LMS BASED ADAPTIVE PREDICTION FOR SCALABLE VIDEO CODING

B. Ugur Toreyin[*], *Maria Trocan*[**]

[*]Bilkent University
Department of Electrical and Electronics Eng.
06800, Bilkent, Ankara, Turkey
{bugur, cetin}@bilkent.edu.tr

Beatrice Pesquet-Popescu[**], *A. Enis Cetin*[*]

[**]E.N.S.T.
Signal and Image Processing Department
46, rue Barrault, 75634 Paris, France
{trocan, pesquet}@tsi.enst.fr

ABSTRACT

$3D$ video codecs have attracted recently a lot of attention, due to their compression performance comparable with that of state-of-art hybrid codecs and due to their scalability features. In this work, we propose a least mean square (LMS) based adaptive prediction for the temporal prediction step in lifting implementation. This approach improves the overall quality of the coded video, by reducing both the blocking and ghosting artefacts. Experimental results show that the video quality as well as PSNR values are greatly improved with the proposed adaptive method, especially for video sequences with large contrast between the moving objects and the background and for sequences with illumination variations.

1. INTRODUCTION

The $t+2D$ wavelet video coding schemes [1], [2, 3] are well known, as they provide spatial, temporal scalability and coding performance competitive with state-of-art codecs. Motion compensated temporal filtering (MCTF) exploits the temporal interframe redundancy by applying an open-loop temporal wavelet transform along the motion trajectories of the frames in a video sequence. The temporal subband frames are further spatially decomposed and can be encoded by different algorithms such as 3D-SPIHT [4], 3D-ESCOT [5] or MC-EZBC [6].

In general, a block-matching algorithm is used for motion estimation. Even though a bidirectional prediction and mode selection can be used, blocks artefacts are still existent. In addition, ringing artefacts appear at low bitrates and ghosting artefacts can be present as well.

In order to avoid such artefacts, motion compensation solutions such as weighted average update operator [7] or overlapped block motion compensation [8] have been proposed, alleviating but not completely solving this problem. In this paper we propose to improve the prediction of the high-frequency temporal subband frames by using an adaptive filter bank structure.

There are various subband adaptive filter structures which perform adaptive filtering in the subbands [9, 10, 11]. We use the least mean squares (LMS) type FIR based adaptive filters proposed in [12]. In [12], adaptation scheme is developed for 2-D image compression. In this work, we extend the adaptation scheme to the motion compensated $t+2D$ video coding case.

The proposed LMS based adaptive prediction method is used in the temporal prediction step in the lifting framework. Note however that it can be as well applied to any temporal prediction scheme. The detail subband frame pixels are predicted using a set of pixels from the neighbouring previous and future frames. This way, the spatio-temporal filters are adapted to better take into account the changing input conditions, in particular moving objects having high contrast with the background or illumination variations. In such cases, fixed coefficient filter structures result in poor image quality with low PSNR values. The proposed scheme substantially improve the image quality while increasing the PSNR, as the number of pixels used in the adaptation is increased. Moreover, a special edge (contrast)-sensitive adaptation methodology is developed for the two-pixel case which introduces a low-cost alternative to adaptation with larger number of pixels as in [13].

This paper is organised as follows: Section 2 describes the adaptive filter bank structure used in the prediction of the high-frequency temporal subband frames. Experimental results, obtained by varying the neighborhood size are presented in Section 3 for several test sequences. Finally, conclusions and future work are drawn in Section 4.

2. ADAPTIVE FILTER BANK STRUCTURE

Motion-compensated temporal filtering (MCTF) coding approach relies on an open-loop subband decomposition. Let us denote by x_t the original frames, t being the time index, by h_t and l_t the high-frequency (detail) and low-frequency (approximation) subband frames, respectively, and by $\mathbf{n}$ the

Part of this work was supported by the European Commission 6th Framework Programme under the grant number FP6-507752 (MUSCLE Network of Excellence).

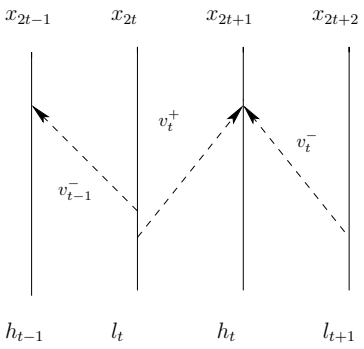

Fig. 1. Motion-compensated temporal filtering with bidirectional lifting steps.

spatial index inside a frame. For the purpose of illustration, we have used in this paper a biorthogonal 5/3 filter bank for our temporal decomposition. The temporal motion compensated filtering in this case is illustrated in Fig.1, where $\mathbf{v}_t^+(\mathbf{n})$ denotes the forward motion vector (MV) predicting the position $\mathbf{n}$ in the $2t+1$-st frame from the $2t$-th frame and $\mathbf{v}_t^-(\mathbf{n})$ denotes the backward MV predicting the same position in the $2t+1$-st frame from the $2t+2$-nd frame.

We first introduce a FIR estimator for $x_{2t+1}(\mathbf{n})$ by using for prediction a set of pixels from the neighboring $x_{2t}(\mathbf{n})$ and $x_{2t+2}(\mathbf{n})$ frames (note that no motion compensation is involved at this point in the prediction):

$$\hat{x}_{2t+1}(\mathbf{n}) = \sum_{\mathbf{k}\in\mathcal{S}} w_{2t,\mathbf{n},\mathbf{k}} x_{2t}(\mathbf{n}-\mathbf{k}) + \sum_{\mathbf{k}'\in\mathcal{S}'} w_{2t+2,\mathbf{n},\mathbf{k}} x_{2t+2}(\mathbf{n}-\mathbf{k}') \tag{1}$$

where the filter coefficients $w's$ are adaptively updated using an LMS-type algorithm [14]. In the above equation, summations are carried out over appropriate neighborhoods $\mathcal{S}$, $\mathcal{S}'$ in the $2t$-th and $2t+2$-nd image frames, respectively. The adaptive estimator for $h_t(\mathbf{n})$ is illustrated in Fig.2.

The FIR normalized LMS adaptation is performed in a conventional manner as follows both at the encoder and the decoder:

$$\hat{\mathbf{w}}(\mathbf{n}+\mathbf{1}) = \hat{\mathbf{w}}(\mathbf{n}) + \mu \frac{\tilde{\mathbf{x}}_t(\mathbf{n}) e(\mathbf{n})}{||\tilde{\mathbf{x}}_t(\mathbf{n})||^2} \tag{2}$$

where $\hat{\mathbf{w}}(\mathbf{n})$ is the filter coefficient vector at image location $\mathbf{n}$, and the vector $\tilde{\mathbf{x}}_t(\mathbf{n})$ contains the pixels within the chosen neighborhoods at the $2t$-th and $2t+2$-nd image-frames. The vector $\mathbf{1}$ represents a unit increment in the image index, in (2). The adaptive algorithm converges when the update parameter μ lies between 0 and 2. The computational cost can be reduced by omitting the normalization by the norm $||\tilde{\mathbf{x}}_\mathbf{t}(\mathbf{n})||^2$ by selecting a μ close to zero.

The detail frame $\mathbf{h_t}$ is given by

$$h_t(\mathbf{n}) = x_{2t+1}(\mathbf{n}) - \hat{x}_{2t+1}(\mathbf{n}) \tag{3}$$

and

$$e(\mathbf{n}) = h_t(\mathbf{n}) = x_{2t+1}(\mathbf{n}) - \tilde{\mathbf{x}}_t^T(\mathbf{n})\hat{\mathbf{w}}(\mathbf{n}) \tag{4}$$

The weights are initialized to the reciprocal of the number of pixels used for adaptation. For example, for 18 pixel adaptation scheme, the weights are initialized to $\frac{1}{18}$. At each iteration the weights are normalized such that their sum equals identity. If at least one of the weights turn out to be less than or equal to zero, then the weights are re-initialized.

In order to take into account the temporal filtering, as illustrated in Fig. 1, we rewrite the prediction equation (1) using the pixels matched to $\mathbf{n}$ by the motion estimation process:

$$\hat{x}_{2t+1}(\mathbf{n}) = \sum_{\mathbf{k}} w_{2t,\mathbf{n},\mathbf{k}} x_{2t}(\mathbf{n}-\mathbf{k}-\mathbf{v}_t^+(\mathbf{n})) + \sum_{\mathbf{k}} w_{2t+2,\mathbf{n},\mathbf{k}} x_{2t+2}(\mathbf{n}-\mathbf{k}-\mathbf{v}_t^-(\mathbf{n})) \tag{5}$$

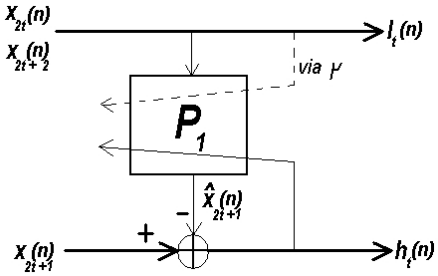

Fig. 2. Adaptive estimator.

A great flexibility for the adaptation scheme is achieved by varying the number of pixels in the selected neighborhood, as illustrated in Fig. 3. Lighter pixels in the left and right images are the corresponding motion-compensated pixels of the pixel $\mathbf{n}$ at the $2t+1$-st subband frame. The adaptation window is extended to the darker pixels to enhance the robustness of the proposed algorithm.

3. EXPERIMENTAL RESULTS AND FUTURE WORK

For our simulations, we have considered four representative test video sequences: "Foreman" (CIF, 30 Hz), "Mobile" (CIF, 30 Hz), "Harbour" (4CIF, 60 Hz) and "Crew" (4CIF, 60 Hz), which have been selected for their different motion, contrast and texture characteristics.

The tests have been made in the framework of the MSRA [15] video codec. This is a fully scalable wavelet-based video codec which supports both spatial ($t+2D$) and inband ($2D+t+2D$) temporal filtering, as well as base-layer coding options. For our simulations we have used only the $t+2D$ video coding approach. The experiments have been run for 5 temporal decomposition levels, considering motion vector-less temporal filtering for the CIF and 4CIF sequences. For comparisons, the video sequences used in our tests have been

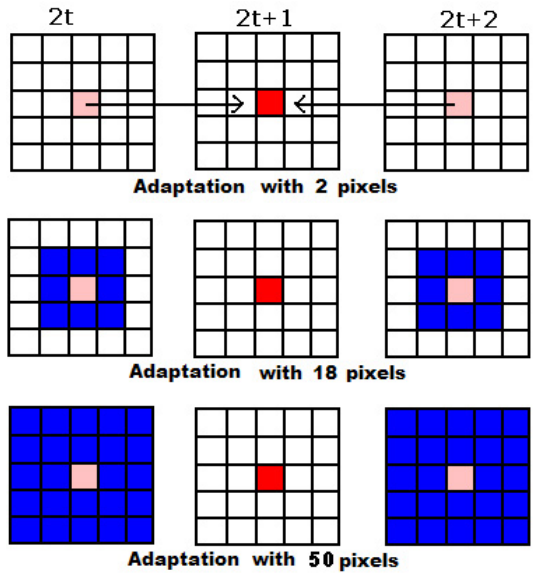

Fig. 3. Adaptation scheme with 2, 18 and 50 pixels.

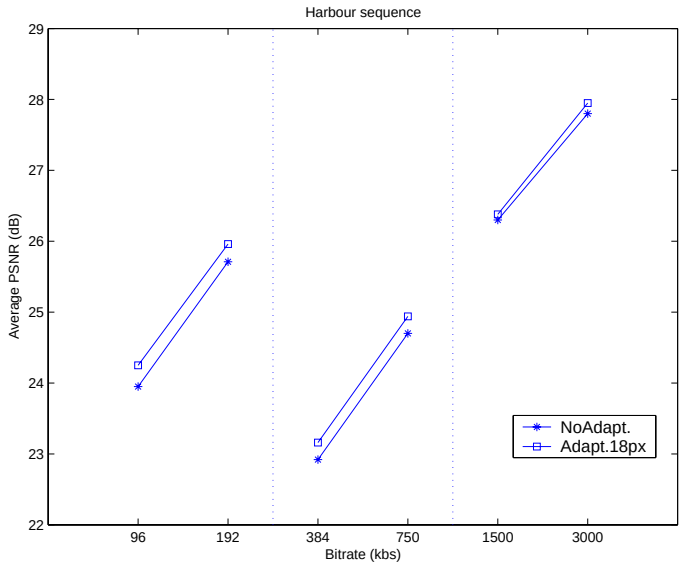

Fig. 4. Rate-distortion comparison for "Harbour" sequence.

decomposed with a 5-level 5/3 temporal decomposition. The temporal approximation subbands have been spatially decomposed over 5 levels with the biorthogonal 9/7 wavelets, for both adaptive LMS and 5/3 filtering schemes.

Rate-distortion curves for the "Harbour" and the "Crew" sequences are presented in Fig. 4 and 5, respectively. In the "Harbour" sequence, several foreground objects move while occluding with the contrasting background. Similarly, in the "Crew" sequence, sudden flashes of light reflects from the crew, resulting in a high contrast between successive frames. For these two situations, the adaptation scheme yields higher PSNR values compared to the no adaptation case.

YSNR	Mobile sequence				
	CIF			QCIF	
	30 Hz	15 Hz		15 Hz	7.5 Hz
BitRate(kbps)	384	256	128	64	48
NoAdapt.(dB)	19.57	19.11	17.92	18.78	18.78
Adapt.2px(dB)	19.65	19.11	17.94	18.91	18.94
Adapt.18px(dB)	20.16	19.57	18.30	19.41	19.35
Adapt.50px(dB)	19.79	19.14	17.92	19.14	19.04

Table 1. Rate-distortion results for "Mobile" sequence.

The PSNRs in Tables 1 and 2 are computed on slightly smaller frames to reduce the inaccuracies due to frame boundaries. There is relatively a small increase in PSNR values for the two-pixel adaptation method when compared with the no adaptation case, for the Mobile sequence. However, the quality of the image frames is substantially improved especially for those segments of the video sequences where there are edges in the image frames and high contrast between moving objects and the background (cf. Fig.6 a, b and c). Indeed, there are many edges defined by the numbers in the calendar and there is a high contrast between the black moving train and the white calendar. The ghosts around the funnel, edges of the train, and numbers in the calendar are drastically removed even with a 2 pixel adaptation. The improvement increases for larger adaptation neighborhood.

YSNR	Foreman sequence				
	CIF			QCIF	
	30 Hz	15 Hz		15 Hz	7.5 Hz
BitRate(kbps)	256	192	96	48	32
NoAdapt.(dB)	27.00	27.44	25.38	24.31	24.70
Adapt.2px(dB)	27.77	28.02	26.11	24.86	25.20
Adapt.18px(dB)	27.53	27.70	25.80	24.70	24.93
Adapt.50px(dB)	27.67	27.80	25.90	24.74	24.94

Table 2. Rate-distortion results for "Foreman" sequence.

In above simulations, we used the scalable video coding software developed Microsoft Research Asia (MSRA). This software implements the lifting structure in conventional manner, in which update step is performed after the prediction step. Due to this and quantization effects, the adaptive algorithm at the decoder has to use quantized and filtered error values. This degrades the adaptation process and the filter weights differ substantially from the filter weights in the encoder. One way to solve this problem is to swap the prediction and update steps in the lifting structure as in adaptive lifting based image coding papers [12], [13].

During the conference, we will present coding results with motion vector information.

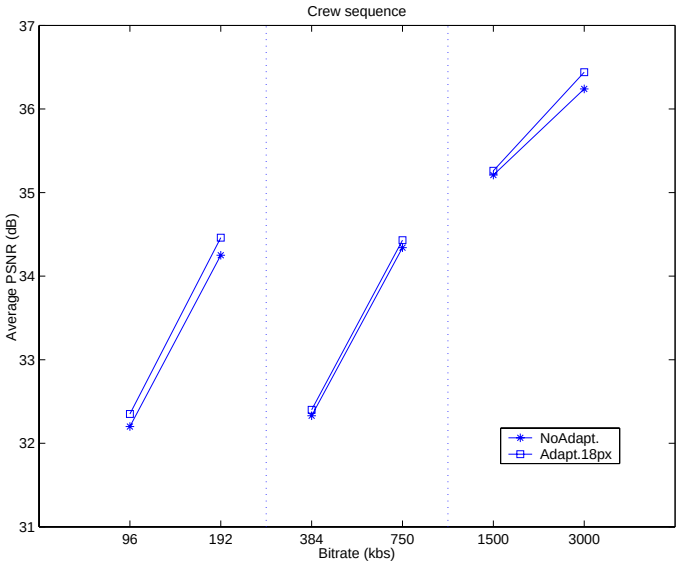

Fig. 5. Rate-distortion comparison for "Crew" sequence.

4. CONCLUSION

We have presented an LMS based adaptive prediction method and used it in the temporal prediction step for scalable video coding. The pixels of temporal detail subband frames are optimally predicted by using a set of pixels from the neighbouring subband frames. We illustrated our purpose on a bidirectional prediction scheme, but the set of pixels for adaptation can be chosen from any number of frames involved in a longer term prediction. Experimental results show that even for two-pixel adaptation case, the visual quality of the reconstructed frames is improved. A trade-off between compression efficiency and additional complexity coming from a larger adaptation window can be done, according to the targeted application. Significant PSNR improvements have been obtained for sequences with high contrast between various segments within the sequence and varying illumination conditions.

(a) original

(b) no adaptation

(c) 2 px. adaptation

Fig. 6. Detail from the Mobile (CIF, 30fps) sequence.

5. REFERENCES

[1] D. Taubman and A. Zakhor, "Multi-rate 3-D subband coding of video," *IEEE Trans. on Image Proc.*, vol. 3, pp. 572–588, 1994.

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

[3] 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.

[4] 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.

[5] S. Li, J. Xu, Z. Xiong, and Y.-Q. Zhang, "3D embedded subband coding with optimal truncation (3D-ESCOT)," *Applied and Computational Harmonic Analysis*, vol. 10, pp. 589, May 2001.

[6] S. Hsiang and J. Woods, "Embedded image coding using zeroblocks of subband/wavelet coefficients and context modeling," in *ISCAS*, Geneva, Switzerland, 2000, pp. 589–595.

[7] C. Tillier, B. Pesquet-Popescu, and M. Van der Schaar, "Weighted average spatio-temporal update operator for subband video coding," *ICIP*, Singapore,Oct. 2004.

[8] R. Xiong, X. Ji, D. Zhang, J. Xu, G. Pau, M. Trocan, S. Brangoulo, and V. Bottreau, "Vidwav wavelet video coding specifications," Tech. Rep. doc. M12339, ISO/IEC JTC1/SC29/WG11, July, 2005.

[9] S. Weiss, M. Harteneck, and R. W. Stewart, "On implementation and design of filter banks for subband adaptive systems," in *IEEE Workshop Signal Processing Systems*, Cambridge, MA, October 1998, p. 172181.

[10] S. Hosur and A. H. Tewfik, "Wavelet transform domain adaptive fir filtering," *IEEE Transactions on Signal Processing*, vol. 45, pp. 617–630, 1997.

[11] S. Attallah and M. Najim, "On the convergence enhancement of the wavelet transform based lms," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Detroit, MI, May 1995, pp. 973–976.

[12] Ö. N. Gerek and A. E. Cetin, "Adaptive polyphase subband decomposition structures for image compression," *IEEE Transactions on Image Processing*, vol. 9, pp. 1649–1659, October 2000.

[13] Ö. N. Gerek and A. E. Cetin, "Edge adaptive lifting structures for image coding," *IEEE Transactions on Image Processing*, vol. 15, January 2006.

[14] C.F.N. Cowan and P.M. Grant, *Adaptive Filters*, Prentice-Hall, Englewood Cliffs, New Jersey, 1985.

[15] "Wavelet codec reference document and software manual," MPEG document N7334, July 2005.