LINEAR AND NONLINEAR TEMPORAL PREDICTION EMPLOYING LIFTING STRUCTURES FOR SCALABLE VIDEO CODING

B. Ugur Toreyin[1], *Maria Trocan*[2], *Beatrice Pesquet-Popescu*[2] *and A. Enis Cetin*[1]

[1]Bilkent University
Department of Electrical and Electronics Eng.
06800, Bilkent, Ankara, Turkey
{bugur, cetin}@bilkent.edu.tr
[2]E.N.S.T.
Signal and Image Processing Department
46, rue Barrault, 75634 Paris, France
{trocan, pesquet}@tsi.enst.fr

ABSTRACT

Scalable 3*D* video codecs based on wavelet lifting structures have attracted recently a lot of attention, due to their compression performance comparable with that of state-of-art hybrid codecs. In this work, we propose a set of linear and nonlinear predictors for the temporal prediction step in lifting implementation. The predictor uses pixels on the motion trajectories of the frames in a window around the pixel to be predicted to improve the quality of prediction. Experimental results show that the video quality as well as PSNR values are improved with the proposed prediction method.

1. INTRODUCTION

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

Problems with the current methods include blocking and ringing artefacts in spite of the bi-directional prediction and block matching based motion estimation. In addition, ringing artefacts become more severe 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 [6] or overlapped block motion compensation [7] 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 predictor structures using pixels around the motion trajectories of the frames.

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

The proposed prediction method is used in the temporal prediction step in the lifting framework. The detail subband frame pixels are predicted from the neighbouring previous and future frames using pixels on the motion trajectories of the frames in a window around the pixel to be predicted. This way, the spatio-temporal filters become larger extent ones taking into account the neighbouring pixels around the motion trajectories. The proposed scheme improve the image quality while increasing the PSNR.

This paper is organised as follows: Section 2 describes the linear and nonlinear filter structures used in the prediction of the high-frequency temporal subband frames. Experimental results are presented in Section 3 for several test sequences. Finally, conclusions and future work are drawn in Section 4.

2. LINEAR AND NONLINEAR FILTER STRUCTURES

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 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.

Instead of the prediction filter, $\{0.5, 0.5\}$, of usual 5/3 filter bank we introduce a FIR estimator for $x_{2t+1}(\mathbf{n})$ by using 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):

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

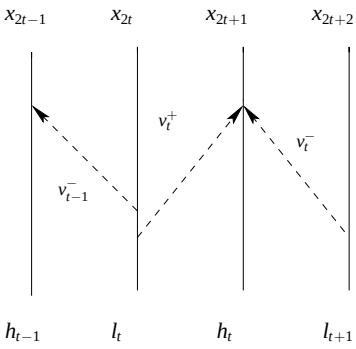

Fig. 1. Motion-compensated temporal filtering with bidirectional lifting steps.

where $w's$ are the filter coefficients which sum up to unity.

In the above equation, summations are carried out over appropriate neighborhoods $\mathscr{S}$, $\mathscr{S}'$ in the $2t$-th and $2t+2$-nd image frames, respectively, as shown in Fig. 2.

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:

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

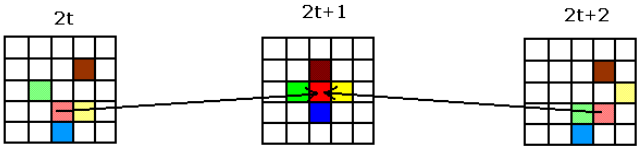

Fig. 2. Pixels utilized in linear and nonlinear temporal prediction filters.

For both the linear and nonlinear filter based prediction scheme, the weights corresponding to the pixels in the $2t$-th and $2t+2$-nd image frames on the motion trajectory of the pixel to be predicted in the detail ($2t+1$-st image) frame are assigned to be 0.45. This sums up to 0.9 for both sides. For the linear filter based prediction scheme the remaining 0.1 is equally distributed among the remaining eight pixels in the $2t$-th and $2t+2$-nd image frames on the motion trajectories of the four pixels around the pixel to be predicted. However, for the nonlinear scheme, two pixels that are closest in value to the pixels in the $2t$-th and $2t+2$-nd image frames on the motion trajectory of the pixel to be predicted in the detail frame are chosen out of these eight pixels. These two pixels share the remaining weights in the nonlinear filter based prediction scheme, as opposed to eight pixels all contribute equally in the linear filter based scheme.

3. EXPERIMENTAL RESULTS

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.

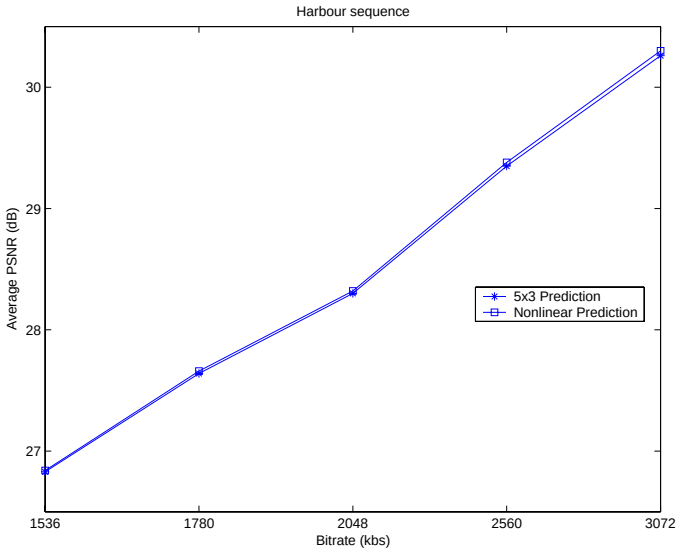

Fig. 3. Rate-distortion comparison for "Harbour" sequence.

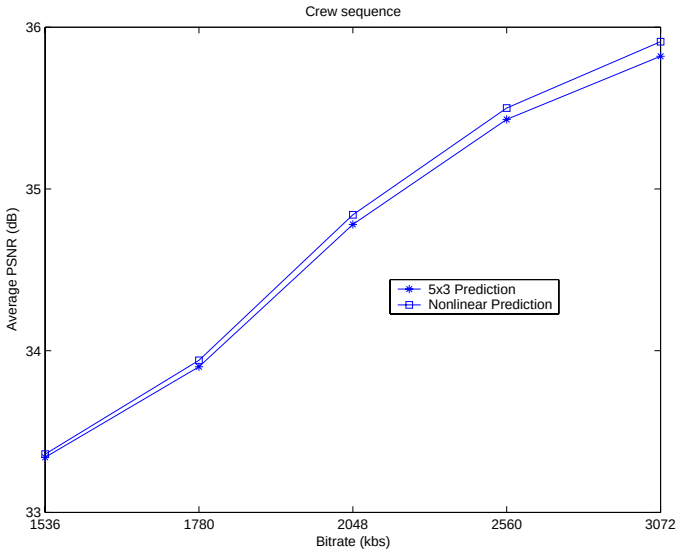

Fig. 4. Rate-distortion comparison for "Crew" sequence.

The tests have been made in the framework of the MSRA [8] 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 a single temporal decomposition level for the CIF and 4CIF sequences. The motion estimation is block-based and the motion vector fields have been estimated with $1/4^{th}$ pixel accuracy.

Rate-distortion curves for the "Harbour" and the "Crew" sequences are presented in Fig. 3 and 4, respectively. In the "Harbour" sequence, several foreground objects move while occluding with the contrasting background. 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 nonlinear filter based prediction scheme yields higher PSNR values compared to the linear filter based prediction scheme although linear filter has a larger support. Both of them produces higher PSNR values than the usual prediction filter of 5/3 filter bank.

YSNR	Mobile sequence (CIF) @30Hz				
BitRate(kbps)	384	320	256	224	192
5/3 Pred.Filt.(dB)	21.64	21.02	20.34	19.92	19.48
Nonlinear(dB)	21.65	21.03	20.34	19.93	19.49
Linear(dB)	21.65	21.03	20.34	19.92	19.48

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

YSNR	Foreman sequence (CIF) @30Hz				
BitRate(kbps)	256	224	192	160	128
5/3 Pred.Filt.(dB)	29.51	28.97	28.38	27.56	26.55
Nonlinear(dB)	29.52	28.98	28.39	27.58	26.57
Linear(dB)	29.52	28.98	28.39	27.57	26.56

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

There is relatively a small increase in PSNR values for both linear and nonlinear method when compared with the usual 5/3 filter bank prediction filter. However, the quality of the image frames is improved especially for the nonlinear filter based prediction (cf. Fig.5 a, b and c). Nonlinear filter has a larger spatial support and only the most relevant pixels are included for prediction. Hence it is more successful in eliminating blocking artefacts. Indeed, the ghosts and blocks on the face and especially around the mouth of the man are greatly reduced with nonlinear temporal prediction.

(a) original

(b) 5/3 filter bank temporal prediction filter

(c) nonlinear temporal prediction

Fig. 5. Frame excerpted from the Foreman (CIF, 30fps) sequence.

4. CONCLUSION

We have presented linear and nonlinear filter based prediction methods and used them in the temporal prediction step for scalable video coding. The pixels of temporal detail subband frames are 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 prediction can be chosen from any number of frames involved in a longer term prediction. Experimental results show that, the visual quality of the reconstructed frames is improved especially for nonlinear filter based prediction scheme. Improvements in PSNR values have also been obtained for the sequences tested.

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] 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] 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.

[4] 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.

[5] 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.

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

[7] 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.

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