HIGHLY SCALABLE VIDEO CODING BY BIDIRECTIONAL PREDICT-UPDATE 3-BAND SCHEMES

Christophe Tillier, Béatrice Pesquet-Popescu

Télécom Paris
Signal and Image Proc. Dept.
46, rue Barrault, 75634 Paris, FRANCE
e-mail : {tillier, pesquet}@tsi.enst.fr

Mihaela van der Schaar

Univ. of California Davis
Dept. of Elect. and Computer Eng.
One Shields Avenue, 3129 Kemper Hall
Davis, CA 95616-5294

ABSTRACT

Motion-compensated temporal filtering is an essential ingredient of recently developed wavelet-based scalable video coding schemes. Lifting implementation of these decompositions represents a versatile tool for spatio-temporal optimizations and numerous improvements have thus been proposed. An additional feature has been lately introduced by providing a 3-band lifting temporal decomposition, allowing non-dyadic scalability factors. This paper presents a motion-compensated scheme which is the 3-band equivalent of the 5/3 temporal filterbank. The proposed structure does not enter the classical lifting framework and therefore its invertibility is studied. Simulation results show higher coding efficiency, improved visual quality and more flexibility compared to the previously introduced Haar-like 2 or 3-band schemes.

1. INTRODUCTION

Recent wavelet-based video coding schemes provide high coding efficiency and full temporal, spatial, and quality scalability, which are key factors enabling multimedia applications over heterogeneous networks. They are based on open-loop motion-compensated (MC) temporal multiresolution decompositions [1, 2, 3], followed by spatial wavelet transforms and performant context-based entropy coding of the coefficients [4, 5]. Lifting implementation of these decompositions [6, 7] represents a powerful tool for spatio-temporal optimizations and numerous improvements based on this idea have been recently proposed, concerning for example weighted spatial filtering in occluded areas [8] or the increase of motion estimation accuracy [9]. Unconstrained prediction using multiple reference frames is another interesting technique that has been proposed in [10], but in this case the scheme does not involve an update step.

An additional feature has been introduced by providing a 3-band temporal decomposition, allowing non-dyadic scalability factors [11, 12, 13, 14]. However, general *non-linear* M-band lifting schemes with perfect reconstruction have been introduced in [15] and lifting implementations of M-channel filterbanks designed [16, 17].

In this paper, we extend the simple operators introduced in [11] to more flexible long-term motion-compensated temporal filters and analyse the invertibility of this structure. If the first scheme can be considered as a 3-band equivalent of the Haar MC wavelet decomposition, this new scheme is the 3-band equivalent of the 5/3 MC temporal filterbank [18]. It involves bidirectional MC in both the predict and update operators. Simulation results show improved coding efficiency compared with dyadic multiresolution analysis, as well as with the 3-band Haar-like scheme.

The paper is organized as follows: in the next section we present the framework of 3-band MCTF. In Section 3 we introduce the new structure in the absence of motion estimation, in Section 4 we analyse the temporal operators involving motion compensation and in Section 5 we derive the synthesis algorithm. Section 6 provides experimental results and some conclusions are given in Section 7.

2. THREE-BAND MCTF SCHEME

First, let us introduce some notations : the frames in the sequence will be denoted by $(x_t(\boldsymbol{n}))$, where t is the temporal index and $\boldsymbol{n}$ is a spatial variable. In the wavelet decomposition, we will denote by h_t the detail ("high-frequency") subband frames and by l_t the approximation ("low-frequency") subband frames. Below we will only describe one transform level, but it is obvious that one can obtain a multiresolution decomposition by subsequent decompositions of the approximation band.

The 3-band (3B) decomposition scheme is presented in Fig. 1. Note that, contrary to the simple scheme described in [11], we are not in the classical case of the lifting scheme and therefore the invertibility is not structurally guaranteed. The corresponding equations describing this analysis structure are:

$$h_t^+ = x_{3t+1} - P^+[x_{3t}, x_{3t-1}], \quad (1)$$

$$h_t^- = x_{3t-1} - P^-[x_{3t}, x_{3t+1}] \quad (2)$$

$$l_t = x_{3t} + U^+[h_t^+] + U^-[h_t^-]. \quad (3)$$

As described in [11], in such a scheme we have two predict operators: P^+ and P^-, leading to two detail subbands, and two update operators for computing the approximation subband: U^+ and U^-, corresponding to forward and backward operations in time. Obviously, as in the case of two-band temporal decompositions, these operators will be non-linear by the nature of temporal prediction, involving ME/MC.

If the inversion at the decoder of the update step is obvious, as in a classical lifting, the inversion of the predict step is not guaranteed by the structure, especially as the two predict operators can involve here various time delays. The perfect reconstruction of the scheme (also involving motion compensation) is discused in Section 5.

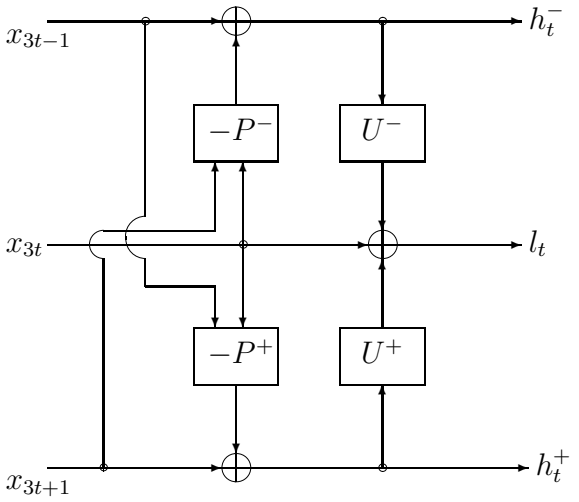

Fig. 1. The temporal lifting scheme with three bands.

3. STRUCTURE WITHOUT ME/MC

The quality of the detail subbands depends on the performance of the predictor involved in the lifting step. In [11], the used predictors were the simplest operators, i.e., identity operators. With this operator, the current frame is predicted only by motion compensation of the previous (resp., the next) frame. However, it is well known from hybrid coding that bidirectional prediction may be useful, especially in areas where occlusions occur. We will therefore introduce bidirectional predictors for both detail subbands. For the sake of simplicity, we constrain the prediction to be performed only from one future and one previous frame (which is usually the case for B-frames in hybrid coding).

Without considering motion estimation/compensation in the structure, the detail subbands are obtained by the following equations:

$$h_t^+ = x_{3t+1} - \beta x_{3t+2} - \big(1-\beta\big)x_{3t} \tag{4}$$

$$h_t^- = x_{3t-1} - \beta x_{3t-2} - \big(1-\beta\big)x_{3t} \tag{5}$$

where $\beta \in \big[0,1\big)$ is a weighting factor. Note that for $\beta = 0$ we find the previously introduced 3B "Haar-like" scheme [11].
If we denote by $H^+(z)$, resp. $H^-(z)$ the z-transform of the previous two filters, we can remark that for any $\beta \in [0,1)$, we have $H^+(1) = H^-(1) = 0$, meaning that we have indeed two high-pass filters. The parameter β can be tuned to take into account irregularities along motion trajectories, but the two detail subbands remain symmetric w.r.t. the central frame.
One can consider linear functions for the update operators:

$$U^+[h_t^+] = \alpha h_t^+, \quad U^-[h_t^-] = \alpha h_t^-$$

where the positive constant α can be determined so that l_t results from a low-pass filtering from the input sequence, that is, its frequency response cancels at the normalized frequency 1/2.

4. STRUCTURE WITH ME/MC

If we want to include the motion estimation and compensation in the two predictors and update operators, we need to consider forward/backward motion vectors, as illustrated in Fig. 2 (upper index "+" refers to forward prediction, while "-" stands for backward prediction).

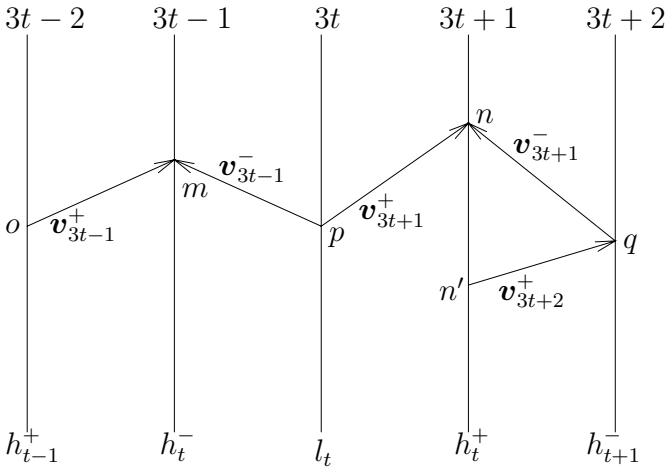

Fig. 2. Temporal prediction in a group of frames using the proposed scheme.

In this case, the analysis relations become:

$$\begin{aligned} h_t^+(\boldsymbol{n}) &= x_{3t+1}(\boldsymbol{n}) - \beta x_{3t+2}(\boldsymbol{n} - \boldsymbol{v}_{3t+1}^-) \\ &\quad - \big(1-\beta\big)x_{3t}(\boldsymbol{n} - \boldsymbol{v}_{3t+1}^+) \end{aligned} \tag{6}$$

$$\begin{aligned} h_t^-(\boldsymbol{m}) &= x_{3t-1}(\boldsymbol{m}) - \beta x_{3t-2}(\boldsymbol{m} - \boldsymbol{v}_{3t-1}^+) \\ &\quad - \big(1-\beta\big)x_{3t}(\boldsymbol{m} - \boldsymbol{v}_{3t-1}^-) \end{aligned} \tag{7}$$

$$\begin{aligned} l_t(\boldsymbol{p}) &= x_{3t}(\boldsymbol{p}) + \alpha h_t^+(\boldsymbol{p} + \boldsymbol{v}_{3t+1}^+) \\ &\quad + \alpha h_t^-(\boldsymbol{p} + \boldsymbol{v}_{3t-1}^-) \end{aligned} \tag{8}$$

Using the expressions (6)-(7) of the details in Eq. (8), we obtain a five-tap motion-compensated low-pass filter:

$$\begin{aligned} l_t(\boldsymbol{p}) &= \big[1 - 2\alpha(1-\beta)\big]x_{3t}(\boldsymbol{p}) + \alpha x_{3t+1}(\boldsymbol{p} + \boldsymbol{v}_{3t+1}^+) \\ &\quad + \alpha x_{3t-1}(\boldsymbol{p} + \boldsymbol{v}_{3t-1}^-) - \alpha\beta x_{3t+2}(\boldsymbol{p} + \boldsymbol{v}_{3t+1}^+ - \boldsymbol{v}_{3t+1}^-) \\ &\quad - \alpha\beta x_{3t-2}(\boldsymbol{p} + \boldsymbol{v}_{3t-1}^- - \boldsymbol{v}_{3t-1}^+) \end{aligned} \tag{9}$$

Remark that, by motion compensation, all the positions to be filtered in the five successive frames are aligned along the same motion trajectory and the filtering is therefore meaningful. We can notice too, that independently of the value of β, the condition of low-pass filtering for l_t yields $\alpha = 1/4$.

5. INVERTIBILITY OF THE PROPOSED SCHEME

As the proposed analysis structure does not stem from a classical lifting scheme, perfect reconstruction needs to be proved. An additional difficulty to establish the invertibility of the decomposition comes from the non-linearity in the operators, introduced by motion compensation.

In order to derive the invertibility of the scheme, we introduce some notations:

$$\begin{aligned}\widetilde{h}_t^+(\boldsymbol{n}) &= h_t^+(\boldsymbol{n}) + (1-\beta)\, x_{3t}(\boldsymbol{n} - \boldsymbol{v}_{3t+1}^+) \\ &= x_{3t+1}(\boldsymbol{n}) - \beta\, x_{3t+2}(\boldsymbol{n} - \boldsymbol{v}_{3t+1}^-) \qquad (10)\\ \widetilde{h}_t^-(\boldsymbol{n}) &= h_t^-(\boldsymbol{n}) + (1-\beta)\, x_{3t}(\boldsymbol{n} - \boldsymbol{v}_{3t-1}^-) \\ &= x_{3t-1}(\boldsymbol{n}) - \beta\, x_{3t-2}(\boldsymbol{n} - \boldsymbol{v}_{3t-1}^+) \qquad (11)\end{aligned}$$

We deduce the expression:

$$\frac{1}{1-\beta^2}\left[\widetilde{h}_t^+(\boldsymbol{n}) + \beta\widetilde{h}_{t+1}^-(\boldsymbol{n} - \boldsymbol{v}_{3t+1}^-)\right] = \frac{1}{1-\beta^2}\left[x_{3t+1}(\boldsymbol{n}) - \beta^2 x_{3t+1}(\boldsymbol{n} - \boldsymbol{v}_{3t+1}^- - \boldsymbol{v}_{3t+2}^+)\right]$$

We remark that it is possible to obtain a reconstruction formula for x_{3t+1} from the above equation, if the forward and backward motion vector fields between two successive frames are identical, with opposite sense, i.e., $\boldsymbol{v}_{3t+1}^- = -\boldsymbol{v}_{3t+2}^+$:

$$x_{3t+1}(\boldsymbol{n}) = \frac{1}{1-\beta^2}\left[\widetilde{h}_t^+(\boldsymbol{n}) + \beta\widetilde{h}_{t+1}^-(\boldsymbol{n} - \boldsymbol{v}_{3t+1}^-)\right] \qquad (12)$$

A similar expression allows to compute x_{3t+2}:

$$x_{3t+2}(\boldsymbol{n}) = \frac{1}{1-\beta^2}\left[\beta\widetilde{h}_t^+(\boldsymbol{n} + \boldsymbol{v}_{3t+1}^-) + \widetilde{h}_{t+1}^-(\boldsymbol{n})\right] \qquad (13)$$

Thus, at the synthesis, from (8) we get l_t and then, based on (10) and (11), we see that the lifting scheme can be completely inverted, under the assumption $\boldsymbol{v}_{3t+1}^- = -\boldsymbol{v}_{3t+2}^+$. This assumption however is not true for every pair of pixels connected by motion between the frames $3t+1$ and $3t+2$. The solution that we have chosen (for complexity and cost of transmission reasons) is to compute only one of these two motion vector fields, for example $\boldsymbol{v}_{3t+2}^+$. Considering this solution, bidirectionally connected pixels in the frame $3t+1$ are obtained by backward prediction with $\boldsymbol{v}_{3t+1}^- = -\boldsymbol{v}_{3t+2}^+$ (see Eq.(6) and (12)). For the pixels not connected with the frame $3t+2$, we apply a simple monodirectional prediction from the frame $3t$, using the vector $\boldsymbol{v}_{3t+1}^+$.

6. EXPERIMENTAL RESULTS

The proposed structure has been compared with other 2-band MCTF schemes and with the previously introduced Haar-like 3-band structure. In all cases, the resulting temporal subband frames have been spatially decomposed with the same 9/7 biorthogonal multiresolution analysis and further encoded using the MC-EZBC software [4].

Since the quantization is performed identically in all the temporal subbands, a re-normalization is necessary in order to be as close as possible to an orthonormal situation. The normalized filters will be obtained as:

$$\widetilde{l_t} = k_l\, l_t, \quad \widetilde{h}_t^+ = k_h\, h_t^+, \quad \widetilde{h}_t^- = k_h\, h_t^-,$$

where l_t, h_t^+ and h_t^- are the filters defined before. Due to the symmetry of the scheme, we consider equal normalizations for h_t^+ and h_t^-.

The condition applied here is to impose a unitary norm condition for the impulse responses of the filters involved in the 3B structure. Thus, with $\alpha = 1/4$, we have:

$$k_l^2\left\{\left[1 - \frac{1}{2}(1-\beta)\right]^2 + 2\cdot\frac{1}{4^2} + 2\cdot\left(\frac{\beta}{4}\right)^2\right\} = k_h^2\left(1 + \beta^2 + (1-\beta)^2\right) = 1.$$

This approach leads to

$$k_l = \sqrt{\frac{8}{3}}\cdot\frac{1}{\sqrt{1 + \frac{4}{3}\beta + \beta^2}}$$

$$k_h = \frac{1}{\sqrt{2}}\cdot\frac{1}{\sqrt{1 - \beta + \beta^2}}$$

Note that the values $k_l = \sqrt{\frac{8}{3}}$ and $k_h = \frac{1}{\sqrt{2}}$ correspond to the re-normalization proposed for Haar-like 3-band scheme [13], [11]. Using these values, the PSNR variation of the reconstructed sequence "mobile" as a function of $\beta \in [0, 0.9)$ is illustrated in Fig. 3. Note first that the optimal value is not $\beta = 0.5$ (corresponding to symmetric predictors and update operator). Secondly, we remark an important coding performance decrease as β goes close to 1. Even though our tests have shown that the energy of the detail frames is minimal for $\beta = 0.5$, the reconstruction error rapidly increases with larger values of β, which explains both phenomena.

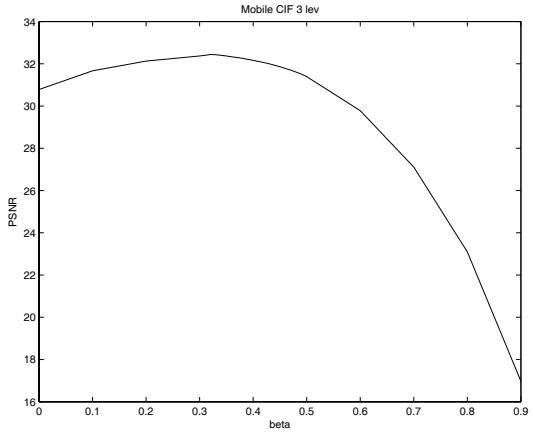

Fig. 3. *PSNR of the reconstructed sequence "mobile" as a function of* $\beta \in [0, 0.9)$.

In Tables 1 and 2 we compare the coding efficiency of the proposed scheme with the equivalent codec based on the Haar-like 3-band structure and also with a 2-band Haar MCTF. The two 3-band decompositions are performed over three temporal levels, which leads to a group of pictures (GOF) of 27 frames. The performance of the dyadic temporal decomposition is tested for GOFs of size 16 (4 decomposition levels) and 32 (5 levels). The optimal value for β, which is $\beta = 0.32$, was used in the 3B scheme with bidirectional operators.

One can remark that the proposed 3B MCTF outperforms of 0.3–0.5 dB the Haar-like 3B scheme and by 0.4–1 dB the dyadic

Bitrate	3B Haar-like	3B bidir	2B 4 lev	2B 5 lev
400	24.77	25.30	24.22	24.43
600	26.15	26.68	25.66	25.78
800	27.08	27.71	26.67	26.78
1024	28.09	28.71	27.62	27.65
1500	29.72	30.59	29.37	29.40
2048	31.45	32.44	31.04	31.04

Table 1. *Rate-distortion comparison of dyadic and triadic schemes for "mobile" CIF 30fps sequence (bitrate in kbs, PSNR in dB).*

Bitrate	3B Haar-like	3B bidir	2B 4 lev	2B 5 lev
400	32.93	33.48	32.58	32.46
600	34.23	34.72	34.08	33.94
800	35.36	35.83	35.11	34.96
1024	36.29	36.72	36.20	36.04
1500	37.93	38.31	37.84	37.70
2048	39.57	39.86	39.59	39.46

Table 2. *Rate-distortion comparison of dyadic and triadic schemes for "foreman" CIF 30fps sequence (bitrate in kbs, PSNR in dB).*

structure on a sequence with irregular motion, like "foreman". This improvement is related to the bidirectional nature of its update and predict operators. Moreover, on a sequence with uniform motion like "mobile", the PSNR difference ranges from 0.5–1 dB with the Haar-like 3B scheme up to 1.4 dB with the dyadic MCTF.

Note that the coding efficiency was improved also by using an optimized prediction for joint estimation of motion vectors like in [19]. This technique leads to better prediction and more coherent MVF. As an example, for the "stefan" CIF sequence at 2Mbs the MV cost with this strategy was representing 11.3%, while for separate search and coding of MVF they would represent 16.9% of the total bitstream.

7. CONCLUSION

In this paper, we have presented a three-band motion-compensated temporal subband decomposition for scalable video compression which is the 3B non-linear counterpart of the dyadic 5/3 multiresolution analysis. We have proven the perfect reconstruction of this scheme which does not enter into the classical lifting formalism. By taking advantage of parametrized bidirectional predict and update operators, it outperforms the Haar-like 3B scheme, while providing flexible temporal scalability factors multiple of 3.

8. REFERENCES

[1] S.J. Choi and J.W. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. on Image Proc.*, vol. 8, pp. 155–167, 1999.

[2] J.-R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Trans. on Image Proc.*, vol. 3, pp. 559–589, 1994.

[3] S.T. Hsiang and J.W. Woods, "Invertible three-dimensional analysis/synthesis system for video coding with half-pixel accurate motion compensation," in *VCIP 99, SPIE Vol. 3653*, 1999, pp. 537–546.

[4] "3D MC-EZBC software package," available on the MPEG CVS repository.

[5] B.-J. Kim, Z. Xiong, and W.A. Pearlman, "Very low bit-rate embedded video coding with 3-D set partitioning in hierarchical trees (3D-SPIHT)," *IEEE Trans on Circ. and Syst. for Video Tech.*, vol. 8, pp. 1365–1374, 2000.

[6] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," in *IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, Salt Lake City, UT, May 2001.

[7] A. Secker and D. Taubman, "Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting," in *Proceedings of the IEEE International Conference on Image Processing*, Thessaloniki, Greece, Oct. 2001.

[8] T. Russert and K. Hanke, "Optimized quantization in interframe wavelet coding," doc. m9003; Shanghai MPEG meeting, Oct. 2002.

[9] J.W. Woods, P. Chen, and S.-T. Hsiang, "Exploration experimental results and software," doc. m8524, Klagenfurt MPEG meeting, July 2002.

[10] D. Turaga and M. van der Schaar, "Unconstrained temporal scalability with multiple reference and bi-directional motion compensated temporal filtering," doc. m8388, Fairfax MPEG meeting, 2002.

[11] C. Tillier and B. Pesquet-Popescu, "3D, 3-band, 3-tap temporal lifting for scalable video coding," in *Proceedings of the IEEE International Conference on Image Processing*, Barcelona, Spain, Sept. 2003.

[12] M. van der Schaar and D. S. Turaga, "Unconstrained motion compensated temporal filtering (UMCTF) framework for wavelet video coding," in *Proc. of IEEE ICASSP*, 2003.

[13] J.-R. Ohm, "Complexity and delay analysis of MCTF interframe wavelet structures," doc. m8520, Klagenfurt MPEG meeting, July 2002.

[14] D. S. Turaga, M. van der Schaar, and B. Pesquet-Popescu, "Complexity scalable motion compensated wavelet video encoding," submitted to IEEE Trans. on Circ. and Syst. for Video Tech., 2003.

[15] F. J. Hampson and J.-C. Pesquet, "M-band nonlinear subband decompositions with perfect reconstruction," *IEEE Trans. on Image Proc.*, vol. 7, pp. 1547–1560, 1998.

[16] T. Tran, "m-channel linear phase perfect reconstruction filter bank with rational coefficients," *IEEE Trans. on Circuits and Systems – I: Fundamental Theory and Applications*, vol. 49, pp. 914–927, 2002.

[17] Y.J. Chen, S. Oraintara, and K. Amaratunga, "m-channel lifting-based design of paraunitary and biorthogonal filter banks with structural regularity," in *Proceedings of the IEEE International Conference on Circuits and Systems*, May 2003, pp. IV 221–IV 224.

[18] Y. Zhan, M. Picard, B. Pesquet-Popescu, and H. Heijmans, "Long temporal filters in lifting schemes for scalable video coding," doc. m8680, Klagenfurt MPEG meeting, July 2002.

[19] G. Pau, C. Tillier, and B. Pesquet-Popescu, "Optimization of the predict operator in lifting-based motion compensated temporal filtering," in *Proceedings of the SPIE VCIP*, San Jose, CA, Jan. 2004.