FULLY-SCALABLE WAVELET VIDEO CODING USING IN-BAND MOTION COMPENSATED TEMPORAL FILTERING

Yiannis Andreopoulos[1], Mihaela van der Schaar[2], Adrian Munteanu[1], Joeri Barbarien[1], Peter Schelkens[1], Jan Cornelis[1]

[1]Vrije Universiteit Brussel [2]Philips Research USA

ABSTRACT

This paper presents a novel fully-scalable wavelet video coding scheme that performs efficient open-loop motion compensated temporal filtering (MCTF) in the wavelet domain (in-band). Unlike the conventional spatial-domain MCTF (SDMCTF) schemes, which apply MCTF on the original image data and *then* encode the residual image using a critically-sampled wavelet transform, the framework presented here applies the in-band MCTF (IBMCTF) after the discrete wavelet transform (DWT) is performed in the spatial dimensions. To overcome the inefficiency of motion estimation (ME) in the wavelet domain, a complete-to-overcomplete DWT (CODWT) is performed. The proposed framework provides improved quality (SNR) and temporal scalability as compared with existing in-band closed-loop temporal prediction schemes with ODWT and improved spatial scalability as compared to SDMCTF. We present a thorough comparison between SDMCTF and the proposed IBMCTF in terms of coding efficiency and scalability. Furthermore, we describe several extensions that enable the filtering of the various bands to be performed independently, based on the resolution, sequence content, complexity requirements and desired scalability.

1. INTRODUCTION

Several scalable coding schemes have recently been proposed that are aimed at enabling the Universal Media Access (UMA) paradigm. For instance, the recently standardised MPEG-4 Fine-Granularity-Scalability (FGS) is able to provide temporal-SNR scalability through a single fine-granular enhancement-layer, but does not provide resolution scalability. However, to serve a broad range of data rates (e.g. from a few Kbit/s to several Mbit/s) on heterogeneous networks, or on a wide selection of terminals with different characteristics, fine-granular spatio-temporal and SNR scalability becomes necessary. Selecting tradeoffs among these three dimensions (spatial/temporal/quality) becomes inevitable in order to support a high degree of content variation with high-quality.

Motion-compensated wavelet video coding schemes can provide full scalability with fine granularity over a large range of bitrates. These schemes can be classified into the following categories [3]:

- **Wavelet in-the-loop:** The conventional predictive coding structure used for FGS is preserved but the residual error in the motion-compensation (MC) loop is coded using DWT instead of DCT.
- **In-band prediction:** First, the spatial wavelet transform is performed on the video data. Subsequently, the temporal redundancy present in each band is exploited using closed-loop ME/MC.
- **Inter-frame wavelet:** Open-loop MCTF is performed first on the spatial-domain video data (SDMCTF) and subsequently, the DWT is performed for the spatial decorrelation of the residual information.

In [3], it was concluded that the first two categories are not appropriate for UMA. The first category ("wavelet-in-loop") preserves the conventional motion-compensation prediction loop. Hence, whenever the entire residual signal is included into the motion-compensation prediction loop, drift occurs if decoding is performed at various bitrates. If the residual signal is *not* entirely included into the motion-compensation prediction loop then there is a considerable coding penalty associated with SNR scalability. Spatial scalability obtained with this scheme also suffers from drift effects and/or significant compression inefficiencies. Furthermore, as mentioned in [3], the block-based MC is less efficient than for DCT-based predictive coders since the discontinuity in the motion boundaries (blocking artefacts) is represented as high-frequency content in the high-frequency wavelet subbands, thereby leading to inefficient MC residual texture coding.

The second category, "in-band prediction", can achieve spatial scalability without experiencing drift, as the MC prediction is applied separately in each spatial resolution level. However, in [3] it was concluded that, if the entire residual signal is included in the prediction loop, drift still occurs *within* these spatial levels as soon as cropping of bits for quality (SNR) scalability is applied. This statement was based on the in-band prediction scheme and results presented in [1].

Alternatively, the third wavelet-coding category based on SDMCTF does not employ a temporal-recursive structure for removing the temporal redundancies. Instead, it performs MCTF that allows perfect reconstruction. During MCTF, frames are filtered temporally in the direction of motion *prior* to performing the spatial transformation and coding, as illustrated in Fig. 1. However, the inter-frame wavelet coding has also several limitations. In particular, the performance of spatial scalability is limited since the ME at full resolution creates drift in lower-resolution decoding. Furthermore, the spatial-domain ME/MC in this scheme also encounters the problem of discontinuity in the motion boundaries mentioned before for the ME/MC in the "Wavelet in-the-loop" scheme. Finally, *no flexibility* exists for the adaptation of the size of the Group Of Frames (GOF), prediction structure and motion accuracy per resolution, thereby limiting the coding efficiency of SDMCTF at different resolutions.

The paper is organized as follows. In section 2, we describe in some detail the limitations of existing in-band prediction scheme and in section 3 we introduce a new fully-scalable wavelet video coding scheme that eliminates these inefficiencies by performing open-loop MCTF in the overcomplete wavelet domain. The results for this new scheme are presented in Section 4, and a comparison with SDMCTF schemes is made in terms of coding efficiency and scalability. Conclusions and directions for further research are presented in Section 5.

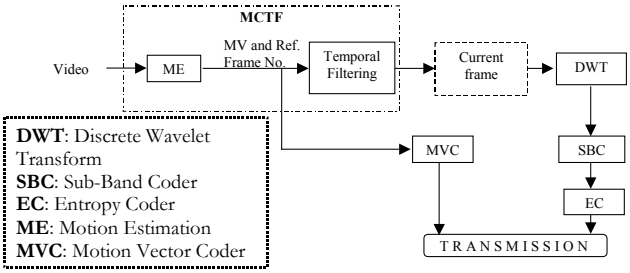


Fig. 1: Block diagram of SDMCTF coding.

2. LIMITATION OF CONVENTIONAL IN-BAND PREDICTION SCHEMES

In [1], a motion-compensated wavelet video coding framework with in-band prediction has been proposed that provides resolution, quality and frame-rate scalability with the classical hybrid-coding

(closed loop) encoding. The encoder performs the ME and MC in the wavelet domain (*in-band*) following a level-by-level refinement of the compressed information. However, since the critically-sampled wavelet decomposition is only periodically shift-invariant, the ME/MC procedures are performed in the overcomplete discrete wavelet transform (ODWT). The produced error frames from the motion compensation are coded using an embedded intra-band coding technique. A block diagram of this coding architecture is depicted in Fig. 2. The embedded nature of the employed wavelet-based compression algorithm and the level-by-level operation of the ME/MC guarantee that the produced bitstream can be decoded at a variety of resolutions and quality levels without drifting problems, as long as every decoder receives a certain portion (i.e. a base-quality layer) of the encoded bitstream. Additionally, depending on the type of successive prediction used in the coding scheme, temporal scalability is supported as well, by simply skipping the frames that are not used as references. In this way, fine-grain scalable video decoding in resolution, quality and frame-rate is achieved.

In [1] and [3], it has been recognized that one of the limitations of this scheme, referred to subsequently as "Overcomplete Predictive Wavelet Coding" (OPWC), is the lower efficiency of SNR (quality) scalability. This inefficiency is due to the fact that a closed-loop MC is performed within each subband, and thus a base-quality layer has to be selected that is used as reference for subsequent frames. Similar with the wavelet in-the-loop schemes, this base-quality layer leads to reduced coding performance, since the residual signal is *not* entirely included into the motion-compensation prediction loop. This is a known problem of all closed-loop motion compensated video coding schemes that limits the SNR scalability in comparison to the open loop wavelet approaches (MCTF).

Another limitation of the scheme in [1] is that limited flexibility is permitted for temporal scalability by using B-frames. Alternatively, in the MCTF-based schemes, due to the use of several temporal decomposition levels, a larger set of frame-rates can be provided.

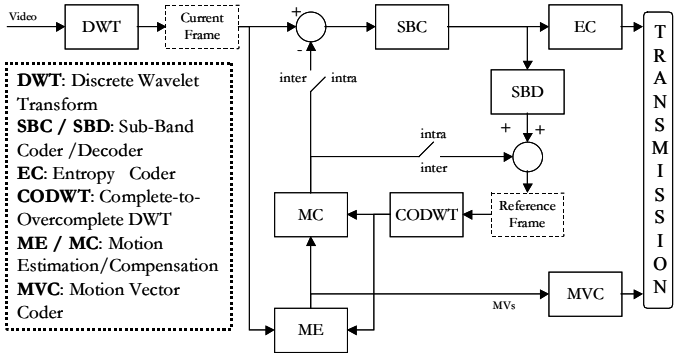


Fig. 2: OPWC using closed-loop in-band motion-compensation.

3. PROPOSED IN-BAND MOTION COMPENSATED TEMPORAL FILTERING (IBMCTF)

In order to overcome the performance loss in SNR scalability and to improve the temporal scalability of OPWC, this paper proposes a combination of OPWC with MCTF. More specifically, the video frames are spatially decomposed into multiple subbands using wavelet filtering, and the temporal correlation for each subband is removed using MCTF (see Fig. 3). The residual signal after the MCTF is coded band-by-band using any desired texture coding technique (DCT-based, wavelet-based, matching pursuit etc.). Also, all the recent advances in MCTF can be employed for the benefit of the OPWC scheme. For instance, the efficient UMCTF scheme proposed in [2] can be employed for temporal filtering.

For efficient ME, the ODWT is constructed from the critically-sampled decomposition of the reference frame(s) assuming resolution scalability, i.e. the codec with k decomposition levels can decode up to k dyadically-reduced resolution levels from the same compressed bitstream. The construction of the ODWT from the DWT, a procedure that is a *complete-to-overcomplete* discrete wavelet transform (CODWT), occurs at both the encoder and decoder side to reconstruct the reference frame(s). In this paper we denote the critically-sampled subbands of the decomposition level l as $LL^l_{(0,0)}, LH^l_{(0,0)}, HL^l_{(0,0)}, HH^l_{(0,0)}$, where the superscript indicates the decomposition (resolution level) and the subscript indicates the polyphase components (*even*=0, *odd*=1) retained after the downsampling in the vertical and horizontal direction.

Under the assumption of resolution scalability, when the (de)coder is processing the decomposition level l, with $l < m \le 1$, we have $LH^m_{(0,0)} = HL^m_{(0,0)} = HH^m_{(0,0)} = \varnothing$, i.e. the finer-resolution levels are zero. With this constraint, the CODWT of the reference frame(s) can be constructed for each resolution level either by using classical techniques such as the LBS algorithm of [6], or more advanced techniques that use a single-rate calculation scheme with reduced computational and delay overhead [4] [5].

After the construction of the ODWT, for each level l we produce a set of critically-sampled subbands $LH^l_{(i,j)}, HL^l_{(i,j)}, HH^l_{(i,j)}$, $0 \le i, j < 2^l$ (and also $LL^l_{(i,j)}$ if l=k). The ME/MC procedures are performed in a level-by-level fashion: for each level, full-search can be performed in order to jointly-minimize the distortion measure for each triplet of blocks from the $LH^l_{(0,0)}, HL^l_{(0,0)}, HH^l_{(0,0)}$ of the current frame that correspond to the same spatial-domain location of a triplet of blocks from $LH^l_{(i,j)}, HL^l_{(i,j)}, HH^l_{(i,j)}, 0 \le i, j < 2^l$ of the reference(s). For the coarsest resolution level, the motion estimation process is performed separately for the *LL* subband. In addition, the motion-vectors of different resolutions can be correlated in order to limit the search range. This leads to reductions in the ME complexity at the expense of affecting the quality of the matches in comparison to the case when ME is performed independently at each level. Other techniques that are typically used in spatial-domain ME/MC, such as interpolation to sub-pixel accuracy, can be performed directly in the ODWT of the reference frame(s), since the linear interpolation and the DWT filtering without downsampling are both linear shift invariant operators and hence their order of application to the input signal can be interchanged.

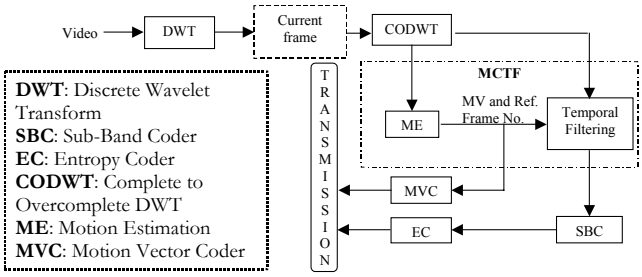


Fig. 3 Proposed IBMCTF scheme.

A simple pictorial example of the previously described ME process is given in the left part of Fig. 4. In this case, the 2-D ODWT of the reference frame contains four critically-sampled low-frequency subbands and 3×4 high-frequency critically-sampled subbands. The gray-shaded blocks indicate the search area in every $LL^1_{(i,j)}$ subband of the reference, with $i, j = \{0.1\}$. In this example, we show that the best match is found in $LL^1_{(0,1)}(x_l, y_l)$ and the corresponding vector is $MV_{low}(x,y)$. Notice that besides this motion vector, the corresponding subband-index (0,1) in which the best match is found should also be transmitted to the decoder. The process is repeated for the high-frequency subbands *HL*, *LH*, *HH*, by grouping each triplet of blocks corresponding to the same spatial-domain location. As shown in the example of Fig. 4 (right), the motion compensation is done in this example using the best match that was found in the triplet of blocks at (x_h, y_h) in the subbands $HL^1_{(1,1)}, LH^1_{(1,1)}, HH^1_{(1,1)}$ of the ODWT of the reference frame. Notice that although the ODWT is used for the

ME process, the MC occurs in the critically-sampled transform of the current frame. Hence, the produced H-frames are critically-sampled. Generalizing the previous example, k+1 ME/MC procedures are performed for k decomposition (resolution) levels. The motion vectors produced for the luminance channel are subsequently used for the chrominance channels as well.

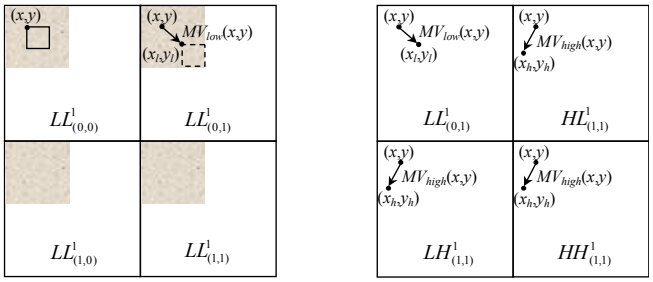


Fig. 4. Left: the ME process for a block of the *LL* subband. Right: The MC process. One level decomposition is shown.

To estimate the increase in bitrate for the in-band motion vectors in comparison to the spatial-domain case, we calculate the uncoded bit budget for the motion vectors of each frame in both cases, under the simple assumption of pixel-accurate motion estimation with a fixed block size $B_R \times B_C$ and a search range of $\pm r$ samples horizontally and vertically.

Starting with the spatial-domain case, for an input frame of $R \times C$ samples and r_n reference frames, for each frame of the SDMCTF system, the bit budget for the motion vectors is:

$$R_{SD} = \left\lfloor \frac{R \cdot C}{B_R \cdot B_C} \right\rfloor \left[2\log_2(2r) + \log_2 r_n \right].$$

Similarly, following the fact that for the first level of the wavelet decomposition there is a downsampling factor of two, if we use half the block size and a dyadically-reduced search range for all the resolution levels of the in-band ME (i.e $\frac{B_R}{2} \times \frac{B_C}{2}$ and a search range of $\pm(r \cdot 2^{-l})$ coefficients with l=1,2,..k, respectively) it can be easily found that the required bits-per-frame for the motion vectors of spatial-decomposition level l is:

$$R_{IB}^{lev}(l) = f(l) \left\lfloor \frac{R \cdot C}{(B_R \cdot B_C) \cdot 4^{l-1}} \right\rfloor \left[2\log_2(2r) + \log_2 r_n \right],$$

where $f(l) = \{1 : 1 \le l < k,\ 2 : l = k\}$ is a factor that takes into account the separate motion estimation in the *LL* subband at the coarsest-resolution level k. As a result for a decoder that decodes all the k spatial decomposition levels:

$$R_{IB}^{tot}(k) = \sum_{l=1}^{k} R_{IB}^{lev}(l).$$

By comparing R_{SD} and $R_{IB}^{tot}(k)$ for various resolution levels k, it can be seen that $R_{IB}^{tot}(k) \le 1.5 R_{SD}$, for $k \ge 2$. In SDMCTF, the decoded spatial-domain motion-vector bit budget R_{SD} is fixed even if fewer resolution levels are decoded. On the other hand, the decoded bit budget for the in-band ME case is:

$$R_{IB}^{res}(l) = \sum_{i=l}^{k} R_{IB}^{lev}(i).$$

The last equation shows that there is a dyadic reduction for the motion-vector bits-per-frame in the in-band case when decoding to coarser resolution levels l, l>1. As a result, $R_{IB}^{res}(l) \le \frac{R_{SD}}{2}$ for l>1, i.e the motion vector bitrate for the lower-resolution decoders is at least half of the corresponding bit budget of the spatial-domain case. In this paper we do not further extend the discussion on the possible coding modes for the multiresolution motion vectors, but spatio-temporal hierarchical coding of motion vectors can be performed for MCTF (see e.g. [7]).

4. RESULTS

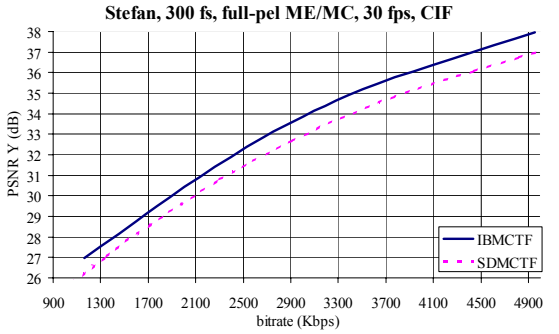
For a fair comparison of the two different frameworks-SDMCTF and IBMCTF- we used the same block-based, pixel-accurate MCTF with a fixed block size; as explained in the previous section, the block size and search range for decomposition level one in IBMCTF are the in-band equivalent of the SDMCTF selections. In addition, both schemes use the same embedded intra-band wavelet coder [1]. While more sophisticated estimation techniques can be employed for an improved performance, in this paper, our main focus is to perform a first comparative study between the two schemes. To ensure resolution scalability, the utilized coder is used separately per resolution level and a target distortion bound is sought in the transform representation of each resolution level. For both systems, the same distortion-based control was used and both used four temporal decomposition levels and three levels for the spatial transform. In addition, for both schemes the bi-directional AHA MCTF scheme of [2] was chosen. In fact, apart from the different processing order (spatial transform followed by temporal filtering) and the different representation in the ME/MC approach (wavelet coefficients instead of input spatial-domain samples), both systems were implemented with the same software.

For the test purposes of this paper we focused on relatively high bitrates. This is done for two reasons. Firstly, as mentioned previously, the employed motion estimation and filtering has only limited accuracy (pixel-accurate) and thus the performance should be evaluated at higher bitrates. The second reason is that we are more interested in examining the *asymptotic behavior* of both schemes under comparison. Specifically, for a first comparison we feel that it is more important to first examine whether both systems provide a comparable degree of steepness in the rate-distortion (R-D) curves at all resolutions and frame-rates and whether they can converge to a very low distortion for high bitrates. In this manner, we can establish a firm answer as to whether there are any *fundamental* problems with resolution scalability in the SDMCTF framework and whether alternative techniques, such as IBMCTF should be pursued.

We present PSNR results in Fig. 5 and Fig. 6 for two CIF sequences. The reference for the lower-resolution is the uncoded LL subband of the input frames, while the reference for the lower frame rate is extracted by frame-skipping. It can be seen that for the typical range of acceptable visual quality range for full-resolution/full frame-rate video (28-38 dB), the IBMCTF scheme achieves comparable performance to the conventional SDMCTF. However, the difference comes when we decode less spatial resolutions; there, the steepness of the R-D curve of the SDMCTF system is significantly lower than the IBMCTF. In addition, it can be seen that the SDMCTF framework converges to a significantly lower PSNR value. The theoretical explanation of this effect is given in Appendix I for the interested reader.

5. CONCLUSIONS AND FUTURE EXTENSIONS

In this paper, we introduced a novel fully scalable framework for the compression of video sequences. This framework presents motion compensated temporal filtering techniques in the overcomplete wavelet domain (IBMCTF) for high performance and scalability. Unlike the conventional spatial-domain wavelet techniques, which apply MCTF on the original frame data and then encode the residual image using a critically-sampled wavelet decomposition, the framework presented here applies the MCTF separately for each subband after the spatial wavelet transform, thereby providing a fully scalable wavelet compression for video. Comparisons between SDMCTF and IBMCTF wavelet coding showed the higher coding efficiency performance of the proposed IBMCTF scheme as well as a much higher efficiency in spatial scalability for a variety of sequences. Further extensions of the proposed framework include employing different-accuracy motion estimation schemes and prediction structures for UMCTF in the various bands [2]. Moreover, different coding schemes can also be employed for the various bands for improved coding efficiency and complexity scalability.



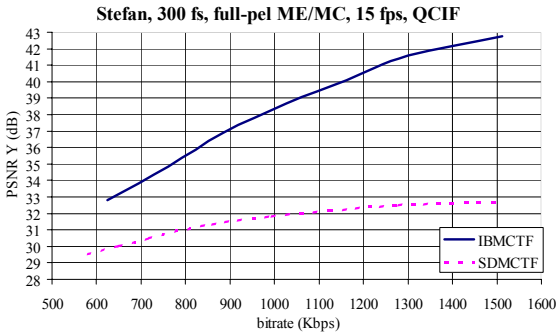
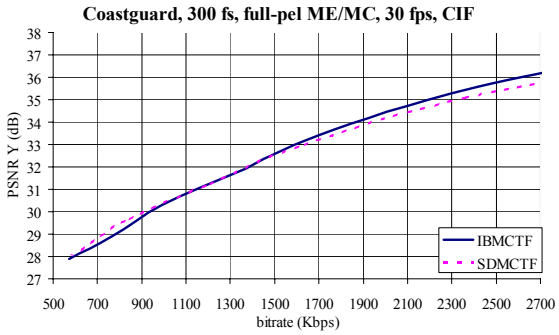


Fig. 5. Experimental results for the Stefan sequence.



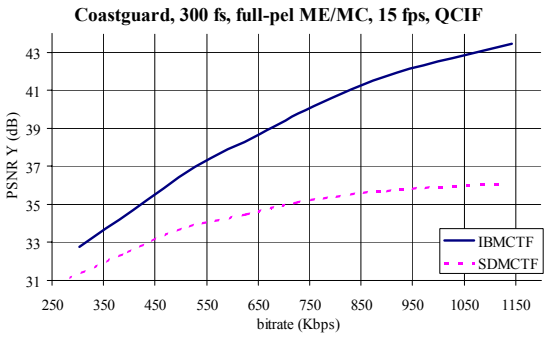


Fig. 6. Experimental results for the Coastguard sequence.

APPENDIX I

The case of a two-level resolution-scalable SDMCTF codec is analyzed below to illustrate the SDMCTF bottleneck for spatial scalability. In addition, although the mathematical description is restricted to the case of 1-D signals for simplicity in the notations, the same reasoning applies to the 2-D case.

The process starts at the encoder by the calculation of the error frame E^0 from the spatial-domain MC of frame B^0 that uses the reference frame A^0 and the corresponding motion-vector set $\mathbf{V}^0$ estimated at full-pixel accuracy from the original full-resolution frames:

$$E^0 = B^0 - \mathcal{MC}(A^0, \mathbf{V}^0)\,. \tag{1}$$

The superscripts always indicate the resolution level (zero equals original/full resolution). Following this notation, all decoded signals X are denoted as X_{dec}, while the subscript *low* or *high* indicates the low or high-frequency subband respectively. Finally, the decomposition and reconstruction low-pass filters are denoted by $U, \widetilde{U}$ respectively.

Under a resolution-scalable framework, the performance of the MC at half resolution in the decoder occurs as described below:

- *step A:* invert the decoded reference and corresponding residual error back to the spatial domain representation by using only the low-frequency subband (since the high-frequency subband is not received by the lower-resolution decoder),
- *step B:* perform the MC in the spatial domain – and
- *step C:* apply the one-level DWT in order to reconstruct the low-frequency subband of the current frame.

In this respect, for the lossless decoding of B^1_{low} (half-resolution frame), in the SDMCTF scheme, *step A* is performed as:

$$A^0_{dec}(z) = \widetilde{U}(z) A^1_{dec,low}(z^2)\,, \tag{2}$$

$$E^0_{dec}(z) = \widetilde{U}(z) E^1_{dec,low}(z^2)\,, \tag{3}$$

since in this case at the decoder we have $A^1_{dec,high} = E^1_{dec,high} \equiv 0$. As a result, *step B* can be written as:

$$B^0_{dec} = E^0_{dec} + \mathcal{MC}(A^0_{dec}, \mathbf{V}^0)\,, \tag{4}$$

Finally, the one-level DWT is performed (*step C*) as:

$$B^1_{dec,low}(z^2) = \tfrac{1}{2}\left(U(z)B^0_{dec}(z) + U(-z)B^0_{dec}(-z)\right). \tag{5}$$

It can be seen that when (2) and (3) are replaced in (4), the resulting expression cannot be factorized by $\widetilde{U}(z)$ due to the fact that $\mathcal{MC}(\widetilde{U}(z)A^1_{dec,low}(z^2)) \neq \widetilde{U}(z)\mathcal{MC}(A^1_{dec,low}(z^2))$, i.e filtering and the MC operation are not commutative. As a result, the replacement of the outcome of (4) in (5) does not give the one-level DWT of (1). This means that even if lossless coding is performed for $A^1_{dec,low}$, $E^1_{dec,low}$, perfect reconstruction is not possible for $B^1_{dec,low}$ (low resolution output). In conclusion, *the SDMCTF creates drift for the resolution-scalable decoding to any level l>0.*

1. REFERENCES

[1] Y. Andreopoulos, A. Munteanu, G. Van der Auwera, P. Schelkens and J. Cornelis "Wavelet-based fine granularity scalable video coding with in-band prediction," *ISO/IEC JTC1/SC29/WG11*, *m7906*, MPEG 59th meeting, Jeju island, Korea, March 2002.

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

[3] T. Ebrahimi, M. van der Schaar, J.-R. Ohm, "The status of Interframe Wavelet Coding Exploration in MPEG," *ISO/IEC JTC1/SC29/WG11*, *n4928*, MPEG 61th meeting, Klagenfurt, Austria, July 2002.

[4] 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*, DSP 2002, Santorini, Greece, pp. 501-504, July 2002.

[5] 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*, to appear.

[6] H.-W. Park and H.-S. Kim, "Motion Estimation Using Low-Band-Shift Method for Wavelet-Based Moving-Picture Coding," *IEEE Trans. Image Processing*, vol. 9, no. 4, pp. 577-587, Apr. 2000.

[7] D. Turaga, M. van der Schaar and B. Pesquet-Popescu, "Differential coding of motion vectors in the MCTF framework," *ISO/IEC JTC1/SC29/WG11*, *m9035*, MPEG 62nd meeting, Shanghai, China, October 2002.