MOTION-COMPENSATED TEMPORAL FILTERING WITHIN THE H.264/AVC STANDARD

Emrah Akyol[1], A.Murat Tekalp[1,2], M.Reha Civanlar[1]

[1] College of Engineering, Koc University, Istanbul, Turkey
[2] Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY 14627

ABSTRACT

We propose an adaptive motion-compensated temporal filtering (MCTF) structure to provide efficient temporal scalability within the H.264./AVC video compression standard. MCTF has traditionally been considered within fully scalable wavelet video coders. However, motion-compensated simple 5/3 lifted temporal wavelet filtering suffers at scene changes, as well as occlusion regions. We note that the bi-directional motion compensation mode in the H.264 standard is best equipped with the state of the art adaptive features such as adaptive block size, mode switching between forward, backward and bidirectional prediction and in-loop deblocking filter. Hence, we propose a GOP structure to implement block-based adaptive MCTF within the H.264 syntax using stored B-pictures, similar to the motion-compensated 5/3 wavelet filtering. We provide experimental results to compare the results of our proposed codec with those of other scalable wavelet video coders which use MCTF. It is also possible to employ the proposed adaptive MCTF structure within fully scalable wavelet video codecs.

1. INTRODUCTION

Motion-compensated simple 5/3 lifted temporal wavelet filtering has been reported as a very effective approach for building scalable video codecs in the literature [1, 2, 3]. The basic idea behind this approach is to interpolate frames from their neighboring (past and future) frames in time domain using motion compensation. Recent predictive coders such as H.264 [4] have advanced features for bidirectional prediction like adaptive block sizes, mode switching between forward, backward and bidirectional prediction, deblocking filter and intra-coded macroblocks in inter frames. These features of the predictive coders should prove to be useful in the MCTF structure to obtain better prediction and hence better compression efficiency.

Given the rich prediction feature set of the H.264 standard, we target implementing an MCTF approach within the H.264 standard to obtain an efficient, easy to produce and effective layering scheme without modifying the standard. Achieving this target is possible by using the stored B pictures which can be used as reference frames for other pictures in H.264. The rest of the implementation is in designing appropriate prediction configurations in the group of pictures (GOP) used.

We compared our results with other scalable video codecs based on MC lifting with invertible motion vectors [2], block based motion estimation without update step [3] and block based motion compensation with update step using inaccurate motion vectors [1].

This paper is organized as follows: Motion compensated lifted 5/3 wavelet temporal filtering is presented in Section 2. Motion compensated temporal filtering in H.264 standard is described in Section 3. Experimental results and comparisons are presented in Section 4. Conclusions are drawn in Section 5.

2. MOTION-COMPENSATED LIFTED 5/3 WAVELET TEMPORAL FILTERING

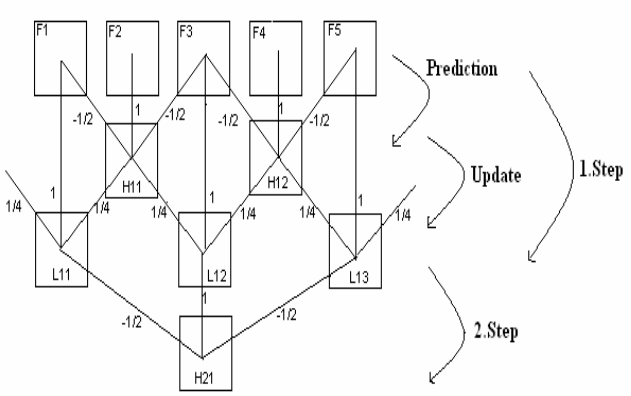

Figure 1: 5-3 lifting scheme for GOP=4

Motion compensated lifted temporal wavelet filtering performs temporal biorthogonal wavelet transform on frames using lifting (that is, prediction) and update steps. Although many biorthogonal wavelet kernels can be used, 5/3 wavelet kernel is reported to have the best experimental performance [1]. Implementation of the motion compensated lifting scheme with 5/3 filters for a

GOP size of 4 frames is shown in Fig. 1. In the prediction step, frames are predicted from their nearest neighbors using motion compensation. In the update step, the reference frames are temporally filtered to prevent aliasing due to subsampling. Motion compensation is also used in the update step, but the direction is reversed. First frame of every GOP, which is intra coded, and the prediction errors are then encoded usually using spatial wavelets. In the encoding part, original frames rather than decoded frames are used.

In this approach, existence of significant motion does not affect the compression performance when motion is compensated effectively. If the motion field used in the motion-compensated lifting step is invertible, the update step does not require new motion vectors. Since sending second set of motion vectors would be very costly, the update step can be performed with the inverse of the motion vectors obtained in the prediction step. However, when the motion is not invertible, the motion vectors will not be correct, significantly deteriorating the compression performance.

When mesh based motion estimation is used in the lifting step, the motion vectors for the update step can be obtained by straightforward inversion [2]. However one-to-one prediction fails when uncovered areas appear in the video sequence. Block based motion estimation is not invertible so the update step is either performed with motion vectors inaccurately obtained by inverting the motion vectors in the lifting step [1] or is not performed at all [3]. Bypassing the update step is reported to achieve better compression performance than using non-exact motion vectors in the literature [3]. Although this trade off is still under investigation of MPEG SVC group in their core experiments, we skipped the update step because MCTF without update can be implemented in H.264 without changing the syntax and has other advantages such as low delay for real time applications and, more importantly, achieving higher quality at temporal sub layers. In our scheme a temporal layer is encoded/decoded independently from higher temporal layers, so, no drift occurs in lower bands when higher temporal frames are unavailable. We will evaluate sublayer quality in the fourth section.

5-3 kernel requires bidirectional motion compensation. In classical lifting scheme, every predicted frame is computed as the average of forward and backward predictions. This averaging results in worse prediction than only forward or backward prediction especially when a scene change occurs in a group of pictures. Adaptive mode switching between forward, backward and bidirectional prediction can make this scheme to avoid such problems. Deblocking filter decreases the blockiness of the prediction which is an inherent problem of block based motion estimation. Adaptive macroblock size significantly increase the motion compensation prediction quality. **The fact that all of these advanced motion compensation features are part of the H.264 syntax motivates us to implement an MCTF structure within the H.264 standard.**

3. MOTION-COMPENSATED TEMPORAL FILTERING IN THE H.264 CODEC

In our configuration, we encode the first frame of a GOP as an intra frame and all others as B frames as shown in Fig. 2 for a GOP consisting of eight frames.

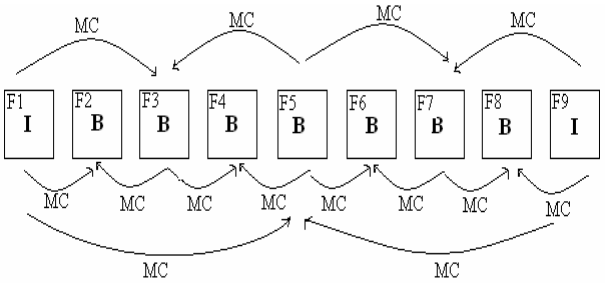

Figure 2: H.264/AVC configuration of lifting scheme with GOP=8

Here, frame **F5** is coded as a B frame estimated from frames **F1** and **F9**. **F5** is used as one of the reference frames for frames **F3** and **F7** which are coded as B pictures also. Frames **F2**, **F4**, **F6**, **F8** are encoded as B frames with reference to neighboring I or B frames. The H.264 syntax permits the use of B frames as reference frames with the feature called *stored B-pictures* [4].

This scheme provides adaptive multilayer temporal scalability to the H.264 compression standard. As an example, for the above configuration, **F1**, **F5** and **F9** are members of the base layer. **F3** and **F7** constitute the first enhancement layer and, the rest of the frames constitute the second enhancement layer, thus providing a three layer bitstream. In [5], it is reported that an H.264 bitstream with missing B pictures should be decodable by a standard decoder. Hence, the described approach produces compliant bitstreams as long as the bits corresponding to the intermediate frames are inserted into their correct place before decoding. The coding efficiency may be less than the ideal case where the best reference frame is used for each prediction but, this loss can be reduced by appropriate selection of the GOPs. Flexibility in choosing the number of layers in every GOP provides better adaptation to varying bitrates.

4. COMPARATIVE RESULTS

We compared the results of our approach with other MCTF based scalable video coders. In [2], mesh based motion estimation is used in bidirectional motion compensation. In [1], the inverse of forward motion vectors are used as the backward motion vectors required for update step. In [3], the update step is not implemented. MC_EZBC is the motion compensated embedded zero block video coder which uses Haar wavelet as lifting kernel [6]. MC_EZBC was also the reference scalable video coder of MPEG. Although MC_EZBC does not implement 5-3 filter, we included it in our comparisons for the sake of completeness.

The results show that our configuration in H.264 outperforms other scalable video coders based on MCTF. Of course, other MCTF implementations may support SNR and spatial scalability, while the proposed H.264 based encoder supports only temporal scalability. This difference is mainly caused by the advanced motion compensated prediction features of the H.264 standard. The PSNR difference gets larger as the bitrate increases in the comparison with other block based 5-3 lifted wavelet video coders. This can be explained by the amount of side information that encoders send. In H.264 standard, the side information costs more because of several encoding modes, flexible macroblock sizes and motion vectors with 1/4 pixel accuracy. Other scalable coders that we used to compare our results, however, have much lighter side information, reducing their overheads.

Block based without update step [1] vs. MCTF in H.264

Foreman QCIF
Block based without update

Bitrate(kbps)	105.56	181.59	333.66
PSNR	31.52	34.33	37.34

MCTF in H.264

Bitrate(kbps)	101.3	155.2	268.7
PSNR	33.16	35.98	39.36

Block based with update step with inaccurate motion vectors[3] vs. MCTF in H.264

Mobile QCIF :
Block based with update step with inaccurate motion vectors

Bitrate(kbps)	550.0	850.0	1100.0
PSNR	~32.50	~34.50	~36.00

(These results are taken from a graph)
MCTF in H.264

Bitrate(kbps)	474.9	815.5	1063.7
PSNR	33.99	37.25	39.08

Mesh based MCTF vs. MCTF in H.264

Football SIF :
Mesh based MCTF [2]

Bitrate(kbps)	500	10000
PSNR	25.32	28.33

MCTF in H.264

Bitrate(kbps)	486.2	980.11
PSNR	26.52	29.74

MC_EZBC vs. MCTF in H.264

Foreman QCIF :
MC_EZBC

Bitrate(kbps)	108.32	181.30	321.31
PSNR	30.83	34.31	37.73

MCTF in H.264

Bitrate(kbps)	101.3	155.2	268.7
PSNR	33.16	35.98	39.36

We also compared our configuration with the latest MPEG SVC group core experiments' reference codec which implements H.264 like features such as adaptive block size, rate distortion optimization and also uses Barbell lifting for temporal decomposition [7]. This codec implements the update step with inexact inverse motion vectors. For a fair comparison we run Barbell codec with only temporal scalability mode with 3 layers. Since this codec utilizes context adaptive binary arithmetic coding in the entropy coding stage, we set the CABAC configuration in the H.264 coder. Barbell codec forms temporal decomposition 'on the fly' to avoid boundary effects [8]. We set GOP=16 to avoid GOP boundary effects as much as possible. The optimum configuration would be to place I frames only at the scene boundaries but memory constraints will effect the implementation.

Also staying in H.264 syntax forces us to use fixed GOP size.

Barbell lifting based vs. MCTF in H.264

Foreman CIF :
Barbell lifting based

Bitrate(kbps)	496.1	577.9	665.2
PSNR	36.42	36.99	37.52

MCTF in H.264

Bitrate(kbps)	489,8	569.5	652.0
PSNR	36.88	37.55	38.09

We also compared corresponding temporal sub layers. We used corresponding original frames in the PSNR calculation.

Low temporal layer comparison

Foreman CIF :
Barbell lifting based

Bitrate(kbps)	394.1	440.3	501.8
PSNR	36.13	36.55	37.07

MCTF in H.264

Bitrate(kbps)	377.2	431.5	497.0
PSNR	36.15	36.82	37.34

5. CONCLUSIONS AND FUTURE WORK

We propose an implementation of the MCTF structure in the H.264 framework to provide adaptive multilayer temporal scalability within the H.264 standard. Our results show that by utilizing H.264 standard's advanced features for motion compensation, we can achieve better compression performance.

Our results may be used as a benchmark for better motion compensated prediction in temporal lifting schemes. Flierl used multi-hypothesis motion compensated prediction in lifted wavelet based video coding framework for this purpose [9]. Multi-hypothesis motion compensated prediction is one of the features that H.264/AVC already utilizes although we did not use it in order to compare our results with those of the lifting scheme. If B pictures are not restricted to select the reference frames that lifting structure selects, it would be analogous to Flier's work in H.264 framework.

As future work, we may employ JPEG2000 spatial wavelets and arithmetic coder implementation [10] instead of the block transform and UVLC to achieve spatial and SNR scalability in addition to temporal scalability using the proposed adaptive motion compensated temporal filtering.

6. REFERENCES

[1] M.Flierl and B.Girod, "Investigation of motion compensated lifted wavelet transforms" in Proceedings of the IEEE International Conference on Image Processing, 2,pp1029-1032 (Thessaloniki, Greece), Oct.2001

[2] 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, Vol. 3, Rochester, NY, 2002, pp. 749–752.

[3] L. Luo, J. Li, S. Li,Z. Zhuang, and Y. Zhang., "Motion compensated lifting wavelet and its application in video coding," in Proceedings of the IEEE International Conference on Multimedia and Expo 2001, Tokyo, Japan, August 2001, pp. 481-484.

[4] T. Wiegand, G. Sullivan, and A. Luthra, "Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496 -10 AVC)," May 27, 2003.

[5] Lukasz Blaszak, Marek Domański, Sławomir Maćkowiak "Spatio-Temporal Scalability in AVC codecs," ISO/IEC JTC1/SC29/WG11 MPEG2003/M9469 Pattaya, March 2003

[6] P. Chen and J. W. Woods, Improved MC-EZBC with quarter-pixel motion vectors, ISO/IEC JTC1/SC29/WG11, MPEG2002/m8366, Fairfax, May 2002.

[7] J. Xu, R.Xiong,B.Feng,G.Sullivan,M.Lee,F.Wu,S.Li, "3D Sub-band Video Coding using Barbell lifting," ISO/IEC JTC/WG11 M10569, S05

[8] Jizheng Xu; Zixiang Xiong; Shipeng Li; Ya-Qin Zhang "Memory-constrained 3D wavelet transform for video coding without boundary effects", IEEE Trans. Circuits and Systems for Video Technology Volume: 12 Issue: 9 , Sept. 2002 Page(s): 812 -818

[9] M.Flierl, "Video Coding with Lifted Wavelet Transforms and Complementary Motion-Compensated Signals," SPIE Conference on Visual Communications and Image Processing, San Jose, CA, Jan.2004.

[10] D.S.Taubman and W.Marcellin, "JPEG2000: Image Compression Fundamentals, Standards and Practice," Kluwer Academic Publishers, Boston 2002.