UNIFORM MOTION-COMPENSATED 5/3 FILTERBANK FOR SUBBAND VIDEO CODING

Grégoire Pau and Béatrice Pesquet-Popescu

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

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. In this paper, we propose an alternative structure for the temporal prediction in the 5/3 filterbank. It significantly reduces the ghosting artefacts in the temporal approximation subband frames, providing a higher quality scalability and improved compression performance, for an equivalent complexity.

1. INTRODUCTION

Subband motion-compensated temporal filtering (MCTF) video codecs have attracted recently a lot of attention, due to their high compression performance comparable to state-of-the-art hybrid codecs and due to their scalability features. The spatio-temporal subband scheme ($2D + t$) exploits the temporal interframe redundancy by applying a temporal wavelet transform in the motion direction over the frames of a video sequence. Temporal filtered subband frames are further spatially decomposed and can be encoded by different algorithms such as 3D-SPIHT [1] or MC-EZBC [2].

Previous works exploiting the Haar motion-compensated temporal transform [3] provided very promising results, but as the Haar filter is a short temporal filter, it has been shown [4, 5, 6] that the use of longer bidirectional filters, like the 5/3 filter bank, can take better advantage of the temporal redundancy existing between frames.

However, a common drawback of the schemes using a block-based motion estimation is that they have to deal with unconnected pixels during the update step of the lifting implementation. Unconnected areas are generally not changed by the temporal low-pass filtering and thus create abrupt changes between the connected areas and the unconnected ones in the approximation frames. This is undesirable from a compression point of view, since the presence of these areas may introduce ghosting artefacts in the approximation frames, and also propagate to the next temporal level, as these zones are generally not selected by the motion estimation algorithm for temporal prediction. Several attempts to avoid these abrupt changes have already been proposed, like applying a smooth transition filtering between unconnected and connected areas [7] or selecting the best candidates for updating [8], alleviating but not completely solving the problem of unconnected pixels.

We present in this paper an alternative prediction scheme based on the skeleton of a classical 5/3 MC temporal lifting filter bank. In contrast with the classical 5/3 motion-compensated temporal filter described in previous works [8, 9, 10], which uses for each lifting step a forward and a backward motion vector field (MVF), this new temporal filter bank uses two MVF in the same direction. A similar idea was also independently proposed in [11]. An advantage of this new scheme is that it does not produce unconnected pixels at all, leading thus to a greatly improved visual quality of the approximation subbands. A main consequence is the higher quality temporal scalability of this new scheme, compared with existing ones.

The paper is organized as follows: in the next section, we review the plain 5/3 MC filter bank. In Section 3 we present the proposed uniform 5/3 MC filter bank and discuss some of its properties. Section 4 illustrates by simulation results the coding performance and the scalability properties of the proposed scheme. We conclude in Section 5.

2. PLAIN MOTION-COMPENSATED 5/3 FILTER BANK

First, we 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, it is usual to denote by $h_{t,j}$ the detail (temporal "high-frequency") subband frames and by $l_{t,j}$ the approximation (temporal "low-frequency") subband frames at temporal resolution level $j \in \mathbb{N}^*$. We will describe only one transform level, but it is clear that one can obtain a multiresolution decomposition by subsequent decomposition of the approximation band. Due to this fact, we will drop the index j in what follows.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

Temporal lifting operations are performed on MC frames, by predicting odd frames x_{2t+1} from even ones, i.e. x_{2t} and x_{2t+2}. In this case, it is clear that for all t, we have to deal with two motion vector fields: the forward MVF, predicting x_{2t+1} from x_{2t} and denoted by $\boldsymbol{v}^{+}_{2t+1}$ and the backward one, predicting x_{2t+1} from x_{2t+2} and denoted by $\boldsymbol{v}^{-}_{2t+1}$. With these notations, the plain 5/3 temporal transform applied along the motion direction can be expressed in lifting form by:

$$h_t(\boldsymbol{n}) = x_{2t+1}(\boldsymbol{n}) - \alpha\, x_{2t}\big(\boldsymbol{n} - \boldsymbol{v}^{+}_{2t+1}(\boldsymbol{n})\big) - \beta\, x_{2t+2}\big(\boldsymbol{n} - \boldsymbol{v}^{-}_{2t+1}(\boldsymbol{n})\big) \quad (1)$$

$$l_t(\boldsymbol{m}) = x_{2t}(\boldsymbol{m}) + \gamma h_{t-1}\big(\boldsymbol{m} + \boldsymbol{v}^{-}_{2t-1}(\boldsymbol{p})\big) + \delta h_t\big(\boldsymbol{m} + \boldsymbol{v}^{+}_{2t+1}(\boldsymbol{q})\big) \quad (2)$$

The update equation (2) involves $\boldsymbol{p}$ and $\boldsymbol{q}$, satisfying $\boldsymbol{p} - \boldsymbol{v}^{-}_{2t-1}(\boldsymbol{p}) = \boldsymbol{m}$ and $\boldsymbol{q} - \boldsymbol{v}^{+}_{2t+1}(\boldsymbol{q}) = \boldsymbol{m}$. As these equations may have no solutions (no connection case) or multiple ones (multiple connection case), the determination of $\boldsymbol{p}$ and $\boldsymbol{q}$ is a tricky point which has been previously discussed [8, 12]. The coefficients $\alpha = \beta = 1/2$ and $\delta = \gamma = 1/4$ correspond to the biorthogonal 5/3 filterbank and they will be used in the sequel, although the biorthogonal properties are not perfectly satisfied in this MC non-linear framework.

If fractional-pel motion estimation (ME) is performed, then the predict and update operators should involve an interpolation, which has been already described [3, 13].

The lifting formalism guarantees the invertibility of the scheme, and therefore samples can be simply retrieved using the same lifting steps in reverse order with opposite signs. Some properties of this filter bank can be drawn:

- The predict operator described by the equation (1) is always bidirectional. This is an important feature, which often leads to a better prediction than a mono-directional operator, especially in case of smooth motion. A suboptimal algorithm able to jointly estimate the best forward and backward MVF minimizing the energy of detail subbands for the plain 5/3 MC filter has been proposed in [10]. However, when the frame cannot be bidirectionally predicted, i.e. in a scene cut, this feature may be undesirable.

- The update operator described by the equation (2) can be either bidirectional, mono-directional or null, depending of the connections of the pixels. Pixels unconnected on both sides are actually not low-pass filtered (they keep their original value, being at most renormalized). As stated before, these pixels create abrupt changes in the approximation frames which may create the so-called "ghosting artefacts" and prevent a smooth prediction at the next temporal level.

Despite of its propensity to create unconnected areas, but thanks to its good bidirectional prediction properties, this filter was shown [10] to perform much better from a compression point of view than a mono-directional Haar filterbank.

3. UNIFORM 5/3 MCTF SCHEME

The motivation in designing a new MC lifting scheme is to avoid the creation of ghosting artefacts in approximation bands. We achieve this goal by an update operator using two MVF with the same orientation, and thus not creating unconnected pixels. An example of temporal filtering in this case is given in Figs. 1 and 2, where backward MVF have been used for illustration.

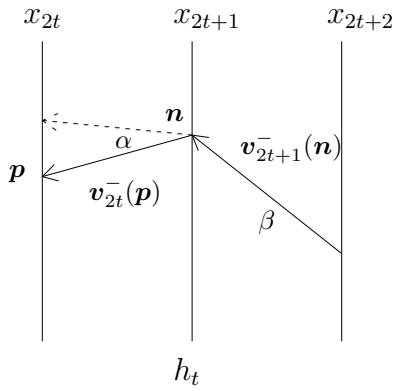

Fig. 1. Predict operator in the spatio-temporal domain. Plain arrows indicate the motion compensation directions. Dashed arrows indicate the possible multiple connections of the pixel $\boldsymbol{n}$.

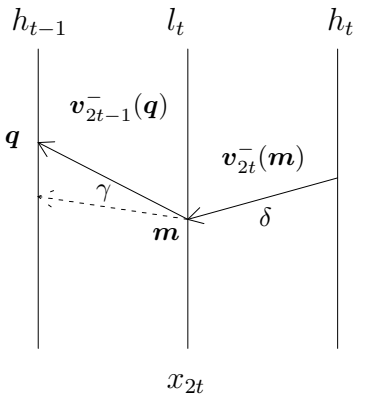

Fig. 2. Update operator in the spatio-temporal domain. Plain arrows indicate the motion compensation directions. Dashed arrows indicate the possible multiple connections to the pixel $\boldsymbol{m}$.

The new analysis lifting equations in this case can be written as:

$$h_t(\boldsymbol{n}) = x_{2t+1}(\boldsymbol{n}) - \alpha\, x_{2t}(\boldsymbol{n} + \boldsymbol{v}_{2t}^{-}(\boldsymbol{p})) - \beta\, x_{2t+2}\big(\boldsymbol{n} - \boldsymbol{v}_{2t+1}^{-}(\boldsymbol{n})\big) \quad (3)$$

$$\text{with } \begin{cases} \alpha = \beta = 1/2 & \text{if } \exists\, \boldsymbol{p} \text{ such } \boldsymbol{p} - \boldsymbol{v}_{2t}^{-}(\boldsymbol{p}) = \boldsymbol{n} \\ \alpha = 0, \quad \beta = 1 & \text{otherwise,} \end{cases}$$

$$l_t(\boldsymbol{m}) = x_{2t}(\boldsymbol{m}) + \gamma h_{t-1}(\boldsymbol{m} + \boldsymbol{v}_{2t-1}^{-}(\boldsymbol{q})) + \delta h_t\big(\boldsymbol{m} - \boldsymbol{v}_{2t}^{-}(\boldsymbol{m})\big) \quad (4)$$

$$\text{with } \begin{cases} \gamma = \delta = 1/4 & \text{if } \exists\, \boldsymbol{q} \text{ such } \boldsymbol{q} - \boldsymbol{v}_{2t-1}^{-}(\boldsymbol{q}) = \boldsymbol{m} \\ \gamma = 0, \quad \delta = 1/2 & \text{otherwise.} \end{cases}$$

An appealing feature of this new temporal lifting scheme is that every pixel is always connected to another one in the previous frame, as well in the predict as in the update step. We can derive some properties of this scheme:

- The temporal prediction described by the equation (3) can be performed either with a bidirectional or with a mono-directional operator. In case of a simple connected pixel, the predict operator turns out to correspond to a Haar predictor. In case of a pixel connected on both sides, the pixel is bidirectionally predicted. Compared to the plain 5/3 MC filter, the prediction is not always bidirectional. However, the lack of connection in one direction is often related to the newly uncovered areas, which can be predicted only in one direction.

- The update operator is very similar to the predict one as it can be either bidirectional or mono-directional. Compared to the plain 5/3 MC filter, no unconnected pixels can appear, since the synchonous frame x_{2t} serves as reference for ME of the backward $\boldsymbol{v}_{2t}^{-}$ MVF.

This filter has the same complexity as the plain 5/3 MC filter as both of them are using the same number of motion vector searches and because Eqs. (1), (2), (3) and (4) have equivalent complexity. Note that backward MVF are employed in the current implementation, however an equivalent scheme is obtained using forward ME/MC. A possible extension of this framework is to alternate the motion vectors estimation on a group of frames (GOF) basis, depending on the dominant motion direction of the GOF.

4. EXPERIMENTAL RESULTS

The implementation of the temporal filtering corresponds to a "sliding window" (or "on-the-fly") technique [4]. Motion estimation is done with the Hierarchical Variable Size Block Matching (HVSBM) algorithm with block sizes varying from 64×64 to 4×4. Window search range is first initialized at $[-2; 2]$, is increased if no good match can be found and is doubled at each temporal level. Motion vector field encoding and bitrate allocation are done within the MC-EZBC framework. Temporal subbands are then spatially decomposed over five resolution levels using biorthogonal 9/7 wavelets and the resulting spatio-temporal wavelet coefficients are encoded using the MC-EZBC [2] algorithm.

In our simulations, we consider two representative test color video sequences in CIF format at 30 fps: "Tempete" and "Mobile", which have been selected for their very different motion and texture characteristics. These video sequences have been decomposed over four temporal levels (five for "Mobile") and motion vectors have been estimated with $1/4^{th}$ pixel accuracy ($1/8^{th}$ for "Mobile"). Both video sequences have been encoded in the YUV color mode, meaning that the bit budget is shared by the luminance and chrominance components, the bitstream headers and the motion vector fields. Performances in Tables 1 and 2 are however only expressed in terms of YSNR, calculated by averaging the Y component PSNR of all decoded frames. We observe that the uniform structure exhibit better performance than the plain 5/3 filterbank, although the latter one benefits from the bidirectional prediction.

YSNR (in dB)	512 kbs	768 kbs	1024 kbs	1536 kbs
Uniform 5/3	31.57	33.35	34.51	36.41
Plain 5/3	30.89	32.74	33.91	35.86
Haar	29.78	31.49	32.76	34.63

Table 1. Rate-distortion comparison of several temporal filters on "Tempete" CIF video sequence (30fps).

YSNR (in dB)	512 kbs	768 kbs	1024 kbs	1536 kbs
Uniform 5/3	30.19	32.36	33.81	36.00
Plain 5/3	29.68	31.97	33.52	35.73
Haar	28.59	30.90	32.44	34.39

Table 2. Rate-distortion comparison of several temporal filters on "Mobile" CIF video sequence (30fps).

An interesting feature of the proposed scheme can be observed from Table 3. The percentage of pixels mono-connected during the update step is decreased at all temporal resolution levels with the new scheme, compared with a plain 5/3 MC filterbank (and, as expected, there are no unconnected pixels).

The importance of this feature for temporal scalability is illustrated in Fig. 3, where two approximation frames at the fourth temporal level are compared: one obtained with the uniform structure, the other resulting from the plain 5/3 filterbank. One can remark the improved visual quality of the former one.

Uniform 5/3	0	1	> 1
Temporal level 1	0.00	4.13	95.87
Temporal level 2	0.00	9.01	90.99
Temporal level 3	0.00	17.42	82.58
Temporal level 4	0.00	26.67	73.33

Plain 5/3	0	1	> 1
Temporal level 1	0.67	6.86	92.47
Temporal level 2	2.23	13.79	83.98
Temporal level 3	5.46	22.80	71.74
Temporal level 4	10.73	29.62	59.65

Table 3. Percentages of unconnected (0), mono-connected (1) and multiple connected pixels (> 1) during the update step of the uniform 5/3 MCTF (up) and plain 5/3 MCTF (down) scheme at several temporal decomposition levels. These results were obtained on a four level decomposition of the CIF sequence "Foreman".

Fig. 3. Approximation frames at the fourth temporal decomposition level obtained with the uniform 5/3 MCTF (up) and the plain 5/3 MCTF (down) on "Stefan" sequence .

5. CONCLUSION AND FUTURE WORK

We have proposed a new temporal prediction structure for the motion-compensated temporal 5/3 filterbank. It exploits a uniform motion estimation strategy to avoid the occurence of unconnected pixels in the update step of the lifting scheme. This strongly reduces the ghosting artefacts in the approximation frames, leading to higher compression performance and an improved temporal scalability compared with the classical 5/3 structure, for an equivalent complexity. The new structure pleads in favor of a truly adaptive structure, combining the advantages of both schemes and providing "on-the-fly" the optimal temporal prediction structure. An efficient estimation and encoding of the motion vectors would be a key factor for the optimality of such a scheme.

6. REFERENCES

[1] 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 Transactions on Circuits and Systems for Video Technology*, vol. 8, pp. 1365–1374, 2000.

[2] S. Hsiang and J. Woods, "Embedded image coding using zeroblocks of subband/wavelet coefficients and context modeling," *IEEE International Symposium on Circuits and Systems*, p. 589, 2000.

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

[4] 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.

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

[6] 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.

[7] K. Hanke, J.-R. Ohm, and T. Rusert, "Adaptation of filters and quantization in spatio-temporal wavelet coding with motion compensation," in *Proc. of the Picture Coding Symposium*, St. Malo, France, April 2003, pp. 49–54.

[8] C. Tillier, B. Pesquet-Popescu, Y. Zhang, and H. Heijmans, "Scalable video compression with temporal lifting using 5/3 filters," in *Proc. of the Picture Coding Symposium*, St. Malo, France, April 2003, pp. 55–58.

[9] A. Secker and D. Taubman, "Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting," in *Proc. of the IEEE Int. Conf. on Image Processing*, Thessaloniki, Greece, Oct. 2001, pp. 1029–1032.

[10] G. Pau, C. Tillier, and B. Pesquet-Popescu, "Optimization of the predict operator in lifting-based motion compensated temporal filtering," in *Proc. of Visual Communications and Image Processing*, San Jose, CA, January 2004.

[11] A. Golwelklar and J. W. Woods, "Scalable video compression using longer motion compensated temporal filters," in *Proc. of Visual Communications and Image Processing*, Lugano, Switzerland, July 2003.

[12] G. Pau, C. Tillier, B. Pesquet-Popescu, and H. Heijmans, "Motion compensation and scalability in lifting-based video coding," to be published in *Signal Processing: Image Communication*.

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