THREE-DIMENSIONAL LIFTING SCHEMES FOR MOTION COMPENSATED VIDEO COMPRESSION

Béatrice Pesquet-Popescu *

ENST, Dept. Signal and Im. Proc.
46, rue Barrault, 75634 Paris, FRANCE
e-mail : pesquet@tsi.enst.fr

Vincent Bottreau

Laboratoires d'Electronique Philips,
22, Avenue Descartes,
94453 Limeil-Brévannes, FRANCE
e-mail : bottreau@philips.com

ABSTRACT

Three dimensional wavelet decompositions are efficient tools for scalable video coding. In this paper, we show the interest of a lifting formulation of these decompositions. The temporal wavelet transform is inherently non-linear, due to the motion estimation step, and the lifting formalism allows us to provide several improvements to the scheme initially proposed by Choi and Woods: a better processing of the uncovered areas is proposed and an overlapped motion compensated temporal filtering method is introduced in the multiresolution decomposition. As shown by simulations, the proposed method results in higher coding efficiency, while keeping the scalability functionalities.

1. INTRODUCTION

With the recent expansion of multimedia applications and the need for delivering compressed bitstreams over heterogeneous networks, scalability has become an important feature for video coders. The 3D wavelet decomposition provides a natural spatial resolution and frame rate scalability. Embedded coding algorithms (like the 3D SPIHT algorithm [4]), lead to the desired SNR scalability, by using in-depth scanning of the coefficients in hierarchical trees and bitplane encoding.

Our global codec scheme includes a temporal multiresolution analysis (MRA) in the direction of the motion in order to take into account large displacements (more precisely, Haar filters are applied at each resolution level on the motion compensated frames). It is followed by a spatial MRA of the resulting temporal subbands. The coding method is a modified 3D SPIHT algorithm followed by a contextual arithmetic coder [6]. In this way, scalability is achieved in temporal and spatial resolutions, as well as in quality.

*This work was performed while the first author was at the head of the "Scalable Subband Video Coding" project at LEP.

While several works addressed scalable coding algorithms and the possible improvements of the spatial wavelet transforms, there was little effort in the direction of exploiting the particularities of the temporal wavelet transform used in 3D decompositions. The temporal wavelet transform is inherently non-linear, due to the motion estimation step. The lifting scheme provides a flexible framework for building wavelet transforms. Its interest for processing monodimensional signals and for providing suitable decompositions for image coding or feature extraction has already been proven. The advantages of this scheme are both in terms of complexity ("in-place" calculation) and additional functionalities: not only every linear wavelet decomposition can be put in this form [2], but it allows the introduction of non-linear operators in the predict-update steps as well [3]. In this paper, we investigate the benefits of the non-linear lifting formalism for motion compensated wavelet decompositions. In particular, we propose an improved method for processing uncovered areas and we introduce an overlapped motion compensation within the temporal decomposition.

The paper is organized as follows: in Section 2 we describe the problems related to motion compensated temporal analysis for 3D video compression. Section 3 introduces the lifting scheme in the context of temporal wavelet decompositions. In Section 4 and 5 we provide several non-linear schemes adapted to our needs and in Section 6 the simulation results are presented. The last section concludes the paper.

2. MOTION COMPENSATED TEMPORAL ANALYSIS FOR 3D VIDEO COMPRESSION

An important issue concerning temporal multiresolution analysis is the choice of the temporal filter length : long filters take better advantage of the temporal correlation existing between successive frames. It was shown in [1] that they do, however, blur the motion, increase buffer memory required

at the decoder (corresponding to the filter length) and the reconstruction delay, which might not be desirable in real-time applications. Moreover, the coding efficiency is not increased significantly by performing a temporal analysis with longer filters. Therefore, in our approach, Haar filters were used for temporal filtering. Moreover, when Haar filters are used for the temporal decomposition, motion estimation and motion compensation (ME/MC) are only performed every two frames of the input sequence due to the temporal downsampling by two. By iterating this procedure over several decomposition levels on the approximation subband, the total number of ME/MC operations is roughly the same as in a predictive scheme.

Motion compensated temporal filtering raises the problem of double-connected/unconnected pixels, i.e., pixels that are filtered several times or not at all. The solution proposed by Choi [5] is to compute a high-pass coefficient at the same location as the pixel in the current frame, and to take as a low-pass coefficient the colocated pixel in the motion-compensated reference frame. For multiple-connected pixels, he proposes to scan the current frame from top to down, from left to right and to consider for computation of the low-pass coefficient the first pixel in the current frame pointing to it. We will show that this is not the best strategy and propose several improvements.

At a given resolution level, let us denote by $H[m,n]$ the pixels in the temporal high frequency subband, by $L[m,n]$ the pixels in the low frequency subband and by $(u_{m,n}, v_{m,n})$ the two components of the motion vector associated to the pixel (m,n). If fractional pel motion estimation is allowed, then the integer part of the motion vector will be denoted by $(\bar{u}_{m,n}, \bar{v}_{m,n})$. Therefore, in Choi's method, the equations that allow to compute the high and low-pass subbands for connected pixels are the following ones:

$$H[m,n] = \left(B[m,n] - \widetilde{A}[m-u_{m,n}, n-v_{m,n}]\right)/\sqrt{2} \tag{1}$$

$$\begin{aligned} L[m-\bar{u}_{m,n}, n-\bar{v}_{m,n}] = \\ \Big(\widetilde{B}[m-\bar{u}_{m,n}+u_{m,n}, n-\bar{v}_{m,n}+v_{m,n}] \\ +A[m-\bar{u}_{m,n}, n-\bar{v}_{m,n}]\Big)/\sqrt{2} \end{aligned} \tag{2}$$

where $\widetilde{X}$ stands for an interpolated value of the field X. For unconnected pixels, the high frequency component is obtained as before and the low frequency values are simply scaled values of the reference pixels:

$$L[m,n] = \sqrt{2}A[m,n]. \tag{3}$$

3. NONLINEAR LIFTING FORMULATION OF THE TEMPORAL HAAR ANALYSIS

As a particular case of the lifting scheme [2] for the Haar transform, we can now write the temporal low-pass filtering in the motion direction as:

$$\begin{aligned} L[m-\bar{u}_{m,n}, n-\bar{v}_{m,n}] = \\ \widetilde{H}[m-\bar{u}_{m,n}+u_{m,n}, n-\bar{v}_{m,n}+v_{m,n}] \\ +\sqrt{2}\,A[m-\bar{u}_{m,n}, n-\bar{v}_{m,n}], \end{aligned} \tag{4}$$

This equation, together with (1), allow us to deduce the form of the non-linear operators $\mathcal{P}$ (predict) and $\mathcal{U}$ (update) used in the temporal Haar lifting. Indeed, we can see that $\mathcal{P}$ is a motion compensation operator ($\mathcal{C}$), followed, in case of a fractional pel motion estimation, by an interpolation ($\mathcal{I}$). In the meantime, $\mathcal{U}$ can be identified as a motion compensation operator, using the same motion vectors as in $\mathcal{P}$, but with opposite sign, followed by an interpolation. In the sequel, we will denote these operations by: $\mathcal{P}\{\cdot\} = \mathcal{I}\{\mathcal{C}\{\cdot\}\}$ and $\mathcal{U}\{\cdot\} = \mathcal{I}\{\bar{\mathcal{C}}\{\cdot\}\}$ and the position $(m-\bar{u}_{m,n}, n-\bar{v}_{m,n})$ by (p,q). With these notations, the temporal analysis of connected pixels can be written as:

$$H[m,n] = \frac{1}{\sqrt{2}}\big(B[m,n] - \mathcal{I}\{\mathcal{C}\{A[m,n]\}\}\big) \tag{5}$$

$$L[p,q] = \mathcal{I}\{\bar{\mathcal{C}}\{H[p,q]\}\} + \sqrt{2}\,A[p,q], \tag{6}$$

while for the synthesis part, we have

$$A[p,q] = \frac{1}{\sqrt{2}}\big(L[p,q] - \mathcal{I}\{\bar{\mathcal{C}}\{H[p,q]\}\}\big) \tag{7}$$

for connected pixels (unconnected pixels in the reference frame are obtained directly from Eq. (3)) and

$$B[m,n] = \sqrt{2}\,H[m,n] + \mathcal{I}\{\mathcal{C}\{A[m,n]\}\}. \tag{8}$$

In the next sections, we will show the interest of this formulation in improving the uncovered zone processing and in reducing the blocking artefacts related to block-based motion compensation.

4. IMPROVEMENT OF THE UNCOVERED ZONE PROCESSING

The problem of unconnected and double-connected pixels is closely related to that of areas uncovered by moving objects. Indeed, consider two objects corresponding to a common part in a frame at time T, and that become separate at time $T+\Delta T$. In this case, two regions in the current frame will correspond by motion compensation to the same region in the reference (previous) frame. For one of the objects, this will be an uncovered area. In our analysis, this area will appear as doubly connected in the reference frame. The approach in [5] associates to these pixels the first block encountered in the motion compensation process. We propose to optimize this choice, by applying some criteria based on the lifting scheme. The main structural property we exploit

is that we can use for the update step (computation of the temporal low frequency subband) all the information available from the predict step (high frequency subband) and causal information in the low frequency subband.

The intuition behind the first criterion we propose is related to the energy of the detail subband of the two moving objects. If the first object was on the foreground at time T, the uncovered region in the second object will give rise to a higher energy of the detail coefficients. The second criterion is a condition of regularization of the motion field: if several pixels are connected to the same pixel in the reference frame, the one with the smallest displacement will be chosen for filtering.

Let us now formalize these ideas within the above nonlinear lifting framework. In the case of multiple-connected pixels in the reference frame, let us consider one of them at the position (p, q) and two pixels found by the motion estimation algorithm at the positions (m_1, n_1) and (m_2, n_2) in the current frame. If the two correponding motion vectors are $(u_{m_1,n_1}, v_{m_1,n_1})$ and $(u_{m_2,n_2}, v_{m_2,n_2})$, we have: $m_1 - \bar{u}_{m_1,n_1} = m_2 - \bar{u}_{m_2,n_2} = p$, $n_1 - \bar{v}_{m_1,n_1} = n_2 - \bar{v}_{m_2,n_2} = q$. Using this observation, equations (5) and (6) can be written for each of the two pixels (m_1, n_1) and (m_2, n_2), yielding two different values $H[m_1, n_1]$ and $H[m_2, n_2]$ in the detail subband. Consequently, the value in the approximation subband can be computed using either of these two values. Note that both values allow perfect reconstruction. Actually, if we denote by $P_{(p,q)}$ the set of all pixels (m, n) in the current frame connected to the pixel (p, q) in the reference frame, we can remark that the perfect reconstruction property is guaranteed for any operator f such that

$$L[p,q] = f\Big(\widetilde{H}\left[m - \bar{u}_{m,n} + u_{m,n}, n - \bar{v}_{m,n} + v_{m,n}\right], \\ (m,n) \in P_{(p,q)}\Big) + \sqrt{2}\, A[p,q].$$

One criterion for the choice of the operator f is to minimize the energy of the detail subband so as to associate $A[p, q]$ to its "closest" value in frame B. This implies using for the low-pass filtering the pixel (m_0, n_0) such that

$$\left|\mathcal{I}\left\{\bar{\mathcal{C}}\left\{H[m_0 - \bar{u}_{m_0,n_0}, n_0 - \bar{v}_{m_0,n_0}]\right\}\right\}\right| \\ = \min_{(m,n)\in P_{(p,q)}} \left|\mathcal{I}\left\{\bar{\mathcal{C}}\left\{H[m - \bar{u}_{m,n}, n - \bar{v}_{m,n}]\right\}\right\}\right|.$$

As for Choi's algorithm, in the proposed algorithm it should not be necessary to transmit the classification map (saying which pixels are connected and which ones are not) to the decoder. Since the decoder follows the symmetric procedure to that of the encoder, there will be *ideally* the same classification map resulting from decisions made on the energy of the high frequency coefficients. However, the decision based on the value of a single pixel is not robust enough. In particular, in the previous example, the two values in the high frequency subband may not be quantized with identical quantization steps, due to the progressive quantization strategy used in the SPIHT algorithm. So, this could lead to an erroneous decision. The above decision can be made more robust by comparing the mean energy of the displaced frame difference (DFD) around the considered pixel, i.e., $e(p,q)^2 = \sum_{(k,l)\in S(p,q)} \left(H(p-k, q-l)u(k,l)\right)^2$, where $S(p, q)$ is a neighborhood around the pixel (p, q) and $u(k, l)$ corresponds to a weighting factor for each pixel in the neighborhood $S(p, q)$, depending on its distance to the central point. For example, we can choose $u(k,l) = \alpha^{-(|k|+|l|)}$, where $\alpha > 0$ is a forgetting factor.

The second term in the minimization criterion is the norm of the motion vector, $\|\mathbf{d}_{m,n}\| = \left(u_{m,n}^2 + v_{m,n}^2\right)^{1/2}$. The regularized criterion can be expressed as $J(p,q) = e(p,q)^2 + \lambda\|\mathbf{d}_{m,n}\|$, λ being a regularization parameter. If the motion vector is too large, its value is not very reliable, so we can choose not to take it into account for the optimization. This yields the following criterion:

$$J(p,q) = \begin{cases} e(p,q)^2 + \lambda\|\mathbf{d}_{m,n}\|, & \text{if} \quad \|\mathbf{d}_{m,n}\| \le s, \\ e(p,q)^2 + \lambda s, & \text{if} \quad \|\mathbf{d}_{m,n}\| > s, \end{cases}$$

where s is a threshold to be determined empirically.

5. OVERLAPPED MOTION COMPENSATED TEMPORAL DECOMPOSITION

Block-based motion estimation (ME) algorithms suffer from blocking artefacts. If the spatial transform is a wavelet analysis, these artefacts lead to undesired large wavelet coefficients and consequently to a reduction of the coding efficiency. Another improvement that can be deduced from the previous nonlinear lifting formulation is related to the possibility of introducing an overlapped motion compensation within the temporal filtering algorithm, so as to reduce blocking artefacts. This operation involves using in the predict step an average of pixels from adjacent windows in the reference frame.

For example, let us consider an overlap of one pixel. In this case, the high-pass filtering of pixels belonging to the first (respectively, the last) row of a block reads:

$$H[m,n] = \frac{1}{\sqrt{2}}\Big[B[m,n] - \Big((1-\beta)\widetilde{A}[m - u_{m,n}, n - v_{m,n}] \\ +\beta\widetilde{A}[m - 1 - u_{m-1,n}, n - v_{m-1,n}]\Big)\Big],$$

$$H[m,n] = \frac{1}{\sqrt{2}}\Big[B[m,n] - \Big((1-\beta)\widetilde{A}[m - u_{m,n}, n - v_{m,n}] \\ +\beta\widetilde{A}[m + 1 - u_{m+1,n}, n + 1 - v_{m+1,n}]\Big)\Big]$$

where β is a constant, $0 < \beta < 1$. A similar processing is applied to the first (resp., last) column of each block.

Fig. 1. Zoom in the reconstructed frame no. 48 from "Hall Monitor" sequence, original method for temporal filtering.

Fig. 2. Zoom in the reconstructed frame no. 48 from "Hall Monitor" sequence, proposed method for temporal filtering.

6. RESULTS

In all our simulations, a full search Block Matching algorithm has been used for motion estimation. The original video sequences are in QCIF, at 10 fps. For the averaging function $u(k,l)$, different choices of the constant α and of the neighborhood $S(p,q)$ have been tested. A good trade-off between complexity and performances has been obtained for a neighborhood of 8 pixels and $\alpha = 0.5$. Experimental results have shown that a threshold $s = 20$ leads to the best results. The overlap between adjacent windows for the motion compensation is here of one pixel (blocks of size 8×8) with $\beta = 0.2$. Fig. 3 compares the PSNR between the original method and the method including the proposed modifications. Note that our method leads mainly to improvements for the frames where large displacements arise. Fig. 1 and 2 show the improvement in visual quality resulting from the use of the proposed method at 30 kbs.

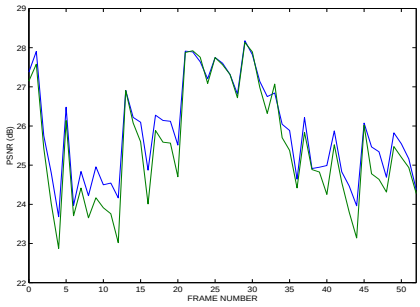

Fig. 3. Comparison of PSNR values for "Carphone" sequence at 30 kbs, 10 fps, frames from 60 to 112: lower curve, the original method; upper curve, the proposed method.

7. CONCLUSION

In this paper, we have introduced a non-linear lifting framework for the temporal wavelet decomposition used in 3D subband video coding. This new formulation allowed us to provide several improvements to the classical motion compensated Haar temporal filtering. Even though the example in this paper was the Haar transform, the extension to more complex non-linear lifting schemes for motion compensated subband video compression is possible and various solutions are currently under investigation.

8. REFERENCES

[1] J. R. Ohm, "Three-dimensional subband coding with motion compensation", IEEE. Trans. on Image Processing, vol. 3, no. 5, pp. 559-571, 1994.

[2] I. Daubechies and W. Sweldens, "Factoring Wavelet Transforms into Lifting Steps", Journal of Fourier Anal. Appl., vol. 4, no. 3, pp. 247-269, 1998.

[3] J. Goutsias and H. Heijmans, "An Axiomatic Approach to Multiresolution Signal Decomposition", Proc. of IEEE Int. Conf. on Image Proc., Chicago, Illinois, Oct. 4-7, 1998.

[4] B.-J. Kim and W. A. Pearlman, "An Embedded Wavelet Video Coder Using Three-Dimensional Set Partitioning in Hierarchical Trees", Proc. DCC'97, pp. 251-260.

[5] S. J. Choi and J. W. Woods, "Motion-compensated 3D subband coding of video", IEEE Trans. on Image Processing, vol. 8, no. 2, pp. 155-164, Feb. 1999.

[6] B. Felts and B. Pesquet-Popescu, "Efficient Context Modeling in Scalable 3D Wavelet-Based Video Compression", Proc. of IEEE Int. Conf. on Image Proc., Vancouver, Canada, Sept. 10-13, 2000.