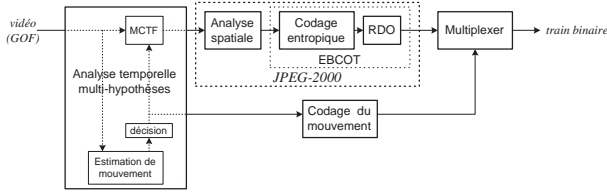


Figure 1 – Architecture de l'encodeur TWAVIX.

2.1 Analyse temporelle

L'efficacité du schéma de codage TWAVIX repose sur une analyse temporelle multi-hypothèses adaptée localement au contenu de la vidéo.

Estimation de mouvement. Successivement à tous les niveaux temporels, une estimation de mouvement bi-directionnelle est effectuée à partir de chaque image impaire vers les images paires précédente et suivante. Chaque estimation est effectuée simultanément sur plusieurs grilles, chacune correspondant à une taille de bloc (de 8×8 à 256×256), avec une précision d'un huitième de pixel. Sur chacune de ces grilles, en fonction des candidats obtenus et selon la latitude laissée par l'utilisateur, le meilleur mode de filtrage est retenu (prédiction *forward*, *backward*, bi-directionnelle ou codage intra). Puis on parcourt la structure d'arbre ("*quad-tree*") formée par l'ensemble des grilles, en sélectionnant les noeuds qui minimisent un critère débit-distorsion (Cf. Fig. 2). La structure élaguée alors obtenue peut être considérée comme la description optimale du déplacement.

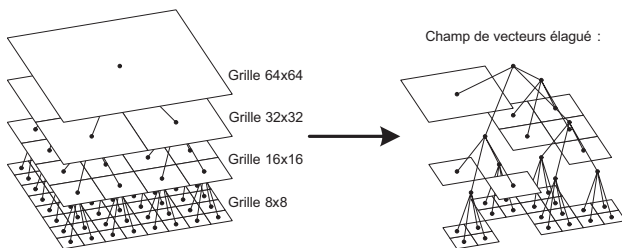


Figure 2 – Élagage d'un champ de mouvement multi-grille.

Filtrage temporel. Le filtrage temporel compensé en mouvement est implémenté selon le schéma *lifting*, c'est-à-dire par combinaison d'un filtrage prédictif (passe-haut) et d'un filtrage de mise-à-jour (passe-bas). Cette factorisation, comparée avec un schéma purement prédictif, permet de se rapprocher de l'idéal de la transformation orthogonale qui optimise l'allocation de débit entre bandes temporelles.

Une fois les décisions prises et les élagages effectués, on procède au filtrage prédictif des images impaires. Puis en fonction des modes de prédiction retenus, on dresse, selon

la méthode préconisée dans [11], les cartes de connexion nécessaires au filtrage de mise-à-jour des images paires, lequel est alors effectué.

Dans le cas le moins contraint, le schéma de filtrage temporel correspond donc à un filtrage 5-3, illustré par les équations de prédiction et de mise-à-jour ci-dessous (dans le cas d'une double connexion dans l'image paire), où l'opérateur \mathcal{C} désigne la compensation en mouvement :

$$h_k = f_{2k+1} - \frac{\mathcal{C}_{2k \rightarrow 2k+1}(f_{2k}) + \mathcal{C}_{2k+1 \leftarrow 2k+2}(f_{2k+2})}{2}$$

$$l_k = f_{2k} + \frac{\mathcal{C}_{2k-1 \rightarrow 2k}(h_{k-1}) + \mathcal{C}_{2k \leftarrow 2k+1}(h_k)}{4}$$

Ce choix de schéma de codage par filtrage 5-3 se révèle coûteux en mémoire et en calculs. A titre indicatif, en négligeant les temps de transmission et d'exécution, une analyse temporelle sur N niveaux temporels induit une latence de reconstruction L de $3 \cdot (2^N - 1)$ images. La figure 3 illustre ce calcul pour le cas simple d'un groupe de 4 images ($L = 9$ lorsque $N = 2$). Les chiffres indiqués correspondent aux instants d'acquisition ou d'obtention des différentes images dans le processus de codage-décodage.

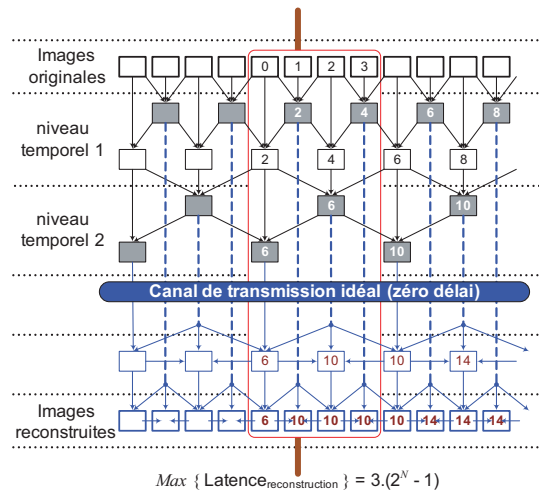


Figure 3 – Latence de reconstruction du filtrage 5-3.

Pour les applications de diffusion, on a classiquement recours à des groupes de 16 images ($N = 4$), ce qui induit des latences de l'ordre de la seconde. Ces valeurs sont malheureusement incompatibles avec les conditions requises par le groupe MPEG en concertation avec les industriels du monde de l'image [12].

C'est la raison pour laquelle TWAVIX a la propriété de pouvoir contraindre son schéma de filtrage de manière adaptative, en particulier en supprimant dans les niveaux temporels supérieurs les mises à jour et les prédictions *backward* en frontière de groupe d'images. Les délais d'encodage et de décodage sont ainsi limités, sans détériorer l'efficacité de compression.

2.2 Codage du mouvement

Pour évaluer les performances d'une technologie de compression scalable ($t+2D$), on peut chercher à la comparer en tous ses points de décodage avec un processus d'encodage-décodage non-scalable dédié. Classiquement, les technologies basées- $(t+2D)$ s'avèrent alors comparativement peu efficaces pour les sous-résolutions spatiales. En effet, dans cette configuration, l'information auxiliaire nécessaire à l'étape de synthèse temporelle est disproportionnée au regard du budget total alloué pour le mouvement et la texture (Cf. Fig. 4).

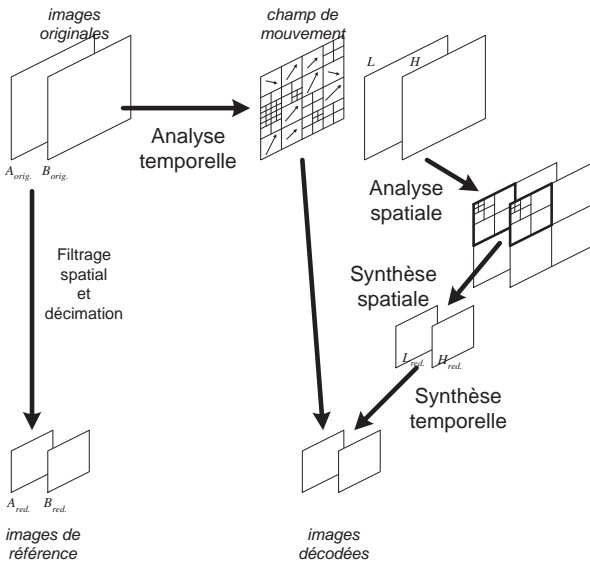


Figure 4 – Scalabilité spatiale dans un schéma $(t+2D)$.

Une solution consiste à partitionner l'information de mouvement en couches successives de précision, adaptées aux différentes résolutions. Les vecteurs ainsi utilisés à la synthèse correspondent au degré d'interpolation adéquat et présentent un débit raisonnable, sans corrompre la structure des déplacements déterminée par l'élagage et utilisée à l'analyse (Cf. Fig. 5 et [10]).

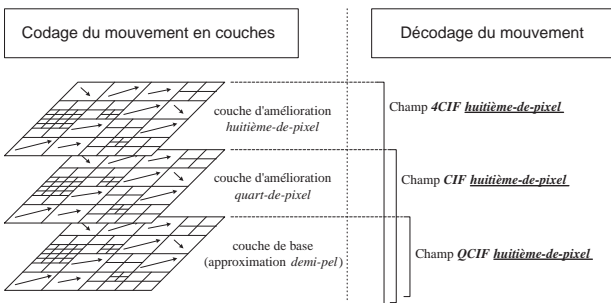


Figure 5 – Codage-décodage scalable en précision.

La figure 6 illustre en terme de PSNR et de qualité visuelle les gains obtenus à une sous-résolution (QCIF 15Hz 64Kbps) grâce au codage de mouvement scalable.

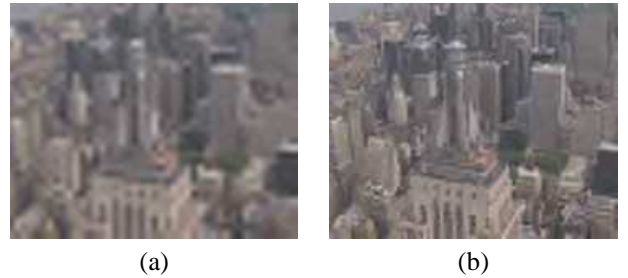
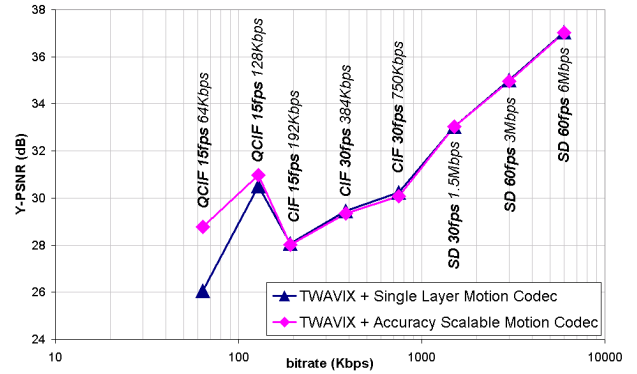


Figure 6 – Séquence CITY : sans (a) et avec (b) codage scalable de mouvement.

3 Résultats du Call for Proposals

Le test principal du *Call for Proposals* (CfP) du groupe MPEG consistait en huit décodages (trois à la résolution 4CIF, trois en CIF et deux en QCIF) à partir d'un unique train binaire, obtenu par encodage d'un contenu 4CIF (Cf. Tab. 1).

Dimension	Cadence	Débit
704×576	60 Hz	6000 Kbps
704×576	60 Hz	3000 Kbps
704×576	30 Hz	1500 Kbps
352×288	30 Hz	750 Kbps
352×288	30 Hz	384 Kbps
352×288	15 Hz	196 Kbps
176×144	15 Hz	128 Kbps
176×144	15 Hz	64 Kbps

Tableau 1 – Décodages évalués pour le CfP : scalabilités spatiale, temporelle, et en débit.

Pour chaque débit-cible, les performances des différents codecs proposés ont été évaluées en terme de perte de qualité subjective par rapport à un encodage-décodage dédié avec le schéma de codage MPEG-4 Part 10 AVC (H264). La description des différentes technologies proposées et l'analyse détaillée des résultats [13] se révèle très instructive. On s'y rend compte que TWAVIX (Cf. Fig. 7 – courbe rouge) fait bonne figure parmi ses concurrents basés- $(t+2D)$ ou -AVC. Mais il apparaît également que certains candidats, qui ne respectent par ailleurs pas le cahier des charges en délivrant un train binaire global non-empoîté (courbes pointillées), présentent d'excellents

résultats, principalement aux basses résolutions.

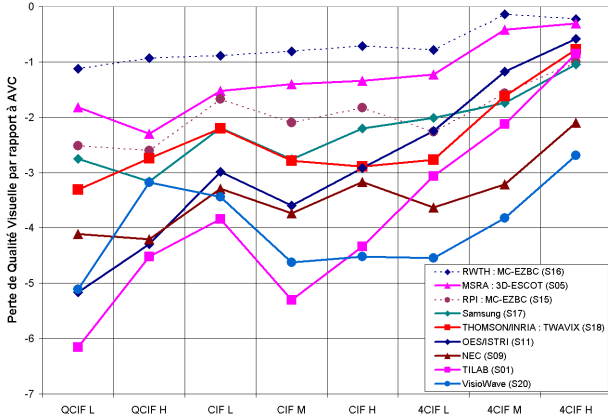


Figure 7 – Résultats finaux des tests subjectifs du CfP.

Cet état de fait, qui préfigure le retour et le succès des techniques pyramidales de *simulcast* dans la compétition MPEG, nous a inspiré de nouvelles évolutions pour le schéma de codage TWAVIX.

4 Codage *simulcast* emboîté

4.1 Principe

Dans la suite, nous noterons respectivement $\mathcal{A} = \begin{pmatrix} \mathcal{A}_{\mathcal{L}} \\ \mathcal{A}_{\mathcal{H}} \end{pmatrix}$ et $\mathcal{S} = (\mathcal{S}_{\mathcal{L}} \ \mathcal{S}_{\mathcal{H}})$ les opérations d'analyse et de synthèse spatiales. S'il est trivial que $\mathcal{S} \circ \mathcal{A} = \mathcal{I}d_{pixel}$, notons que l'on a également $\mathcal{A}_{\mathcal{L}} \circ \mathcal{S}_{\mathcal{L}} = \mathcal{I}d_{BF}$.

Le schéma de codage est le suivant :

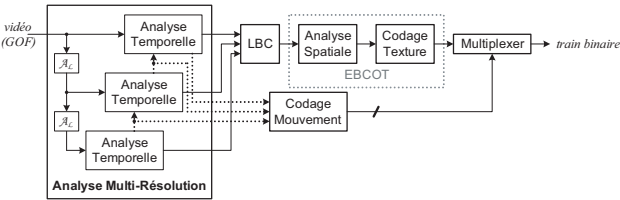


Figure 8 – Architecture TWAVIX de codage *simulcast* emboîté.

Après analyse temporelle multi-résolution, les groupes d'images transformées sont emboîtés avec la technique LBC (*Low-Band Correction*) décrite pour la première fois par Han dans [14].

Analyse temporelle multi-résolution. L'analyse temporelle repose sur une approche purement prédictive. À chaque résolution i , les images impaires sont codées par prédiction – ici simplement *forward* pour alléger les notations – (Cf. Fig. 9) :

$$h_k^i = f_{2k+1}^i - C_{2k \rightarrow 2k+1}^i(f_{2k}^i) \quad (1)$$

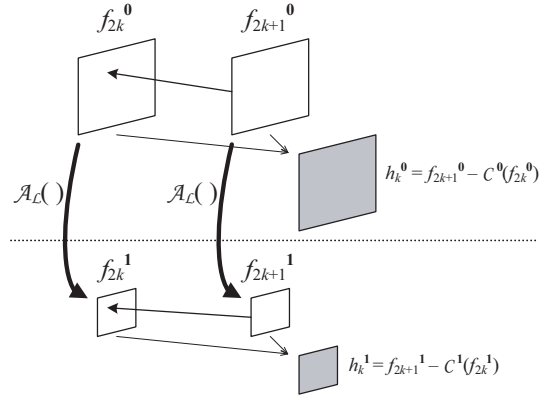


Figure 9 – Analyse multi-résolution.

Emboîtement des résolutions. Pour supprimer la redondance spatiale, on transmet d'une part la résolution originale de l'image de référence (l'image paire, codée intra, sans filtrage de mise-à-jour), qui contient naturellement sa version sous-échantillonnée, et d'autre part un mélange de hautes fréquences temporelles, à savoir l'image h_k^0 dont les basses fréquences spatiales auraient été substituées par h_k^1 :

$$\begin{cases} l_k &= f_{2k}^0 \\ h_k &= h_k^0 - (\mathcal{S}_{\mathcal{L}} \circ \mathcal{A}_{\mathcal{L}})(h_k^0) + \mathcal{S}_{\mathcal{L}}(h_k^1) \end{cases}$$

Notons que les hautes fréquences temporelles transmises peuvent s'écrire sous la forme :

$$h_k = \mathcal{S} \begin{pmatrix} h_k^1 \\ \mathcal{A}_{\mathcal{H}}(h_k^0) \end{pmatrix}$$

Composition spectrale des informations transmises.

En terme de sous-bandes, cela revient à transmettre, en plus de l'image de référence, les hautes fréquences temporelles des basses fréquences spatiales et les hautes fréquences spatiales de l'erreur de prédiction à résolution originale (Cf. Fig. 10).

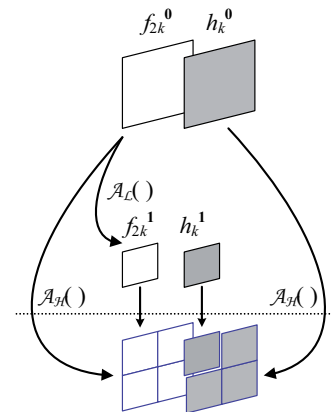


Figure 10 – Emboîtement des résolutions : informations transmises.

Ce mécanisme d'emboîtement d'analyse multi-résolution s'étend trivialement au nombre désiré de niveaux spatiaux.

Décodages. Le décodage à résolution minimale est trivial :

$$\begin{cases} f_{2k}^1 &= \mathcal{A}_{\mathcal{L}}(l_k) \\ f_{2k+1}^1 &= \mathcal{C}_{2k \rightarrow 2k+1}^1(f_{2k}^1) + \mathcal{A}_{\mathcal{L}}(h_k) \end{cases}$$

Le décodage à résolution originale (ou intermédiaire) est un peu plus problématique du fait de l'apparente disparition des basses fréquences spatiales de l'erreur de prédiction lors de l'emboîtement. Cette information peut néanmoins être recouverte à partir de f_{2k+1}^1 et de l'image f_{2k}^0 compensée en mouvement. En effet, d'après l'équation (1) :

$$\begin{aligned} \mathcal{A}_{\mathcal{L}}(h_k^0) &= \mathcal{A}_{\mathcal{L}}(f_{2k+1}^0 - \mathcal{C}_{2k \rightarrow 2k+1}^0(f_{2k}^0)) \\ &= f_{2k+1}^1 - (\mathcal{A}_{\mathcal{L}} \circ \mathcal{C}_{2k \rightarrow 2k+1}^0)(f_{2k}^0) \end{aligned}$$

Les équations du décodage à résolution originale (ou intermédiaire) sont donc les suivantes :

$$\begin{cases} f_{2k}^0 &= l_k \\ f_{2k+1}^0 &= \mathcal{C}_{2k \rightarrow 2k+1}^0(f_{2k}^0) + \\ &+ \mathcal{S} \left(f_{2k+1}^1 - (\mathcal{A}_{\mathcal{L}} \circ \mathcal{C}_{2k \rightarrow 2k+1}^0)(f_{2k}^0) \right) \end{cases}$$

L'ensemble du processus de synthèse multi-résolution est illustré Fig. 11.

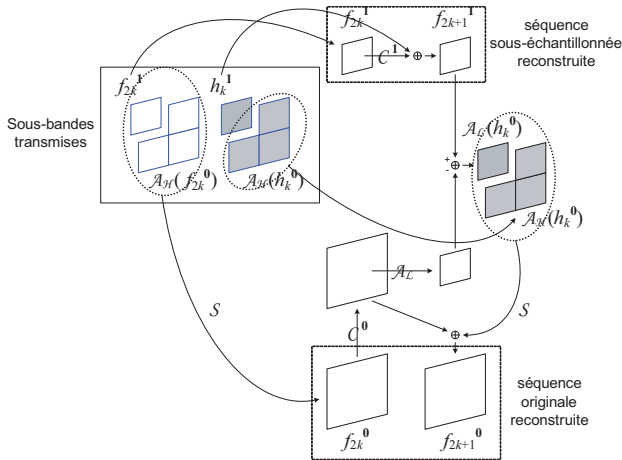


Figure 11 – Synthèse multi-résolution.

Notons enfin qu'au mépris des apparences, qui laisseraient supposer que la reconstruction des basses fréquences spatiales manquantes nécessite le codage intra de f_{2k}^0 , il devient concevable d'introduire une étape de mise-à-jour temporelle dans le schéma LBC grâce aux récents travaux de Mehrseresht & Taubman [15]. Ce prolongement fera l'objet de travaux ultérieurs.

4.2 Résultats

Le codage par *simulcast* emboîté présente la propriété de répartir avantageusement la qualité entre les différents

niveaux spatiaux grâce à l'analyse temporelle multi-résolution. En effet l'information de mouvement utilisée à la synthèse correspond exactement à l'information employée à l'analyse, sans troncation d'aucune sorte, et son contenu a été optimisé directement pour la résolution décodée.

En comparaison avec les codecs (t+2D) de l'état de l'art, cela se traduit par une perte raisonnable de compression à résolution originale et un gain significatif aux autres résolutions (Cf. Fig. 12). De plus, dans certains cas, lorsque les séquences présentent une forte activité, le codage multi-résolution permet de pallier la faiblesse de l'analyse temporelle classique (due à un degré limité de corrélation temporelle) en recourant à une prédiction inter-résolution intra-image. Le codage par *simulcast* emboîté se montre alors supérieur à l'approche classique sur toute la gamme de décodage (Cf. Fig. 13). Soulignons encore une fois que ce dispositif ne se traduit pas par une description globale redondante, comme en présentent la plupart des approches pyramidales.

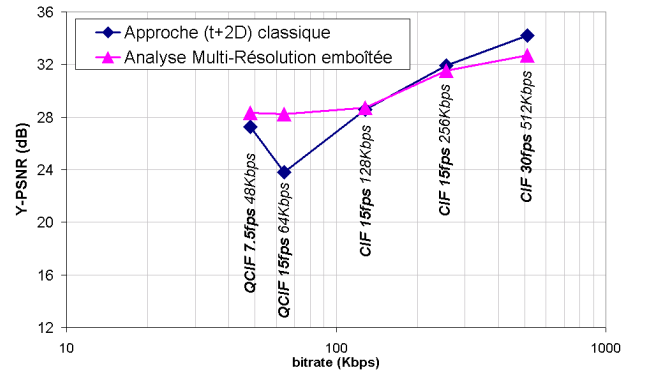


Figure 12 – Séquence FOREMAN.

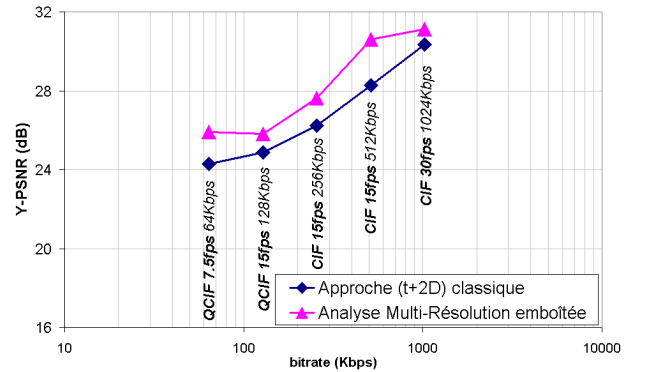


Figure 13 – Séquence FOOTBALL.

Enfin, la figure 14 illustre le gain de qualité visuelle obtenu pour un décodage à une sous-résolution (QCIF 15Hz 64Kbps).



(a) Approche (t+2D) classique



(b) avec Analyse Multi-résolution emboîtée

Figure 14 – Séquence BUS : Comparaison de la qualité visuelle obtenue à faible résolution et bas débit.

5 Conclusion

Nous avons présenté le schéma de codage vidéo scalable basé-ondelettes TWAVIX dans deux configurations. La première, classique et appartenant à la famille des solutions (t+2D), a contribué à montrer lors de la compétition MPEG qu'il était possible de délivrer un flux vidéo scalable couvrant trois résolutions spatiales, de 64Kbps à 6Mbps, sans grande pénalité à résolution originale par rapport à un codec non-scalable. La seconde, reposant sur une approche *simulcast* emboîtée, c'est-à-dire sans redondance, en respectant le principe d'échantillonnage critique, permet de répartir d'une manière plus équitable la qualité entre les différents niveaux spatiaux et se montre très efficace dans le cas des séquences à fort mouvement.

Références

- [1] S. Mallat. Multiresolution approximation and wavelet orthonormal basis. *Transactions of the American Mathematical Society*, 315 :69–87, Septembre 1989.
- [2] M. Antonini, P. Barlaud, et I. Daubechies. Image coding using wavelet transform. *IEEE Transactions on Image Processing*, 1(2) :205–220, Avril 1992.
- [3] D. Taubman et A. Zakhor. Multirate-3D subband coding of video. *IEEE Transactions on Image Processing*, 3(5) :572–588, Septembre 1994.
- [4] J.-R. Ohm. Three-dimensional subband coding with motion compensation. *IEEE Transactions on Image Processing*, 3(5) :559–571, Septembre 1994.
- [5] S.J Choi et J.W. Woods. Motion-Compensated 3-D Subband coding of video. *IEEE Transactions on Image Processing*, 8(2) :155–167, Février 1999.
- [6] S.-T. Hsiang et J.W. Woods. Invertible three-dimensional analysis-synthesis system for video coding with half-pixel-accurate motion compensation. Dans *SPIE Conference on Visual Communication and Image Processing*, pages 537–546, San Jose, California, 1999.
- [7] B. Pesquet-Popescu et V. Bottreau. Three-dimensional lifting schemes for motion compensated video compression. Dans *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'01*, Salt Lake City, Utah, Mai 2001.
- [8] Call for Proposals on Scalable Video Coding Technology. Dans *ISO/IEC JTC1/SC29/WG11 MPEG2003/N6193*, Waikoloa, Hawaii, Décembre 2003.
- [9] Report on Call for Evidence on Scalable Video Coding Advances. Dans *ISO/IEC JTC1/SC29/WG11 MPEG2003/N5701*, Trondheim, Norway, Juillet 2003.
- [10] G. Boisson, E. François, et C. Guillemot. Accuracy scalable motion coding for efficient scalable video compression. Dans *Proceedings of 11th IEEE International Conference on Image Processing, ICIP'2004*, Singapore, Octobre 2004.
- [11] Y. Zhan, M. Picard, B. Pesquet-Popescu, et H. Heijmans. Long temporal filtering in lifting schemes for scalable video coding. Dans *ISO/IEC JTC1/SC29/WG11 MPEG2002/M8680*, Klagenfurt, Germany, Juillet 2002.
- [12] Applications and Requirements for Scalable Video Coding. Dans *ISO/IEC JTC1/SC29/WG11 MPEG2004/N6830*, Palma de Mallorca, Spain, Octobre 2004.
- [13] V. Baroncini et T. Oelbaum. Subjective test results for the CfP on Scalable Video Coding Technology. Dans *ISO/IEC JTC1/SC29/WG11 MPEG2004/M10737*, Muenchen, Germany, Mars 2004.
- [14] W.-J. Han. Responses of Call-for-Proposal for Scalable Video Coding. Dans *ISO/IEC JTC1/SC29/WG11 MPEG2004/M10569/S17*, Muenchen, Germany, Mars 2004.
- [15] N. Mehrseresht et D. Taubman. Spatial scalability and compression efficiency within a flexible motion compensated 3D-DWT. Dans *Proceedings of 11th IEEE International Conference on Image Processing, ICIP'2004*, Singapore, Octobre 2004.