A comparative study of scalable video coding schemes utilizing wavelet technology

Peter Schelkens[a*], Yiannis Andreopoulos[a], Joeri Barbarien[a], Tom Clerckx[a], Fabio Verdicchio[a],
Adrian Munteanu[a], Mihaela van der Schaar[b]
[a]Vrije Universiteit Brussel-IMEC, Dept. of Electronics and Information Processing (ETRO),
Pleinlaan 2, B-1050 Brussels, Belgium
[b]University of California Davis, Dept. of Electrical and Computer Engineering
One Shields Avenue, 3129 Kemper Hall, Davis, CA 95616-5294, USA

ABSTRACT

Video transmission over variable-bandwidth networks requires instantaneous bit-rate adaptation at the server site to provide an acceptable decoding quality. For this purpose, recent developments in video coding aim at providing a fully embedded bit-stream with seamless adaptation capabilities in bit-rate, frame-rate and resolution. A new promising technology in this context is wavelet-based video coding. Wavelets have already demonstrated their potential for quality and resolution scalability in still-image coding. This led to the investigation of various schemes for the compression of video, exploiting similar principles to generate embedded bit-streams. In this paper we present scalable wavelet-based video-coding technology with competitive rate-distortion behavior compared to standardized non-scalable technology.

Keywords: Scalable wavelet video coding, motion compensated temporal filtering, in-band wavelet video coding, MPEG-AVC.

1. INTRODUCTION

The increasing demand for multimedia over networks and the heterogeneous profile of today's playback devices (from low-resolution portable devices to HDTV platforms) and networks imposes the need for scalable video coding. In practice, scalability implies the ability to decompress to different quality, resolution and frame-rate layers from subsets of a single compressed bit-stream.

Recently, the MPEG-committee has started an exploration of new technologies to support scalable video coding while providing competitive rate-distortion performance compared to single-layer coding schemes (i.e. MPEG-2,4, H.26L, MPEG-AVC). Solutions proposed based on existing standards using the hybrid coding architecture, on which those standards are based (e.g. MPEG-4 Fine Grain Scalability-FGS framework), have never met this criterion in a satisfactory way[1] (see section 2). Hence, to meet this target a specific effort is necessary and new available technologies have to be involved.

A new promising technology is wavelet-based video coding. Wavelets have already demonstrated their potential for quality and resolution scalability in still-image coding – see for instance JPEG2000[2, 3]. This led to the investigation of various schemes for the compression of video, exploiting similar principles to generate embedded bit-streams: in-band wavelet video coding, video coding based on motion compensated temporal filtering (MCTF) and in-band MCTF video coding. These schemes will be described shortly in section 3.

In section 4, we investigate the rate-distortion performance of these different approaches. In comparison to non-scalable MPEG-4 coding technology, it will be shown that wavelet-based video codecs yield excellent coding results, proving that scalable video coding can be achieved without sacrificing compression efficiency.

2. ITU-T AND ISO/IEC CODING STANDARDS

Current codecs standardized by ITU-T and ISO/IEC, respectively the H.26x and MPEG-x standards, are based on a hybrid, closed prediction-loop coding architecture[4] (Figure 1). With these schemes the spatial redundancies are typically removed by using a block-based discrete cosine transform (DCT), while the temporal correlations in the video sequence are exploited by inserting a motion estimation and motion compensation stage in a feedback loop. As shown in this figure, in the intra-frame coding mode, the DCT coefficients are quantized and entropy coded, similar to a DCT-based still-image coder. Subsequently, the inverse quantization process and the inverse DCT are used to reconstruct a lossy

[*] Peter.Schelkens@vub.ac.be; phone +32-2-6293955; fax +32-2-6292883; www.etro.vub.ac.be

coded version of the input intra-frame, used as a reference at the next coding stage. In the inter-frame coding mode, the motion estimation determines on a block basis the motion between the new incoming frame and the reference frame that was kept in a frame memory. The obtained motion vectors allow for predicting the new incoming frame based on the reference frame. The resulting motion-compensated frame is then subtracted from the new incoming frame, and the residual prediction error (error-frame) is encoded. The obtained DCT coefficients for both the inter-frames (i.e. error-frames) and intra-frames are quantized and entropy encoded with a lossless variable length encoder (VLC, e.g. Huffman coding or arithmetic encoding, usually in conjunction with run length encoding). The motion vectors produced in the motion estimation stage are losslessly entropy coded as well. Finally, the reconstructed error-frame together with the motion compensated frame are used to compose the new reference frame for the next coding stage.

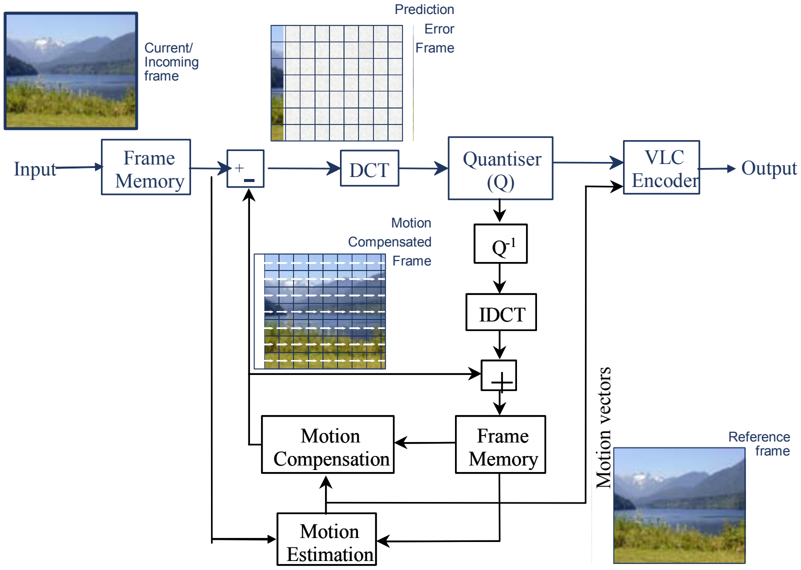


Figure 1 – Basic structure of a hybrid DCT-based coding scheme typically issued for classical ITU-T and ISO/IEC video coding standards.

All current ISO/IEC and ITU-T standards (Figure 2) are based on this hybrid architecture. ITU-T developed their H.26x series of coding standards targeting videoconferencing applications. H.261 was developed to allow audiovisual communication over ISDN channels and supports the use of non-interlaced video in CIF or QCIF format at 30 fps. Its popular successor H.263 targets low bit-rates ranging from 20 kbps up to 320 kbps and supports a bigger range of image formats scaling from sub-QCIF to 16CIF.

In parallel ISO and IEC developed their MPEG-standards. The MPEG-1 standard focuses on the compression of non-interlaced video for storage on CD-ROM, while MPEG-2 was designed for HDTV, DVD and other high-end applications that consume interlaced/non-interlaced video material. MPEG-2 was jointly standardized with the ITU-T experts (resulting in the H.262 specification). Digital television uses for instance MPEG-2 streams to transmit the video. Hence, seen the recent introduction of interactive digital television (iDTV) and the fact that most hardware (set-top boxes, TV satellites etc.) is still using this technology, it can be expected that MPEG-2 will remain to be an important standard the coming years, notwithstanding the introduction of newer and more powerful compression standards. For example, baseline MPEG-4 offers besides a slightly enhanced compression performance also support for object-based video coding.

Recently, both committees decided to define a new joint standard, being MPEG-4 Advanced Video Coding (AVC) and H.264 (H.26L). This new standard provides excellent coding performance over a wide range of bit-rates and is giving a rate-distortion performance boost of 50% over baseline MPEG-4 at the expense of a significantly higher implementation complexity.

By using a layered model for the image information (Figure 3), scalability can be supported by these hybrid coding schemes at the expense of a lower rate-distortion performance[1]. The base layer allows then for obtaining a base quality (in terms of Peak-Signal-to-Noise-Ratio – PSNR, frame-rate or resolution) of the decoded sequence (Figure 3). Decoding the additional layer(s), or enhancement layer(s), improves consequently the quality of the reconstructed video sequence. For Fine Grain Scalability (FGS) in MPEG-4[5] the enhancement layers are obtained by encoding the DCT-coefficients in a bit-plane-by-bit-plane fashion. The usage of this technology allows a more continuous scaling of the image quality with respect to the available bit-rate. Nonetheless, its significant disadvantage consists of the poor rate-distortion performance in comparison with its non-scalable equivalents[1]. Hence, while temporal scalability is typically well supported by the usage of B-frames, quality or resolution scalability should be provided using completely different coding strategies. Scalable video coding solutions that overcome these drawbacks are the wavelet-based video coding technologies, which will be our focus in the following sections.

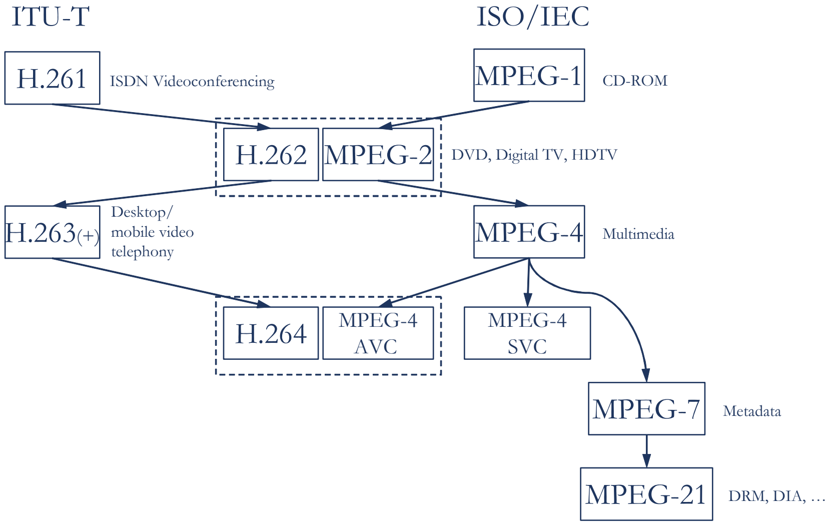


Figure 2 – Overview of ITU-T and ISO/IEC standardization pipelines.

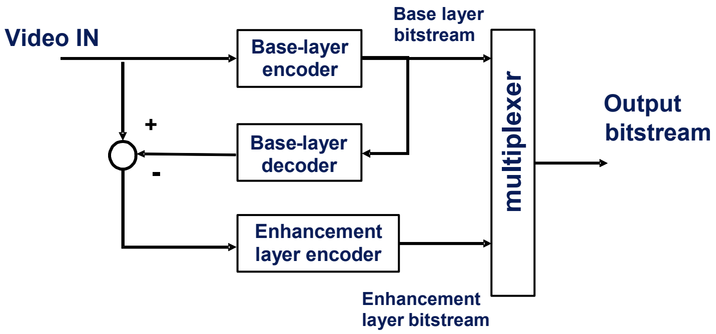


Figure 3 – Realizing a quality scalable hybrid video coder using a base layer/enhancement layer representation of the input sequence.

3. MPEG-4 SCALABLE VIDEO CODING (SVC)

Recently, the MPEG-committee has started the exploration of new technologies to support scalable video coding without jeopardizing the rate-distortion performance compared to single-layer coding schemes (i.e. MPEG-x, H.26x, MPEG-4 AVC). Solutions proposed for the hybrid coding schemes on which those standards are based (e.g. MPEG-4 Fine Grain Scalability-FGS framework), have – as said – never met this criterion in a satisfactory way[1]. Hence, a specific effort is necessary to meet this target by involving new available technologies[6].

As mentioned earlier, a new promising technology is wavelet-based video coding. The new developments that emerge from this research can be situated in a few categories.

The first set of compression technologies is based on **spatial-domain motion compensated temporal filtering (SDMCTF)**[7-9]. These video codecs replace the classical closed-loop with a front-end containing a motion-compensated temporal wavelet transform: during this stage the temporal redundancies are removed by using the temporal (wavelet) filtering, which follows the trajectory indicated by the motion vectors obtained via a spatial domain motion estimation. Thereafter, in order to reduce the spatial redundancies, the frames resulting from the temporal filtering are subsequently spatially-decomposed using a two-dimensional wavelet transform (Figure 4).

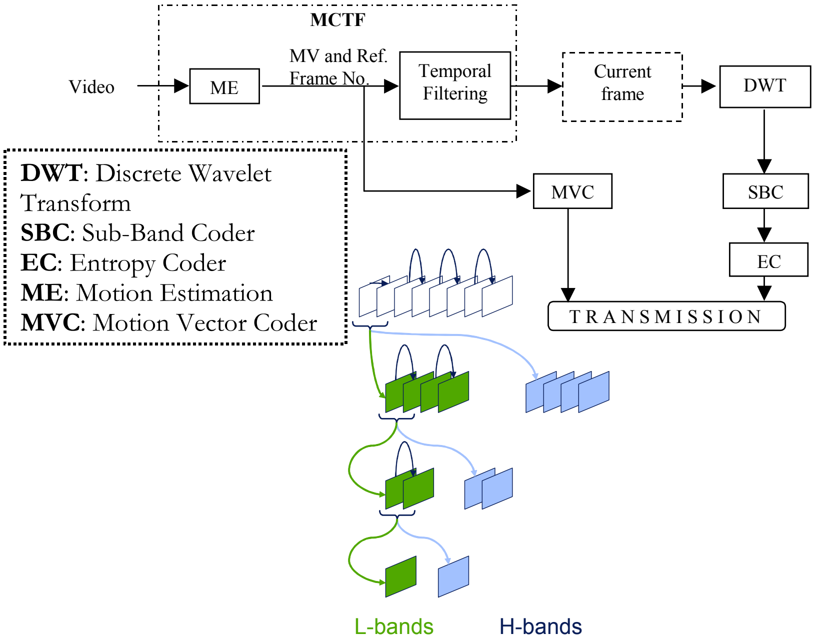


Figure 4 – Video Coding based on Spatial-domain Motion Compensated Temporal Filtering (SDMCTF)

A second category of video compression technologies is based on motion compensation in the wavelet domain, the so-called **in-band predictive schemes**[10, 11]. These schemes use a classical hybrid video codec architecture, with that exception that the motion estimation and compensation take place after the spatial wavelet transform (i.e. in the wavelet domain) and not before (Figure 5). The advantage of this approach is that the multiresolution nature of the wavelet domain representation can be fully exploited. For a long time, the main bottleneck blocking the wider acceptance of these techniques has been the shift-variance problem, characteristic for critically sampled wavelet transforms, which hampers efficient motion estimation. Recently, this bottleneck has been removed by using an overcomplete wavelet representation[10, 11] of the reference frames in the prediction process. Such an overcomplete representation can be constructed from the critically-sampled wavelet representation of the reference frames using a complete-to-overcomplete discrete wavelet transform (CODWT)[10,11] – see Figure 5. Inside the prediction loop, all phases of the overcomplete representation are taken into account during motion estimation. This approach removes the aforementioned bottleneck due to the shift-invariant nature of the overcomplete discrete wavelet transform, but introduces an augmented computational complexity. Recently, efficient solutions have been proposed to alleviate this drawback[10, 12, 13].

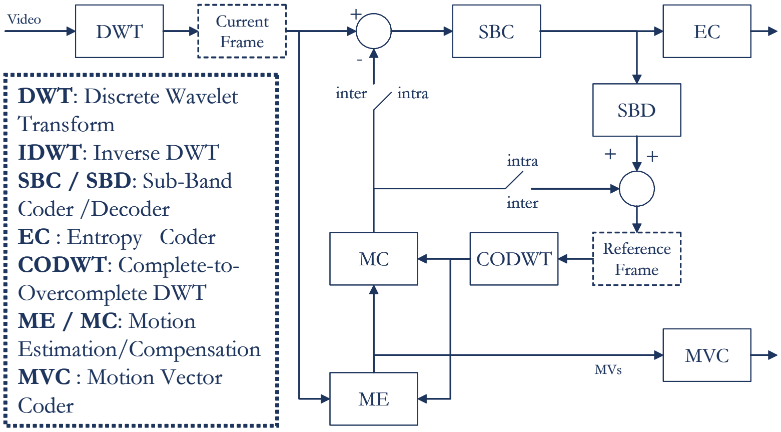

Video
DWT
Current Frame
+
-
SBC
EC
inter
intra
SBD
+
+
intra
inter
MC
CODWT
Reference Frame
MVs
MVC
ME
DWT: Discrete Wavelet Transform
IDWT: Inverse DWT
SBC / SBD: Sub-Band Coder /Decoder
EC : Entropy Coder
CODWT: Complete-to-Overcomplete DWT
ME / MC: Motion Estimation/Compensation
MVC : Motion Vector Coder

Figure 5 – In-band, hybrid wavelet video coding. Notice the complete to overcomplete discrete wavelet transform (CODWT) module in the prediction loop.

Very recent research efforts have been focused on a third category of architectures, combining the MCTF and in-band wavelet coding approaches, yielding competitive results compared to the single-layer switching (SLS) mode of MPEG-4[14] (Figure 6). This architecture is based on a spatial wavelet transform front-end, followed by a wavelet domain, subband-based MCTF or **in-band MCTF (IBMCTF)**. The advantage of this architecture is that it combines the open-loop structure of the SDMCTF-based codecs, yielding excellent performance for quality scalability and an efficient temporal decomposition of the video sequence, with the superior performance for resolution scalability provided by the in-band hybrid video coding architectures. Moreover, the IBMCTF architecture allows for a very flexible subband-based configuration of the codec, enabling the choice of different GOF structures, different motion estimation accuracies and different prediction structures (bi-directional, forward/backward prediction, longer temporal filters) per subband or resolution (see Figure 7).

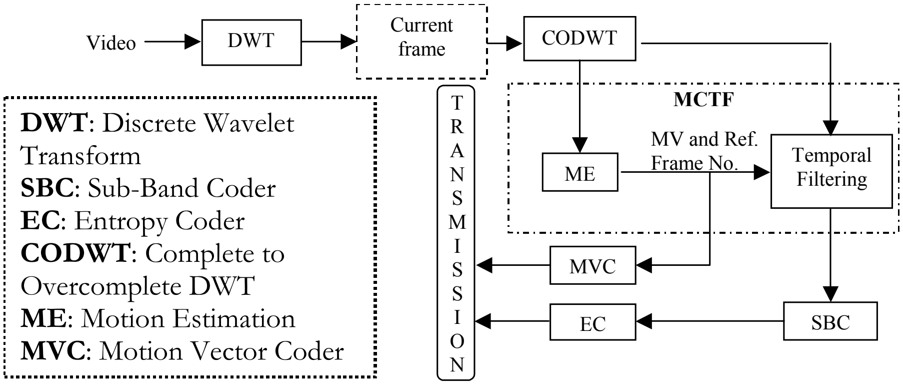


Figure 6 – In-band motion compensated temporal filtering (IBMCTF).

Recent rate-distortion results prove that wavelet-based video coding delivers promising rate-distortion behavior, while offering at the same time the required functionalities of resolution, quality and temporal scalability. In the next section, coding results that demonstrate these findings will be given.

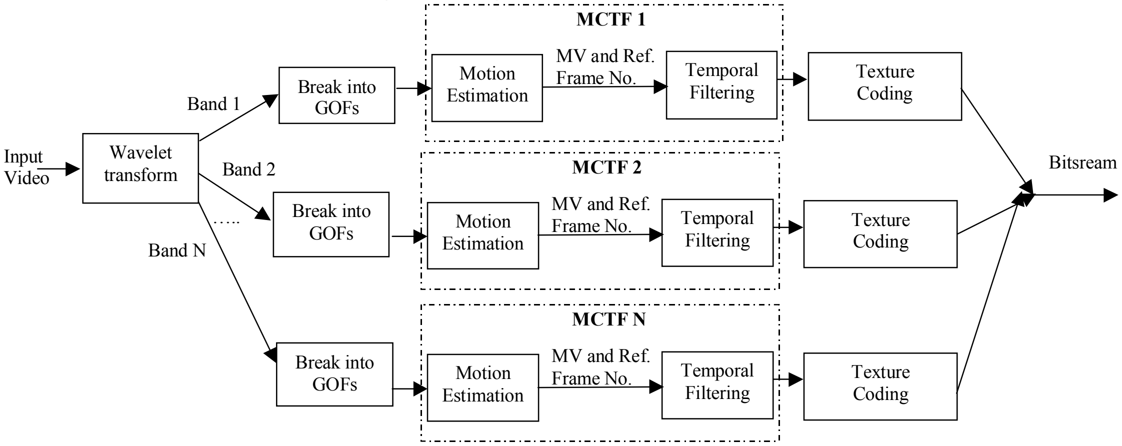


Figure 7 – The IBMCTF codec offers extremely flexible configuration possibilities on a subband basis. The low-frequency subband can even be encoded with another MPEG or H.26x-codec providing in this way backward compatibility for specific low-end applications.

4. CODING RESULTS

We evaluate the coding performance obtained with our experimental instantiations of the SDMCTF and IBMCTF video-coding architectures[18]. For results on the in-band prediction schemes we refer to our earlier work[10,13].

The discussed wavelet codecs have been equipped with successive approximation quantization and with an efficient entropy coding system, called QuadTree-Limited (QT-L) codec[3], which combines quadtree coding and block-based coding of the significance maps with context-based entropy coding[3,15]. The motion vectors are encoded with a combination of predictive coding and context-based arithmetic coding[16, 17]. Our MCTF codecs were also equipped with an efficient multi-hypothesis block-based motion estimation technique[18]. For reasons of comparison, we will also compare the performance of these codecs with the MC-EZBC codec[19], which utilizes bi-directional motion compensation.

First, we evaluate our SDMCTF codec against (1) the MC-EZBC[19], which is a state-of-the-art fully-scalable instantiation of SDMCTF-based coding, and (2) the non-scalable Advanced Video Coder (AVC) jointly standardized by ISO/IEC and ITU-T[20]. In addition, we compare the spatial-domain MCTF with our in-band MCTF instantiation and determine the efficiency of each coding architecture for resolution and temporal scalability.

For our experiments, we use the 9/7 filter-pair with a three level spatial decomposition. The chosen parameters for the multi-hypothesis motion estimation (ME) correspond to the bidirectional Haar temporal filter without the use of the update step[21]. Four temporal decomposition levels were performed, with the search range dyadically increasing per level. After the ME algorithm, the motion vectors are coded using a prediction formed by the neighboring vectors[17] and the residual error is compressed using adaptive arithmetic entropy coding..

Figure 8 illustrates the coding results obtained with the spatial-domain MCTF using multi-hypothesis ME against the fully-scalable MC-EZBC coder[19] and the new MPEG-4 AVC coder[20]. For the CIF-resolution sequence, ME with 1/8 pixel accuracy was performed, while retaining 1/4 pixel accuracy for the larger-size sequence in order to limit the complexity of the simulation. The AVC coder operated with two references, variable block sizes and CABAC entropy coding; also, a GOP structure of 16 frames using three B-frames between consecutive P-frames was used. Due to its non-scalable nature, the AVC coder had to perform coding/decoding operations for every experimental point, while the two scalable algorithms produced embedded bit-streams, from which sub-streams were extracted at the bit-rates produced by the AVC.

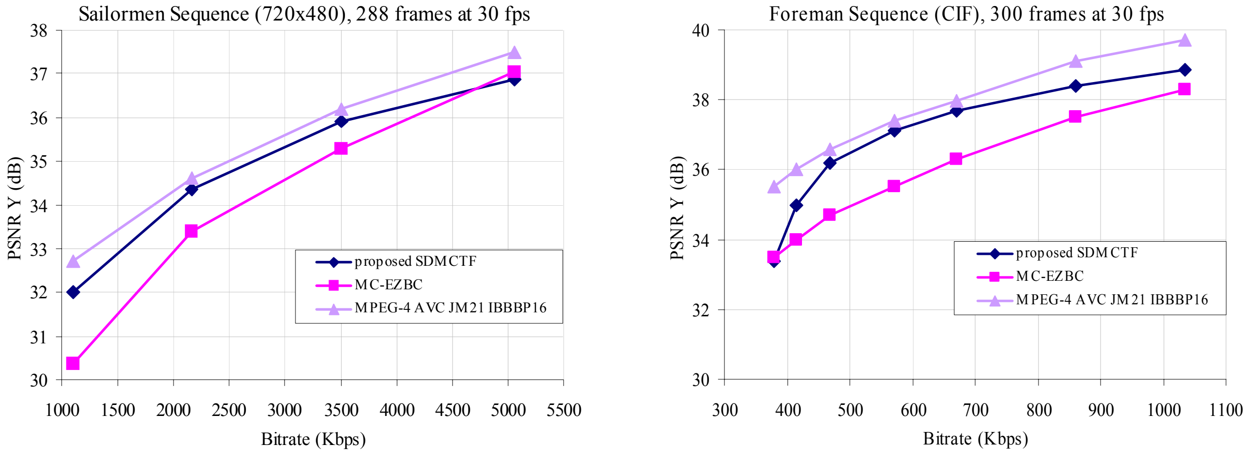


Figure 8 – Comparison between SDMCTF, MCEZBC and MPEG-4 AVC.

The results show that the SDMCTF using multi-hypothesis prediction is comparable with the highly optimized AVC over a large range of bit-rates, while retaining the advantages offered by embedded coding. Additionally, significant coding gains are observed in comparison to the state-of-the-art MC-EZBC over a large range of bit-rates. Notice that in comparison to AVC's ME, the employed ME algorithm does not include selective intra prediction modes, the rate-distortion optimization is simpler, and only two block sizes are considered. We plan to address these issues in the future because preliminary experiments suggest that the performance of the MCTF-based approaches improves with the use of such tools.

To compare the performance between the spatial domain and in-band MCTF approaches we present two typical examples, using two sequences of high and medium motion content. Half-pixel accurate ME was used in these experiments. The results for the full-resolution/full frame-rate decoding are given in Figure 9 (top). We find that the use of multiple motion vectors for each resolution level increases the coding performance of IBMCTF in the case of complex motion, as seen in the Football sequence. Nevertheless, without a rate-distortion optimization framework, the increased motion-vector bit-rate may overturn this advantage for sequences with medium motion activity, like the Foreman sequence.

The full-scalability capability of the IBMCTF codec is demonstrated in Figure 9 (bottom), which depicts the performance for decoding at half resolution/half frame-rate and different quality levels. The IBMCTF operates using the level-by-level CODWT. The reference sequences used in the rate-distortion comparison of Figure 9 (bottom) are obtained by one-level spatial DWT performed on a frame-by-frame basis, followed by retaining the LL subbands and frame-skipping. Both IBMCTF and SDMCTF can achieve resolution-scalable decoding. However, it is important to notice that, in a scalable coding scenario corresponding to coarse-to-fine progressive transmission of different resolutions, only the IBMCTF architecture guarantees a lossy-to-lossless decoding at all resolutions provided that the LL subbands are used as the best possible approximations of the original video sequence at any given resolution[22]. This is observed experimentally in Figure 9 (bottom), where a large PSNR difference exists between the different alternatives. A typical visual comparison is also given in Figure 10.

Additionally, a visual comparison between the MPEG-4 AVC and our SDMCTF codec (Figure 11) shows that both codecs yield a similar visual quality, each though with their own artifacts.

5. CONCLUSION

We have demonstrated that the recently introduced spatial domain and in-band motion compensated temporal filtering video coders deliver a rate-distortion behavior that is competitive with non-scalable state-of-the-art coding technology, such as MPEG-4's Advanced Video Coding. The results presented in this paper suggest that scalability can be reached at little expense in terms of rate-distortion behavior. Moreover, it can be expected that significant improvements are feasible with respect to – for example – more advanced motion estimation techniques and rate-allocation mechanisms. Finally, implementation aspects need to be investigated, since it is unclear (and hard to predict) at this moment how the complexity of the MCTF solutions will compare against their non-scalable competitors.

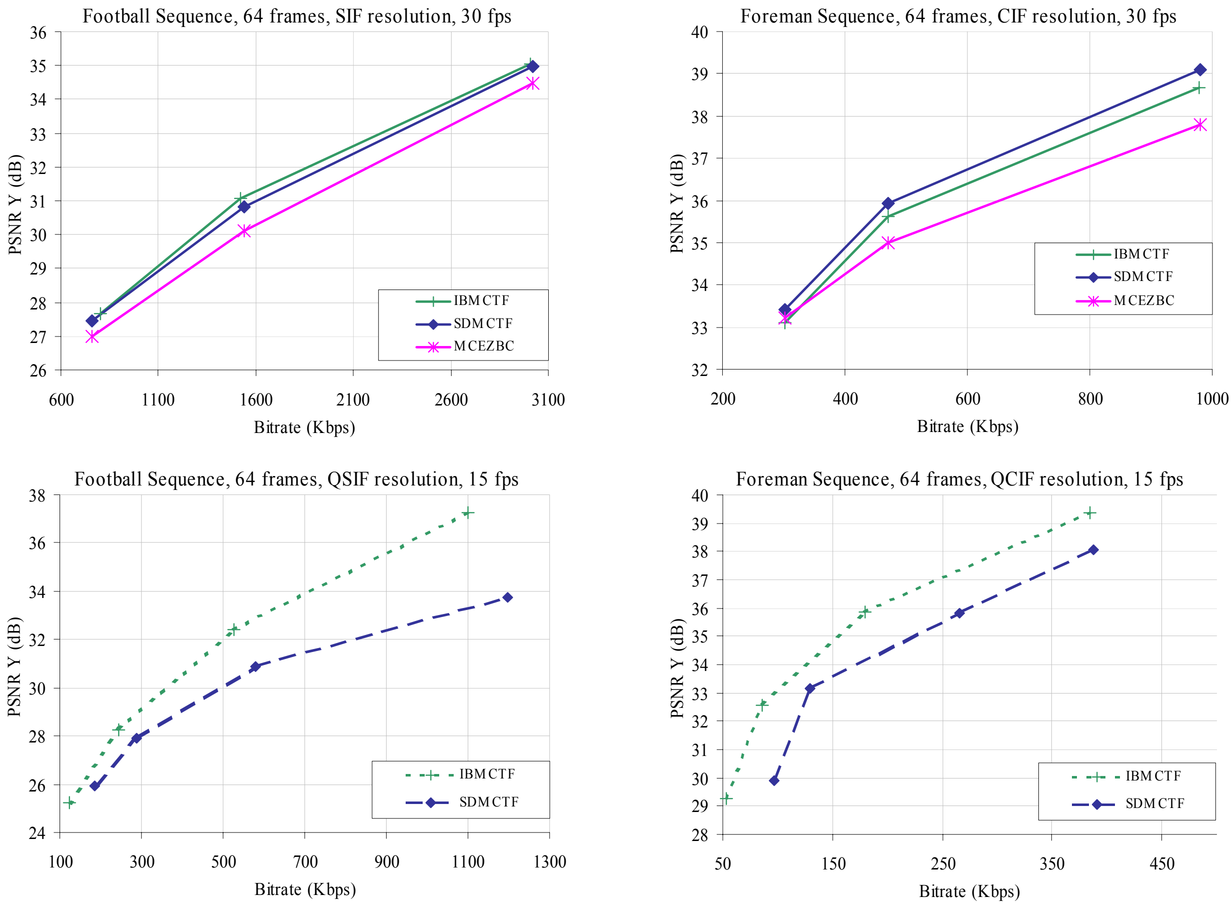
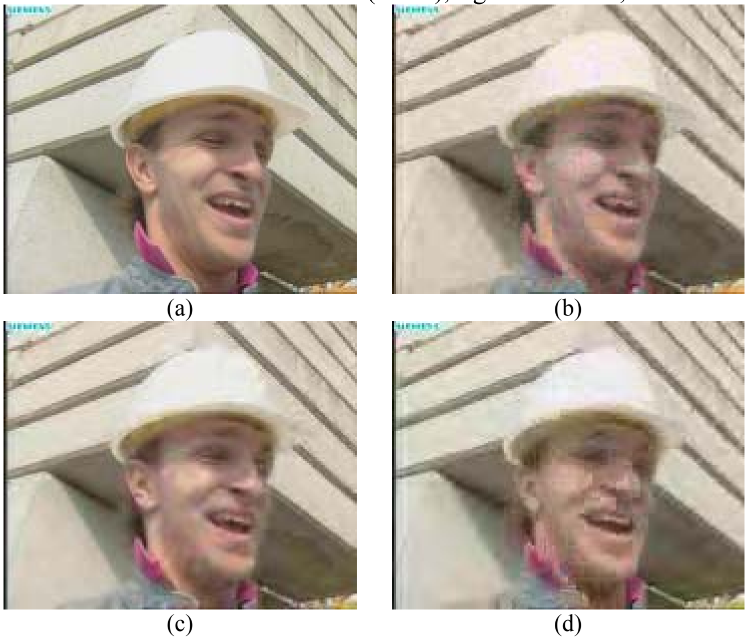


Figure 9 – Comparison between in-band MCTF and spatial-domain MCTF. All points were obtained by decoding a different number of resolution, quality or temporal levels from the temporal decomposition. For the PSNR comparison at the lower spatial resolution, the uncoded LL subband of the original sequence is used, while frame-skipping provides the reference sequence at lower frame-rates.

ACKNOWLEDGEMENTS

This work was supported in part by the Flemish Institute for the Promotion of Innovation by Science and Technology (PhD bursary Joeri Barbarien and GBOU RESUME project), the Federal Office for Scientific, Technical and Cultural Affairs (IAP Phase V – "Mobile Multimedia"), the Fund for Scientific Research – Flanders ("Reconfigurable Hardware for Embedded Systems" and "Design and Management of Metadata-based Multimedia Systems using Active Networks") 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.



(a) (b)

(c) (d)

Figure 10 – Visual comparison for the half-resolution/half frame-rate decoding at 97 Kbps (Foreman sequence, decoded frame no. 7): (a) Original, (b) SDMCTF, (c) IBMCTF, (d) MC-EZBC.

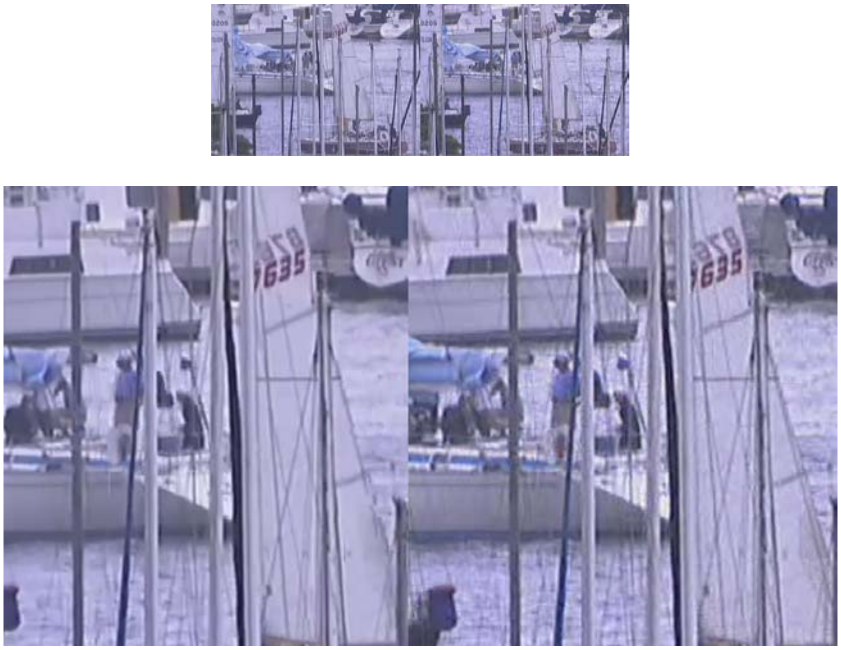


Figure 11 – Visual results obtained with the MPEG-4 AVC (left) and SDMCTF (right) codecs for the Harbour sequence at 384kbps for a resolution of 360x240pels and 15fps (upper pictures) and at 1.5Mbps for a resolution of 720x480 and 30fps (lower pictures). Both codecs yield comparable visual results.

REFERENCES

1. Y. He, R. Yan, F. Wu, and S. Li, "H.26L-based fine granularity scalable video coding," ISO/IEC JTC1/SC29/WG11, Pattaya, Thailand, Report MPEG2001/m7788, December 2001.
2. D. Taubman, "High Performance Scalable Image Compression with EBCOT," *IEEE Transactions on Image Processing*, vol. 9, no. 7, pp. 1158-1170, July 2000.
3. P. Schelkens, A. Munteanu, J. Barbarien, M. Galca, X. Giro i Nieto, and J. Cornelis, "Wavelet Coding of Volumetric Medical Datasets," *IEEE Transactions on Medical Imaging*, vol. 22, no. 3, pp. 441-458, March 2003.
4. A. Puri and T. Chen, *Multimedia Systems, Standards, and Networks*, Signal Processing and Communications Series, vol. 2. New York - Basel: Marcel Dekker, Inc., 2000.
5. M. van der Schaar and H. Radha, "Adaptive motion-compensation fine-granular-scalability (AMC-FGS) for wireless video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 12, no. 6, pp. 360-371, June 2002.
6. D. Taubman and M. W. Marcellin, *JPEG2000 - Image Compression: Fundamentals, Standards and Practice*. Hingham, MA: Kluwer Academic Publishers, 2001.
7. J.-R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 559-571, September 1994.
8. D. Taubman and A. Zakhor, "Multirate 3-D Subband Coding of Video," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 572-588, September 1994.
9. J. W. Woods and G. Lilienfield, "A resolution and frame-rate scalable subband/wavelet video coder," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 9, pp. 1035-1044, September 2001.
10. G. Van der Auwera, A. Munteanu, P. Schelkens, and J. Cornelis, "Bottom-up motion compensated prediction in the wavelet domain for spatially scalable video coding," *IEE Electronics Letters*, vol. 38, no. 21, pp. 1251-1253, October 2002.
11. H.-W. Park and H.-S. Kim, "Motion Estimation Using Low-Band-Shift Method for Wavelet-Based Moving-Picture Coding," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 577-587, April 2000.
12. I. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens, and J. Cornelis, "Fast Level-By-Level Calculation of Overcomplete DWT for Scalable Video-Coding Applications," Vrije Universiteit Brussel, Department of Electronics and Information Processing (ETRO), Brussel, ETRO/IRIS Technical Report IRIS-TR-81, November 2001.
13. F. Verdicchio, I. Andreopoulos, A. Munteanu, J. Barbarien, P. Schelkens, J. Cornelis, and A. Pepino, "Scalable video coding with in-band prediction in the complex wavelet transform," Proceedings of Advanced Concepts for Intelligent Vision Systems (ACIVS2002), Gent, Belgium, pp. 6, September 9-11, 2002.
14. I. Andreopoulos, M. van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelis, "Complete-to-overcomplete discrete wavelet transforms for scalable video coding with MCTF," Proceedings of SPIE Visual Communications and Image Processing (VCIP), Lugano, Switzerland, Vol. 5150, pp. 719-731, July 8-11, 2003.
15. A. Munteanu, J. Cornelis, G. Van der Auwera, P. Cristea, "Wavelet-based lossless compression scheme with progressive transmission capability," *International Journal of Imaging Systems and Technology*, special issue on "Image and Video Coding", Eds. J. Robinson and R. D. Dony, John Wiley&Sons, vol. 10, no. 1, pp. 76-85, January 1999.
16. J. Barbarien, I. Andreopoulos, A. Munteanu, P. Schelkens, and J. Cornelis, "Coding of motion vectors produced by wavelet-domain motion estimation," Proceedings of IEEE Picture Coding Symposium (PCS), Saint Malo, France, April 23-25, 2003.
17. J. Barbarien, I. Andreopoulos, A. Munteanu, P. Schelkens, and J. Cornelis, "Coding of motion vectors produced by wavelet-domain motion estimation," ISO/IEC JTC1/SC29/WG11 (MPEG), Awaji island, Japan, MPEG Report M9249, December 7-12, 2002.
18. I. 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 (MPEG), Trondheim, Norway, MPEG Report M9911, July 20-25, 2003.
19. P. Chen and J. W. Woods, "Bidirectional MC-EZBC with lifting implementations," *IEEE Transactions on Circuits and Systems for Video Technology*, to appear.
20. T. Wiegand and G. Sullivan, "Draft ITU-T recommendation and final draft international standard of joint video specification," ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6

21. D. S. Turaga and M. van der Schaar, "Wavelet coding for video streaming using new unconstrained motion compensated temporal filtering," Proceedings of International Workshop on Digital Communications: Advanced Methods for Multimedia Signal Processing, Capri, Italy, pp. 41-48, September 2002.

22. I. 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," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Hong Kong, Vol. III, pp. 417-420, April 6-10, 2003.