Architectural Aspects of Scalable Wavelet Video Coding

Fabio Verdicchio, Yiannis Andreopoulos, Adrian Munteanu, Jan Cornelis and Peter Schelkens

***Abstract*— Recently, wavelet-based fully-scalable video coding has received increased attention from the research and standardization communities [1]. This is mainly due to new and efficient techniques for open-loop video coding based on motion-compensated temporal filtering (MCTF). In this paper, we review the evolution of video coding from the architectural perspective and highlight the different trade-offs that are expected to consist design challenges for future optimized implementations of scalable video coders.**

***Index Terms*— scalable video coding, motion compensated temporal filtering, adaptive lifting, in-band motion estimation and compensation, shift invariance, overcomplete discrete wavelet transforms.**

I. INTRODUCTION

VIDEO TRANSMISSION over variable-bandwidth or IP-based networks requires instantaneous bitrate adaptation of the compressed bitstream to provide an acceptable decoding quality. For that purpose, recent developments in video coding aim at providing a fully-embedded bitstream with seamless adaptation capabilities in bitrate, frame-rate and resolution level. This is achieved with techniques that are based on motion-compensated temporal filtering (MCTF) [2] [3], which essentially perform 3-D wavelet-based coding with motion compensation. The recent MPEG activity in this area has revealed a large number of applications [4], ranging from video surveillance to MPEG 21 DIA-based adaptation.

Motion-compensated (MC) 3-D transforms for video coding have been originally proposed by Kronander [5] [6]. The idea was further refined by Ohm [2], who considered open-loop systems with block-based, integer-pel, motion estimation and compensation. For such open-loop systems, Taubman and Zakhor [7] identified the possibility for seamless bitrate scalability with the use of embedded (layered) coding strategies. Pesquet-Popescu and Bottreau [8] demonstrated improved performance for MC 3-D transforms with the use of lifting-based wavelet decompositions that allow for full adaptability in the selection of reference pictures, MC mode selection, and advanced motion models for the motion estimation. Further-extended methods and results along these lines have been shown by Turaga and Van der Schaar [9], Secker and Taubman [10] and Chen and Woods [3]. These approaches have been termed *t*+2D or spatial-domain MCTF techniques due to the application-order of the transform decomposition (temporal and then spatial).

An alternative design for the MC 3-D transforms was recently proposed by Andreopoulos et. al. [11] [12]; it consists the 2D+*t* or in-band MCTF approach since the spatial transform precedes the temporal filtering. As a result, the application of temporal prediction and temporal update of the lifting decomposition occurs in the wavelet-domain. To address the aspect of shift-variance, characteristic of any critically-sampled wavelet decomposition, a complete-to-overcomplete discrete wavelet transform is performed [13] [14] [15]. The 2D+*t* approach presents the potential advantage of adaptive tuning of the lifting decomposition *across* resolution levels according to different criteria for complexity, coding efficiency and scalability, something that is not possible with the conventional *t*+2D approaches.

In this work, we review the architectures for fully-scalable video coding based on MCTF. Furthermore, we present the different trade-offs that one can investigate in such systems both from the algorithmic and the implementation perspective.

II. MOTION COMPENSATED TEMPORAL FILTERING

We begin by reviewing the new open-loop video coding schemes that perform a temporal decomposition using temporal filtering. Both the spatial-domain (*t*+2D) and in-band (2D+*t*) approaches are presented.

A. Spatial-Domain Motion Compensated Temporal Filtering

To address the issues of robust adaptation of the compressed video content to transmission conditions, several proposals suggested an open-loop system, depicted in Figure 1, which incorporates a recursive temporal filtering. This can be perceived as a temporal wavelet transform with motion compensation [2], i.e. *motion-compensated temporal filtering*. Similar to the polyphase separation of the conventional lifting-based transform [16], this scheme begins with a separation of the input into even and odd temporal frames (temporal split). Then the temporal predictor performs motion estimation and motion compensation to match the information of frame A_{2t+1} with the information present in frame A_{2t}. Subsequently the update step inverts the information of the prediction error back to frame A_{2t}, thereby producing, for each pair of input frames, an error frame H_t and an updated frame L_t. The update operator performs either MC using the inverse vector set produced by the predictor [8], or generates a new vector

This work was supported in part by the Federal Office for Scientific, Technical and Cultural Affairs (IAP Phase V - Mobile Multimedia) and by the European Community under the IST Program (Mascot, IST-2000-26467). P. Schelkens has a post-doctoral fellowship with the Fund for Scientific Research -Flanders (FWO), Egmontstraat 5, B-1000 Brussels, Belgium.

The authors are with the Vrije Universiteit Brussel, Dept. of Electronics and Information Processing (ETRO), Pleinlaan 2, B1050, Brussels, Belgium, tel: +32-2-629-3951; fax: +32-2-629-2883; e-mail:

{ fverdicc, yandreop , acmuntea, jpcornel, pschelke } @etro.vub.ac.be

set by backward ME [10]. The process recursively iterates on the L_t frames, which are now at half temporal-sampling rate (following the multilevel operation of the conventional lifting), thereby forming a hierarchy of *temporal levels* for the input video. The decoder performs the mirror operation of the scheme depicted in Figure 1: operating from right to left, the signs of operators $\mathcal{P}$, $\mathcal{U}$ are inverted and a temporal merging occurs to join the reconstructed frames. As a result, having performed the reconstruction of the L_t, denoted by $\widetilde{L}_t$, at the decoder we have:

$$\widetilde{A}_{2t} = \widetilde{L}_t - \mathcal{U}\mathcal{T}_S^{-1}\mathcal{Q}_S^{-1}C_t,\ \widetilde{A}_{2t+1} = \mathcal{P}\widetilde{A}_{2t} + \mathcal{T}_S^{-1}\mathcal{Q}_S^{-1}C_t, \quad (1)$$

where $\widetilde{A}_{2t}$, $\widetilde{A}_{2t+1}$ denote the reconstructed frames at time instants $2t$, $2t+1$. As seen from (1), even if $C_t \neq \mathcal{Q}_S\mathcal{T}_S H_t$ in the decoder (i.e. the compressed information is incompletely received or received with errors), the error affects locally the reconstructed frames $\widetilde{A}_{2t}$, $\widetilde{A}_{2t+1}$ and does not propagate linearly in time over the reconstructed video. Error-propagation may occur only across the temporal levels through the reconstructed $\widetilde{L}_t$ frames. Upon completion of the temporal decomposition, embedded coding may be applied in each GOP by prioritizing the information of the higher temporal levels based on a dyadic-scaling framework, i.e. following the same principle of prioritization of information used in wavelet-based SNR-scalable image coding [7]. Hence, the effect of error propagation in the temporal pyramid is limited and seamless video-quality adaptation occurs during the process of bitrate adjustment for SNR scalability [7] [2]. In fact, experimental results obtained with SNR-scalable MCTF video coders suggest that this coding architecture can be comparable or superior in rate-distortion sense to an optimized non-scalable coder that uses the closed-loop structure [3].

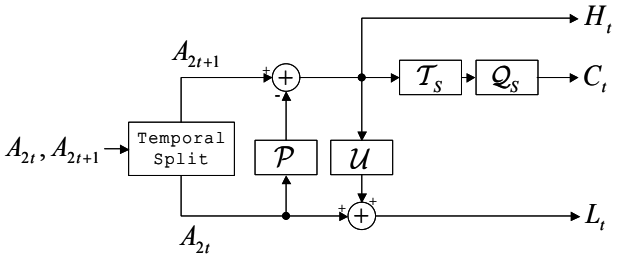

Figure 1: Motion-compensated temporal filtering. Notations: A_τ consists the input video frame at time instant $\tau = 0,1,\cdots,2t,2t+1$; H_t consists the error frame and L_t is the updated frame; C_t denotes the transformed and quantized error frame using the spatial operators $\mathcal{T}_S$ and $\mathcal{Q}_S$ respectively; $\mathcal{P}$ denotes temporal prediction, while $\mathcal{U}$ denotes the temporal update.

B. In-Band Motion-Compensated Temporal Filtering

In this section, we present a modification of the conventional MCTF video coding architecture that allows for independent temporal filtering operations across different resolutions of the video content. This may be a desirable functionality for MCTF since, in this way, all the advanced features discussed previously may be applicable with different configurations for each resolution of the input video. For example, different update and predict operators may be applied for each resolution, thereby allowing for additional levels of optimization or complexity reduction. In addition, since the multiresolution MCTF permits the complete decoupling of the various decodable resolutions, the use of different temporal decompositions and a variable number of temporal levels for each resolution becomes possible. This creates an additional degree of freedom for compact scalable video representations across resolution.

In general, a multiresolution MCTF is achievable if the $\mathcal{T}_S$ operator is a multiresolution discrete wavelet transform and the process of temporal filtering occurs in-band, i.e. after the spatial analysis of the input video frames by the DWT. Such a scheme is shown in Figure 2. In the proposed architecture, first a spatial transform $\mathcal{T}_S^l$ splits the input video into a discrete set of resolutions l, $1 \le l \le k$: for each resolution, the process of temporal splitting separates the subbands of the input frames and the prediction and update operations are performed in the wavelet domain. Since the critically-sampled wavelet transform (complete DWT) is a shift-variant representation, it is not suitable for efficient performance of in-band prediction [17] [18] [14], hence the operator $\mathcal{S}_S^l$ is with the subbands of each resolution l in order to construct the overcomplete wavelet representation of the reference frame $\mathcal{T}_S^l A_{2t}$ (CODWT). Nevertheless, the predicted subbands of resolution l remain critically-sampled: as a result, the subsequently-produced error-frame subbands are critically-sampled as well, similar to the spatial domain approach. The process then continues with the performance of the update step. Note that, in the case of an update step using backward ME, one would need to perform the CODWT to $\mathcal{T}_S^l H_t$ before the application of temporal update. However, since this type of update is not usually selected in practical implementations [3] [8], we do not elaborate on this option.

Decoding occurs following the principle of inverse MCTF, i.e. for each resolution level l, $1 \le l \le k$, the structure of Figure 2 is inverted by operating from right to left, inverting the sign of both predict and update operators, and performing a temporal merging. When the necessary number of resolutions is collected, the inverse transform performs the synthesis of the accumulated set of subbands to the spatial-domain representation.

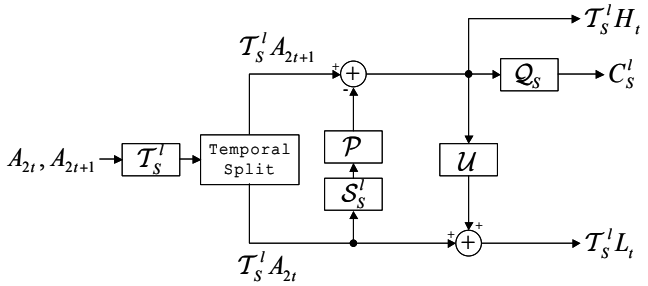

Figure 2. In-Band Motion Compensated Temporal Filtering. A_τ represents the input video frame at time instant $\tau = 0,1,\cdots,2t,2t+1$, H_t is the error frame, while L_t consists the updated frame; C_t denotes the transformed and quantized error frame; $\mathcal{T}_S^l W$ denotes the resolution level l of the DWT of frame W, $W = \{A, L, H\}$; $\mathcal{S}_S^l$ is the CODWT of resolution level l; $\mathcal{P}$ denotes temporal prediction, while $\mathcal{U}$ denotes the temporal update.

III. ALGORITHMIC AND IMPLEMENTATION PERSPECTIVES

A. Algorithmic Extensions and Capabilities of the MCTF Structure

Similar to the extensions that have been proposed for classical hybrid video coding structure, that allow for improved functionality and higher coding efficiency, relevant work was performed recently on MCTF-based video coding. For instance, newly proposed MCTF structures [8] allow for adaptive temporal splitting operators that can process the input in sets of frames that are larger

than two in order to allow for non-dyadic temporal decompositions. Similar to the conventional lifting [16], more complex series of predict-and-update steps may be envisaged thereby leading to longer temporal filters for MCTF [8]; on the other hand, temporal filtering may be performed even without an update operator [9]. This may be necessary in order to reduce visual artifacts that occur in the L-frames due to the poor prediction performance of the commonly-employed block-based ME methods. To this end, several proposals attempt to improve the prediction performance in MCTF, based on bidirectional ME and variable block-sizes [3], or multihypothesis prediction [19] i.e. by incorporating in the MCTF some of the advanced prediction tools proposed for the hybrid video-coders.

B. Implementation Aspects

By considering only the prediction step of Figure 1, it can be seen that, for each group-of-pictures (GOP), the number of motion estimations required is $\sum_{t=1}^{T}\left(N \cdot 2^{-t}\right)$ for a GOP with N frames, which is similar to a predictive (closed-loop) scheme. The complexity for each motion estimation depends on the number of reference frames used and the specific algorithm (variable block sizes, multihypothesis, etc.). For example, if two reference frames are used, from the complexity point of view this corresponds to the use of only I and B-frames within each GOP in a predictive framework.

In addition, in the in-band MCTF, Figure 2, each motion estimation is performed in the wavelet-domain which means that:

- Coefficients with a larger dynamic range than the image coefficients are used. As a result, a representation of 2 bytes or higher is used in the motion estimation algorithm
- A multiresolution motion estimation algorithm is used. This contains a class of algorithms for subband-by-subband or level-by-level or tree-by-tree motion estimation [20] that can potentially allow for a higher degree of optimization with the expense of some increased complexity.
- The suitable representation for the motion estimation has to be constructed (CODWT). The reader is referred to [13] [14] [15] [18] for further details on the complexity issues involving this construction.

Concerning the update step, if the inverse set of motion vectors produced by the prediction is used, the complexity requirement corresponds to roughly one additional motion compensation operation per frame.

With respect to the required delay, the reader is referred to [21] for further analysis on this issue. It can be generally shown that without the use of an update step, the required codec delay is comparable to the delay of a classical predictive scheme. However, the use of the update makes delay requirement approach the size of a GOP (N frames) [21].

IV. CONCLUSIONS

The new framework for open-loop fully-scalable video coding based on MCTF was presented in this paper. The architectures of its two instantiations, spatial-domain and in-band MCTF, were presented and algorithmic and implementation topics were discussed.

REFERENCES

[1] *ISO/IEC JTC1/SC29/WG11*, *n5559*, "Call for evidence on scalable video coding advances", March 2003.

[2] J. R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Trans. Image Processing*, vol. 3, no. 5, pp. 559-571, Sept. 1994.

[3] P. Chen and J. W. Woods, "Bidirectional MC-EZBC with lifting implementation," *IEEE Trans. Circ. and Syst. for Video Tech.*, to appear.

[4] *ISO/IEC JTC1/SC29/WG11*, *n5540*, "Applications and requirements for scalable video coding", March 2003.

[5] T. Kronander, "Motion compensated 3-dimensional wave-form image coding," *Proc. International Conf. on Accoustics Speech and Signal Proc.*, ICASSP 1989, Glasgow, UK, vol. 3, pp. 1921-1924, May 1989.

[6] T. Kronander, "New results on 3-dimensional motion-compensated subband coding, *Proc. Picture Coding Symposium*, PCS 1990, Cambridge MA, USA, pp. 8-10.

[7] D. Taubman and A. Zakhor, "Multirate 3-D subband coding of video," *IEEE Trans. Image Processing*, vol. 3, pp. 572-588, Sept. 1994.

[8] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," *Proc. International Conf. on Accoustics Speech and Signal Proc.*, ICASSP 2001, Salt Lake City, Utah, US, vol. 3, pp. 1793-1796, May 2001.

[9] D. Turaga and M. van der Schaar, "Wavelet coding for video streaming using new unconstrained motion compensated temporal filtering," *Proc. Internat. Workshop on Dig. Communications: Advanced Methods for Multimedia Signal Processing*, Capri, IT, pp. 41-48, Sept. 2002.

[10] A. Secker and D. Taubman, "Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting," *Proc. International Conf. on Image Proc.*, ICIP 2001, Thessaloniki, GR, vol. 2, pp. 1029-1032, Oct. 2001.

[11] Y. Andreopoulos, M. van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens and J. Cornelis, "Open-loop, in-band motion-compensated temporal filtering for objective full-scalability in wavelet video coding," ISO/IEC JTC1/SC29/WG11, m9026, MPEG 62nd meeting, Shanghai, China, October 2002.

[12] Y. Andreopoulos, M. Van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens and J. Cornelis, "Fully-scalable wavelet video coding using in-band motion compensated temporal filtering," *Proc. International Conf. on Accoustics Speech and Signal Proc.*, ICASSP 2003, Hong-Kong, CN, vol. 3, pp. 417-420, March 2003.

[13] Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens and J. Cornelis, "A new method for complete-to-overcomplete discrete wavelet transforms," *Proc. International Conference on Digital Signal Processing*, Santorini, GR, pp. 501-504, July 2002.

[14] G. Van der Auwera, A. Munteanu, P. Schelkens and J. Cornelis, "Bottom-up motion compensated prediction in the wavelet domain for spatially-scalable video", *IEE Electronics Letters*, vol. 38, no. 21, pp. 1251-1253, Oct. 2002.

[15] X. Li, "New results of phase-shifting in the wavelet space," *IEEE Signal Processing Letters*, vol. 10, no 7, pp. 193-195, July 2003.

[16] I. Daubechies and W. Sweldens, "Factoring wavelet transforms into lifting steps," Journal of Fourier Analysis and Applications, vol. 4, no. 3, pp. 247-269, March 1998.

[17] X. Li, L. Kerofski and S. Lei, "All-phase motion compensated prediction in the wavelet domain for high performance video coding," *Proc. IEEE Int. Conf. on Image Processing*, Thessaloniki, GR, vol. 3,pp. 538-541 Oct. 2001.

[18] Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens and J. Cornelis "Wavelet-based fully-scalable video coding with in-band prediction," *Proc. IEEE Benelux Signal Processing Symposium*, SPS'02, Leuven, BE, pp. 217-220, March 2002.

[19] Y. Andreopoulos, J. Barbarien, F. Verdicchio, A. Munteanu, M. Van der Schaar, J. Cornelis and P. Schelkens, "Response to call for evidence on scalable video coding," *ISO/IEC JTC1/SC29/WG11*, *m9911*, MPEG 65th meeting, Trondheim, Norway, July 2003.

[20] H. S. Kim and H. W. Park, "Wavelet-based moving-picture coding using shift-invariant motion estimation in wavelet domain," *Signal Proc.*: *Image Commun.*, vol. 16, pp. 669-679, Mar. 2001.

[21] J.-R. Ohm, "Complexity and delay analysis of MCTF interframe wavelet structures," *ISO/IEC JTC1/SC29/WG11*, *m8520*, MPEG 61st meeting, Klagenfurt, Austria, July 2002.