WEIGHTED AVERAGE SPATIO-TEMPORAL UPDATE OPERATOR FOR SUBBAND VIDEO CODING

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

Spatio-temporal motion-compensated wavelet decomposition is an increasingly popular method for scalable video coding, with coding ef£ciency which is competitive with state-of-the-art non-scalable codecs. In this paper, we propose a new spatio-temporal update operator in the lifting scheme allowing ef£cient implementation of these temporal decompositions. We demonstrate its improved performance both theoretically, by exhibiting a decrease in the reconstruction error, and by simulation results.

1. INTRODUCTION

Scalable video coding based on motion-compensated (MC) spatio-temporal (2D+t) wavelet decompositions is becoming increasingly popular, as it provides similar coding ef£ciency as state-of-the-art non-scalable codecs, while also providing adaptivity to bandwidth variations and receiver device capabilities (framerate, display size, CPU, ...). Most of proposed wavelet video schemes are using a block-based motion compensation (with £xed or variable block sizes), due to its simplicity and popularity. However, this technique raises in the case of subband temporal £ltering several problems related to the non-symmetry of the prediction relation: the motion vector £eld (MVF) estimated from a frame A to a frame B is not exactly the opposite of the MVF estimated from B to A. This leads to the so-called "unconnected" and "multiple connected" pixels [1], [2], which lead to annoying coding artefacts and reduced texture coding ef£ciency (as these pixels result in discontinuities in the wavelet domain). A solution to alleviate this problem could be the use of mesh models for motion estimation and compensation [3]. However, the popularity of block-based motion estimation techniques pleads in favor of £nding an ef£cient solution for incorporating these schemes into wavelet video codecs. The case of unconnected pixels was considered, and a solution has recently been proposed that reduces the annoying visual artifacts by employing a lowpass transition spatial £ltering [4], [5]. However, this solution does not solve the problem of "multiple connected" pixels. In [6], we have shown that using a non-linear lifting framework, any transformation can be applied to the various multiple connected coef£cients, while preserving the perfect reconstruction. In this paper, we investigate novel processing strategies for multiple connected pixels that lead to an improved coding ef£ciency. We start with an analysis of the reconstruction error of the different types of pixels in a MC temporal 2-band Haar scheme. We then propose a new spatio-temporal update operator, performing a weighted average of all pixels connected to a single pixel in the approximation frame. Compared to a previous approach, where "the best" pixel was chosen for £ltering according to certain criteria [6], our current strategy has the advantage of an increased robustness with respect to the quantization effects. The cases of normalized mean and unconstrained linear combination of values are analyzed. We prove that this averaging leads to a decrease of the reconstruction error globally and also individually on most of the pixels involved in the spatio-temporal £ltering. We con£rm these theoretical results by simulations both on 2-band MCTF and on 3-band MCTF codecs. Indeed, the same problem appears in non-dyadic subband coding schemes involving a block-based motion compensation [7].

The paper is organized as follows. In the next section, we analyse the quantization errors in the existing 2-band scheme. In Section 3, we introduce the proposed average operator and analyze the resulting improvements. Section 4 illustrates via simulation results on both 2-band and 3-band systems the coding performance of the proposed method. We conclude in Section 5.

2. ANALYSIS OF THE EXISTING 2-BAND SCHEME

Consider two pixels, $\boldsymbol{n}$ and $\boldsymbol{m}$ in the odd frame x_{2t+1}, predicted by motion compensation from the same pixel $\boldsymbol{p}$ in the even frame x_{2t} (see Fig.1). We say that $\boldsymbol{p}$ is connected with $\boldsymbol{n}$ and $\boldsymbol{m}$. Let us denote by $\boldsymbol{q}$ an unconnected pixel in x_{2t} (not used for the prediction of any pixel in x_{2t+1}). Consider now that only the pixel $\boldsymbol{m}$ is chosen for £ltering (the choice being often done by the scanning order in the frame x_{2t+1} [2] or according to any other criterion, as those exposed in [6]). Then, considering the lifting formulation of the Haar MCTF [6], the analysis equations corresponding to these pixels are:

$$H(\boldsymbol{m}) = \frac{1}{\sqrt{2}}(\boldsymbol{m} - \boldsymbol{p}), \quad H(\boldsymbol{n}) = \frac{1}{\sqrt{2}}(\boldsymbol{n} - \boldsymbol{p})$$
$$L(\boldsymbol{p}) = \sqrt{2}\left[\boldsymbol{p} + \frac{1}{\sqrt{2}}H(\boldsymbol{m})\right], \quad L(\boldsymbol{q}) = \sqrt{2}\boldsymbol{q},$$

where H and L are respectively the temporal detail (high-pass) and approximation (low-pass) subband frames. The synthesis equa-

0-7803-8554-3/04/$20.00 ©2004 IEEE.

tions read:

$$\boldsymbol{p} = \frac{1}{\sqrt{2}}\big(L(\boldsymbol{p}) - H(\boldsymbol{m})\big), \quad \boldsymbol{q} = \frac{1}{\sqrt{2}}L(\boldsymbol{q}) \tag{1}$$

$$\boldsymbol{m} = \sqrt{2}H(\boldsymbol{m}) + \boldsymbol{p}, \quad \boldsymbol{n} = \sqrt{2}H(\boldsymbol{n}) + \boldsymbol{p} \tag{2}$$

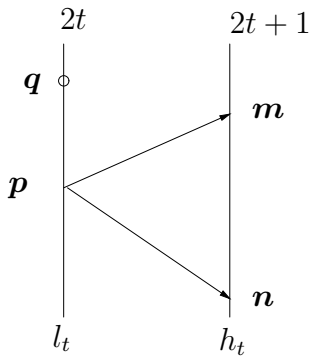

Fig. 1. Motion compensated prediction in the 2-band Haar scheme.

Let us denote by $\varepsilon_m,\ \varepsilon_n,\ \varepsilon_p$ and ε_q the quantization errors respectively on $H(\boldsymbol{m}),\ H(\boldsymbol{n}),\ L(\boldsymbol{p})$ and $L(\boldsymbol{q})$, which will be considered statistically independent of the original signal. Then, from Eqs. (1)-(2), the variances of the reconstruction errors will be related to the variances of the quantization errors as follows:

$$\sigma_p^2 = \frac{1}{2}\big(\sigma_{\varepsilon_p}^2 + \sigma_{\varepsilon_m}^2\big), \quad \sigma_q^2 = \frac{1}{2}\sigma_{\varepsilon_q}^2 \tag{3}$$

$$\sigma_m^2 = \frac{1}{2}\big(\sigma_{\varepsilon_p}^2 + \sigma_{\varepsilon_m}^2\big) \tag{4}$$

$$\sigma_n^2 = 2\sigma_{\varepsilon_n}^2 + \frac{1}{2}\big(\sigma_{\varepsilon_p}^2 + \sigma_{\varepsilon_m}^2\big). \tag{5}$$

Remark from Eqs. (4) and (5) that the reconstruction error of two pixels in the frame x_{2t+1} is different, depending on whether it has been used in the update operation or not. In the sequel, we describe an averaging method that leads to equal reconstruction errors for all pixels, as well as to a decreased global reconstruction error.

Before proceeding with the analysis of the new scheme, note that if all the coef£cients in the detail/approximation frames are quantized with the same quantization step, then $\sigma_{\varepsilon_m}^2 = \sigma_{\varepsilon_n}^2 = \sigma_\varepsilon^2$ and $\sigma_{\varepsilon_p}^2 = \sigma_{\varepsilon_q}^2$. Therefore, we have:

$$\sigma_m^2 = \frac{1}{2}\big(\sigma_\varepsilon^2 + \sigma_{\varepsilon_p}^2\big), \quad \sigma_n^2 = \frac{1}{2}\big(5\sigma_\varepsilon^2 + \sigma_{\varepsilon_p}^2\big) \tag{6}$$

$$\sigma_p^2 = \frac{1}{2}\big(\sigma_\varepsilon^2 + \sigma_{\varepsilon_p}^2\big), \quad \sigma_q^2 = \frac{1}{2}\sigma_{\varepsilon_p}^2.$$

If there are N pixels in frame x_{2t+1} that are connected to the same pixel $\boldsymbol{p}$, then one of them will have the reconstruction error $\boldsymbol{m}$, while the remanining pixels will have the same reconstruction error as $\boldsymbol{n}$. The sum of all their reconstruction errors will be:

$$\sigma_H^2 = \frac{1}{2}\big(N\sigma_{\varepsilon_p}^2 + [5(N-1)+1]\sigma_\varepsilon^2\big). \tag{7}$$

3. PROPOSED WEIGHTED AVERAGE OPERATOR

Let us denote by $(\boldsymbol{m}_i)_{i\in\{1,\ldots,N\}}$ the N pixels in frame x_{2t+1} connected to $\boldsymbol{p}$. Instead of choosing a single pixel for updating, the weighted sum of the details in these points is now used for update:

$$H(\boldsymbol{m}_i) = \frac{1}{\sqrt{2}}\big(\boldsymbol{m}_i - \boldsymbol{p}\big)$$

$$L(\boldsymbol{q}) = \sqrt{2}\boldsymbol{q}, \quad L(\boldsymbol{p}) = \sqrt{2}\boldsymbol{p} + \sum_{i=1}^{N}\alpha_i H(\boldsymbol{m}_i)$$

where $(\alpha_i)_{i\in\{1,\ldots,N\}}$ are some real constants. The synthesis equations are as before, except for $\boldsymbol{p}$, which is obtained from:

$$\boldsymbol{p} = \frac{1}{\sqrt{2}}\left(L(\boldsymbol{p}) - \sum_{i=1}^{N}\alpha_i H(\boldsymbol{m}_i)\right).$$

Accordingly, the reconstruction error for $\boldsymbol{p}$ becomes:

$$\sigma_p^2 = \frac{1}{2}\big(\sigma_{\varepsilon_p}^2 + \sum_{i=1}^{N}\alpha_i^2\sigma_\varepsilon^2\big) \tag{8}$$

while that of one of the connected pixels, $\boldsymbol{m}_{i_0}$, will be:

$$\sigma_{\boldsymbol{m}_{i_0}}^2 = \frac{1}{2}\sigma_{\varepsilon_p}^2 + 2\big(1-\alpha_{i_0}\big)\sigma_\varepsilon^2 + \frac{1}{2}\sum_{i=1}^{N}\alpha_i^2\sigma_\varepsilon^2.$$

The sum of reconstruction errors over all connected pixels is:

$$\sum_{i=1}^{N}\sigma_{\boldsymbol{m}_i}^2 = \left(\frac{N}{2}\sum_{i=1}^{N}\alpha_i^2 + 2N - 2\sum_{i=1}^{N}\alpha_i\right)\sigma_\varepsilon^2 + \frac{N}{2}\sigma_{\varepsilon_p}^2$$

or, equivalently, the mean error on connected pixels is:

$$\frac{1}{N}\sum_{i=1}^{N}\sigma_{\boldsymbol{m}_i}^2 = \left(\frac{1}{2}\sum_{i=1}^{N}\alpha_i^2 + 2 - 2\frac{\sum_{i=1}^{N}\alpha_i}{N}\right)\sigma_\varepsilon^2 + \frac{1}{2}\sigma_{\varepsilon_p}^2. \tag{9}$$

Note that the reconstruction error on the unconnected pixel $\boldsymbol{q}$ does not change when using the new strategy, so it will not be considered in the sequel.

3.1. Averaging connected pixels

If we set the condition $\sum_{i=1}^{N}\alpha_i = 1$ (this means we really update with an average of all connected pixels and do not change the mean value of data), then minimizing the mean error on these pixels amounts to minimizing the expression $\frac{1}{2}\sum_{i=1}^{N}\alpha_i^2 + 2 - \frac{2}{N}$. In the meantime, this also minimizes the reconstruction error on $\boldsymbol{p}$ in (8). It is easy to see that the weigths minimizing this expression are

$$\alpha_i = \frac{1}{N}. \tag{10}$$

In this case, the error compared with the previous case will be, for all $N \geq 1$:

- larger on pixel $\boldsymbol{m}$ (the only one used in the previous scenario for updating). Indeed,

$$\sigma_m^2 = \frac{1}{2}\sigma_{\varepsilon_p}^2 + \big(2 - \frac{3}{2N}\big)\sigma_\varepsilon^2 \geq \frac{1}{2}\big(\sigma_{\varepsilon_p}^2 + \sigma_\varepsilon^2\big)$$

- smaller on all the other pixels connected to $\boldsymbol{p}$ (denoted by $\boldsymbol{m}_i$):

$$\sigma^2_{m_i} = \frac{1}{2}\sigma^2_{\varepsilon_p} + \big(2 - \frac{3}{2N}\big)\sigma^2_{\varepsilon} \leq \frac{1}{2}\sigma^2_{\varepsilon_p} + \frac{5}{2}\sigma^2_{\varepsilon}$$

- smaller on $\boldsymbol{p}$:

$$\sigma^2_{p} = \frac{1}{2}\sigma^2_{\varepsilon_p} + \frac{1}{2N}\sigma^2_{\varepsilon} \leq \frac{1}{2}\big(\sigma^2_{\varepsilon_p} + \sigma^2_{\varepsilon}\big)$$

- the sum of reconstruction errors of all N pixels connected to $\boldsymbol{p}$ and that of $\boldsymbol{p}$ is reduced. The total error becomes in this case:

$$\sigma^2_{t} = \frac{N+1}{2}\sigma^2_{\varepsilon_p} + \big(2N + \frac{1}{2N} - \frac{3}{2}\big)\sigma^2_{\varepsilon},$$

while in the previous case, it was:

$$\sigma^2_{t'} = \frac{N+1}{2}\sigma^2_{\varepsilon_p} + \frac{5N-3}{2}\sigma^2_{\varepsilon} \geq \sigma^2_{t}.$$

3.2. Unconstrained update operator

If we do not constrain the weigths to sum up to 1, then two options exist for optimizing this scheme:

- Minimize the total error on the connected pixels, by differentiating the expression (9) with respect to the different variables. This leads to

$$\alpha_i = \frac{2}{N}, \qquad \text{for all} \quad i \in \{1, \ldots, N\}. \tag{11}$$

- An improved performance can be expected by minimizing the total reconstruction error:

$$\sigma^2_{t} = \left(\frac{N+1}{2}\sum_{i=1}^{N}\alpha_i^2 + 2N - 2\sum_{i=1}^{N}\alpha_i\right)\sigma^2_{\varepsilon} + \frac{N+1}{2}\sigma^2_{\varepsilon_p}. \tag{12}$$

By deriving this expression w.r.t. each α_i, we get

$$\alpha_i = \frac{2}{N+1}, \qquad \text{for all} \quad i \in \{1, \ldots, N\}, \tag{13}$$

which is slightly different from the values obtained in (11). In particular, for $N = 1$, we get $\alpha_1 = 1$, which is identical with the previous processing of simple connected pixels. For $N = 2$ however, the weights will be $\alpha_1 = \alpha_2 = 2/3$. In general, the error on a connected pixel will be:

$$\sigma^2_{m_{i_0}} = \frac{1}{2}\sigma^2_{\varepsilon_p} + 2\left[1 - \frac{N+2}{(N+1)^2}\right]\sigma^2_{\varepsilon}. \tag{14}$$

Compared with the original situation, we have

$$2 - \frac{2(N+2)}{(N+1)^2} \geq \frac{1}{2}, \quad \forall\, N \geq 1.$$

The pixel $\boldsymbol{m}$ has therefore a larger reconstruction error than in the original case.
All the other pixels connected with $\boldsymbol{p}$ have to be compared with the error we got on $\boldsymbol{n}$ in (6):

$$2 - \frac{2(N+2)}{(N+1)^2} \leq \frac{5}{2}, \quad \forall\, N \geq 1,$$

which shows that all the other $\boldsymbol{m}_i$ have smaller reconstruction errors than before.

In the same way, the pixel $\boldsymbol{p}$ has also a smaller error than before:

$$\sigma^2_{p} = \frac{1}{2}\sigma^2_{\varepsilon_p} + \frac{2N}{(N+1)^2}\sigma^2_{\varepsilon} \;\leq\; \frac{1}{2}\big(\sigma^2_{\varepsilon_p} + \sigma^2_{\varepsilon}\big), \quad \forall\, N \geq 1.$$

Globally, the reconstruction error has been minimized by the constraint we imposed:

$$\sigma^2_{t} = \frac{N+1}{2}\sigma^2_{\varepsilon_p} + 2N\left(1 - \frac{1}{N+1}\right)\sigma^2_{\varepsilon} \leq \sigma^2_{t'}.$$

3.3. Remarks on the application of the proposed operator in non-dyadic schemes

A very similar analysis can be applied on non-dyadic lifting schemes involving uni-directional MC temporal prediction like for example, the 3-band "Haar-like" structure [7], whose temporal £ltering structure is depicted in Fig. 2.

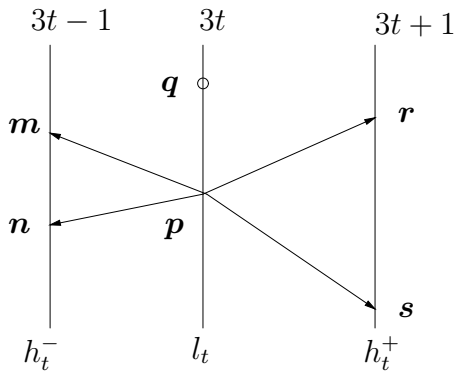

Fig. 2. Motion compensated temporal prediction in a 3-band lifting scheme.

The difference with the previous 2-band case consists of the fact that there two temporal detail subbands (h_t^+, h_t^-) and a pixel $\boldsymbol{p}$ in frame $3t$ can be used for the prediction of multiple pixels both in frame $3t-1$ (like $\boldsymbol{m}$, $\boldsymbol{n}$) and frame $3t+1$ (like $\boldsymbol{r}$ and $\boldsymbol{s}$). Therefore, the average operator can be computed either by averaging the connected pixels in each detail subband independently and then taking the mean of the two results, or by averaging all the connected pixels from the two frames.

4. EXPERIMENTAL RESULTS

We present simulation results for both the 2-band and 3-band schemes on two popular CIF sequences ("foreman" and "mobile") at 30fps, selected for their different motion characteristics. The video codec used in experiments consists in the MCTF (2-band or 3-band), followed by a spatial decomposition of the temporal subbands with a 9/7 biorthogonal multiresolution analysis. The spatio-temporal wavelet coef£cients are encoded with the MC-EZBC software [8]. The same software is used to perform the motion estimation by Hierarchical Variable Size Block Matching

(HVBSM) with 1/8th pel accuracy and motion vector arithmetic encoding.
The proposed spatio-temporal operator is compared with the equivalent system (2-band or 3-band) where the MC temporal prediction is done using the £rst connected pixel in scanning order.

Bitrate(kbs)	200	400	600	800	1500
First	29.33	33.40	35.33	36.62	39.66
Mean	29.41	33.70	35.70	36.99	39.99
NN Mean	29.38	33.67	35.60	36.94	39.94

Table 1. *RD comparison between several updating strategies in the 2-band MC £lterbank: £rst pixel used ("First"), the average ("Mean") and the non normalized mean ("NN Mean") on "foreman" CIF sequence, 30 fps.*

Bitrate(kbs)	200	400	600	800	1500
First	30.34	34.29	35.98	37.33	40.23
MMean	30.40	34.40	36.09	37.43	40.32
BMean	30.42	34.41	36.11	37.44	40.32

Table 2. *RD comparison between several updating strategies in the 3-band MC £lterbank: £rst pixel used ("First"), the average of the two values computed with connected pixels on each side ("MMean") and the global mean of all connected pixels ("BMean") on "foreman" CIF sequence, 30 fps.*

Bitrate(kbs)	200	400	600	800	1500
First	18.59	26.18	28.99	30.76	34.23
Mean	18.67	26.31	29.17	30.91	34.37
NN Mean	18.65	26.30	29.15	30.89	34.35

Table 3. *RD comparison between several updating strategies in the 2-band MC £lterbank: £rst pixel used ("First"), the average ("Mean") and the non normalized mean ("NN Mean") on "mobile" CIF sequence, 30 fps.*

One can remark a PSNR improvement of 0.2-0.3dB on "foreman" when using the updating strategy for both 2-band and 3-band MC £lterbanks, with a very small difference between normalized and non normalized average (cf. Eq. 13), and a slight improvement when using the global mean of the pixels in the 3-band case. The PSNR results are less impressive on "mobile" sequence, which has a more uniform motion, giving rise to less multiple connected pixels and thus reducing the £eld of action of the proposed update strategy. However, we observed that the visual quality in the concerned areas is improved.

Bitrate(kbs)	200	400	600	800	1500
First	22.26	28.48	30.64	32.12	35.10
MMean	22.29	28.53	30.68	32.16	35.14
BMean	22.28	28.54	30.69	32.16	35.14

Table 4. *RD comparison between several updating strategies in the 3-band MC £lterbank: £rst pixel used ("First"), the average of the two values computed with connected pixels on each side ("MMean") and the global mean of all connected pixels ("BMean") on "mobile" CIF sequence, 30 fps.*

5. CONCLUSION AND FUTURE WORK

In this paper, we have provided a new spatio-temporal update operator for lifting-based motion-compensated temporal £ltering, performing an average of multiple connected pixels. We have then proved both theoretically and by simulations its improved coding performance. Moreover, in the future, we plan to employ such spatio-temporal processing schemes for improving the resiliency of wavelet video compression schemes depending on the content characteristics. Also, various averaging techniques in combination with a multi-hypothesis approach can be employed at different spatio-temporal resolutions.

6. REFERENCES

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

[2] 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.

[3] A. Secker and D. Taubman, "Highly scalable video compression using a lifting-based 3D wavelet transform with deformable mesh motion compensation," in *Proceedings of the IEEE International Conference on Image Processing*, Oct. 2002.

[4] K. Hanke, J.-R. Ohm, and T. Rusert, "Adaptation of £lters 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.

[5] K. Hanke, "Interframe wavelet video coding with lowpass transition," doc. m8997, Shanghai MPEG meeting, Oct. 2002.

[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] 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.

[8] "3D MC-EZBC software package," available on the MPEG CVS repository.