Title:
OPEN-LOOP, IN-BAND, MOTION-COMPENSATED TEMPORAL FILTERING FOR OBJECTIVE FULL-SCALABILITY IN WAVELET VIDEO CODING

Source: Yiannis Andreopoulos\textsuperscript{1}, Mihaela van der Schaar\textsuperscript{2}, Adrian Munteanu\textsuperscript{1}, Joeri Barbarien\textsuperscript{1}, Peter Schelkens\textsuperscript{1}, Jan Cornelis\textsuperscript{1}

\textsuperscript{1}Vrije Universiteit Brussel \hspace{1cm} \textsuperscript{2}Philips Research USA

Subgroup: Requirements and Video

Purpose:
Proposal for the AHG on Interframe Wavelet Coding
(Response to Exploration Experiments EE1 of document w4926)

Abstract
In this contribution, we investigate the alternative of performing the wavelet decomposition in the spatial direction before the temporal filtering step for open-loop motion-compensated temporal-filtering (MCTF) systems. In order to efficiently perform the temporal decomposition in the produced spatial wavelet representation of the input frames, we utilize the framework presented in a previous contribution [1], which involves wavelet-domain (in-band) motion estimation and compensation (ME/MC). Additionally, in order to evaluate the trade-offs between the in-band MCTF (IBMCTF) and the conventional spatial-domain MCTF (SDMCTF) systems, we utilize the framework of unconstraint MCTF (UMCTF) proposed in [2]. In the UMCTF framework, by allowing temporal decompositions without the update lifting step, one can attain a set of undistorted frames for the lower temporal resolutions, thereby allowing PSNR comparisons between the different alternatives. This gives a framework for objective assessment of the performance of the two solutions for all the degrees of freedom for scalability, i.e temporal, spatial and SNR.
Introduction – Objective assessment of scalability

Previous experiments in the context of fully-scalable video coding have identified three potential approaches [3]:

**Category A:** "Wavelet in loop" : Conventional coder structure, replacement of DCT by Wavelet transform of the residual error in the MC prediction loop.

**Category B:** "In-band prediction" : Spatial Wavelet transform for each frame performed first, then exploitation of interframe redundancy by prediction of Wavelet coefficient values, or by definition of temporal contexts in entropy coding. This should also include MC.

**Category C:** "Interframe Wavelet" : Wavelet filtering along the temporal axis followed by 2D spatial wavelet transform or vice versa. This is possible with or without MC. Also, temporal contexts can be utilized for entropy coding.

In the present contribution we try to combine the merits of the systems of category B into the framework of category C. Notice that the systems of category C can potentially operate in the transform domain and such an example is given in [4]. Nevertheless, the ME/MC process of that case is usually restricted to the LL subband [4] and in addition it uses only the critically-sampled part of the transform, which contains considerable aliasing due to the downsampling operation of the wavelet transform. In contradiction to this, the in-band prediction techniques of category B [1] are typically utilizing *overcomplete wavelet representations* for the purpose of wavelet-domain ME/MC and separate motion vectors are generated per resolution level, directly from the low or high-frequency subband content. This allows the tuning of the motion-vector bandwidth according to the desired budget and also permits the purely independent operation of the system across different spatial resolutions, potentially allowing for more efficient resolution scalability.

One bottleneck for the objective assessment of the performance of a generic MCTF-based video coder at lower temporal and/or spatial levels is associated to the update step performed during the temporal filtering. This is pictorially shown in Figure 1. There, the simple example of the Haar temporal filtering is shown for one temporal decomposition level.

It can be seen from Figure 1 that the produced H-frames are simply the MC prediction error from the ME between each original B-frame and its previous A-frame. After this step the update step is performed [5] and the original A-frames are replaced with the L-frames, which contain unchanged areas (that potentially went through a scaling operation) and parts that contain a weighted sum of the original with the prediction error of the H-frame, warped-back by the update lifting-step coefficient and the inverted motion vector [5]. The pixels belonging to each of these two areas are called unconnected and connected pixels respectively. These L-frames consist the temporally-downsampled frames (half frame-rate) and are further decomposed to achieve the desired temporal decomposition levels. In addition, after the temporal decomposition is completed, the spatial wavelet decomposition generates the desired number of spatial levels that can be independently compressed and packed in the embedded bitstream.
This procedure leads to the conclusion that each coder that utilizes a different transform in the temporal direction or a different motion estimation process, will generate a unique set of uncoded L-frames, due to the existence of connected pixels. As a result, this does not permit PSNR comparisons between two systems for lower frame-rates and lower spatial levels.

Nevertheless, following the reasoning of [2], one can opt for the use of a temporal decomposition without the update lifting step. The example of such a temporal decomposition is shown in the right side of Figure 2 and it must be noted that it still ensures the invertibility of the temporal transform. In this manner, the uncoded output at each different temporal or spatial decomposition consists of the original A frames or the LL subband of their spatial wavelet decomposition respectively. Naturally, this uncoded information can be used as a common reference between two systems, as long as a fixed filter-bank is used for the spatial decomposition.

This allows the objective comparison between two different approaches. In addition, it ensures that the uncoded result at a given temporal or spatial level will be visually pleasing, which is not always the case for the current MCTF schemes that include the update lifting step in the temporal decomposition [2].
SDMCTF and IBMCTF – the chosen temporal decomposition

In this section we present the basic systems for SDMCTF and IBMCTF. The IBMCTF structure is based on the extension of the classical MCTF concept in the wavelet domain. For an introduction to the basic spatial-domain MCTF and UMCTF systems, the reader is referred to [5] and [2].

As explained in the previous section, by using the temporal filtering without the update step, we can produce a temporal decomposition of the form seen in Figure 3. There, an example of 8 input frames is seen and for each row \( i = 1, 2, 3 \) (temporal decomposition level) each original \((k \cdot 2^i)\)-th A-frame (with \( k=1, 2, ..., 8/2^i \) for the particular example) is predicted by its neighboring A-frames (original frames) using bi-directional ME/MC. Then it is replaced, as seen in Figure 3, by the produced error frame \( H^i \). It can be noticed that with the absence of the update step that inverts the prediction-error information back to the A-frames, the scheme of Figure 3 is essentially predictive in its nature [2].

For the encoder, the procedure occurs in a top-bottom manner; the first ME/MC step is denoted in the dotted circles of Figure 3. Conversely, the decoder operates in a bottom-up manner; the procedure is initiated by receiving the motion vectors of the temporal level 3 and the error-frame \( H^3 \), at which point the reconstructed frame at position 5 is immediately created. Similarly the motion vectors of temporal level 2 are received along with the error frames \( H^2 \) and so on. In order to have also resolution scalable decoding in this process, the error frames of each temporal level are decoded in a resolution scalable manner,

Figure 2. An example of the MCTF scheme without the update lifting step [2].
following also a bottom-up decoding (coarse-to-fine decomposition level). In this way, temporal and spatial scalability are completely independent of each other. For example, the full temporal pyramid can be reconstructed using only the 3\textsuperscript{rd} spatial level of the wavelet decomposition thereby generating full frame-rate video at reduced resolution. Conversely, if one reconstructs in the decoder only the 3\textsuperscript{rd} temporal level but uses all the spatial resolution levels, video at quarter frame-rate and full resolution is received.

<table>
<thead>
<tr>
<th>Temporal level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. A basic form of MCTF scheme without the update lifting-step.

**The SDMCTF system**

Up to the point of temporal scalability no distinction exists between the two systems of interest, i.e SDMCTF and IBMCTF. The differences occurs on the actual process of ME/MC in the codec structure. Starting with the encoder, in the SDMCTF, motion estimation is performed during the temporal decomposition in the spatial domain using the full-resolution neighboring A-frames prior to the spatial decomposition. This vector-set is used for the production of the H-frames (spatial-domain error frames). Subsequently the spatial transform is applied to the produced H-frames and to the A frame of temporal level 3 and coding in a bottom-up manner occurs, i.e from coarse to fine spatial resolution.

At the decoder side, when a certain resolution level is decoded in the transform domain for each H-frame, the motion compensation can be performed in two ways:

- The first approach inverts both the references and the decoded error frame back to the spatial-domain representation (using zero as the contribution for the high-frequency subbands that have not been decoded) and then uses the fully-accurate motion-vector set to produce the reconstructed current A-frame. Afterwards, the spatial decomposition occurs to reconstruct the spatial resolution level required for decoding.

- The second approach inverts the decoded transform representation of the references and the error frame only up to the point where the LL subband of the immediate finer-resolution level is constructed. Then, this approach uses the subsampled motion vectors (with the subsampling of the current resolution level) to reconstruct the predicted frame of the current resolution level.
From the two possible approaches we opted for the first approach in our experiments. Although it represents a more complex method, since it requires inverting back to the full-resolution and using the same amount of motion-vector data for each resolution level, from the theoretical point of view it resolves the ambiguity of applying subsampled motion vectors to the low-frequency subband (which is subsampled and contains considerable aliasing in comparison to the full resolution). Nevertheless, it must be noted that even in this case, all the lower-resolution decoders reconstruct a spatial-domain representation that lacks certain high-frequency content (which was captured by the spatial DWT into the high-frequency subbands that were not decoded). As a result, even though the motion vectors are used for the motion compensation in the reconstructed full-resolution spatial-domain representation, it is highly probable that in this case these vectors are not representing the optimal block-matching operation under the utilized distortion measure, since the motion estimation process was utilizing the full-frequency content of the full-resolution video frames.

The IBMCTF system

The other alternative for MCTF, starts by performing the spatial transform to each input frame. The wavelet representation of each input frame is used during the temporal-filtering stage. In particular, the motion estimation process is performed in a multiresolution manner.

The proposed framework utilizes a classical biorthogonal filter-pair (like the 9/7 filter-pair) for the spatial transform, with a maximum of \( k=3 \) decomposition levels. In every case, for the performance on ME, the overcomplete discrete wavelet transform (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 bit-stream. The construction of the ODWT from the DWT, a procedure that is a complete-to-overcomplete discrete wavelet transform (CODWT) \([7]\), occurs at both the encoder and decoder side for the reference frame(s). In this contribution 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 \( l \) we have \( LH^m_{(0,0)} = HL^m_{(0,0)} = HH^m_{(0,0)} = \emptyset \) for \( l < m \leq 1 \), 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 \([8]\), or more advanced techniques that use a single-rate calculation scheme with reduced computational and delay overhead \([7]\). Since this process is performed in the decoder side as well and it occurs separately for each resolution level and for each reference frame, fast and single-rate calculation schemes are imperative for an optimized, low-complexity implementation.
After the construction of the ODWT, for each level \( l \) we produce a set of critically-sampled subbands \( LH_{i,j}^l, HL_{i,j}^l, HH_{i,j}^l, 0 \leq i < 2^l, 0 \leq j < 2^l \) (and also \( LL_{i,j}^l \) 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 \( LH_{(0,0)}^l, HL_{(0,0)}^l, HH_{(0,0)}^l \) of the current frame that correspond to the same spatial-domain location with a triplet of blocks from each of the \( LH_{i,j}^1, HL_{i,j}^1, HH_{i,j}^1, 0 \leq i < 2^l, 0 \leq j < 2^l \) of the reference(s). For the coarsest resolution level, the motion estimation process is performed separately for the \( LL \) subband. In this approach, similar to the spatial-domain methods, variable block sizes and search ranges can be used per resolution level. In addition, the motion-vectors of different resolutions can be correlated in order to limit the search region, however this is also going to affect the motion estimation quality in comparison to the full-freedom search per resolution 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 LTI operators and hence their order of application to the input signal can be interchanged.

A simple pictorial example of the previously described ME process is given in the left part of Figure 4 for the case of the low-frequency subband of an one level decomposition of one reference frame. In this case, the 2-D ODWT of the reference frame contains four critically-sampled low-frequency subbands and \( 3 \times 4 \) high-frequency critically-sampled subbands. The gray-shaded blocks indicate the search area in every \( LL_{i,j}^1 \) subband of the reference, with \( i, j = \{0,1\} \). In this example, we show that the best match is found in \( LL_{(0,0)}^1(x,y) \) and the corresponding vector is \( MV_{low}(x,y) \). Notice that apart of encoding this motion vector, one needs also to encode the corresponding subband-index \((0,1)\) in which the best match is found. 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 Figure 4 (right side), the motion compensation is done in this example with the minimum 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 process 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.

To estimate the increase in the bitrate from 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 \( B_x \times B_y \) 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 \( r_a \) reference frames, for each frame of the SDMCTF system, the bit budget for the motion vectors is:
\[ R_{SD} = 2 \left[ \frac{R \cdot C}{\log_2 r} \right] \left[ \log_2 (2 \cdot r + \log_2 r) \right]. \]

Similarly, following the fact that for the first level of the wavelet decomposition there is a downsampling factor of 2, 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. \( \frac{b_x}{2} \times \frac{b_y}{2} \) 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 bit budget for the motion vectors of each frame for spatial-decomposition level \( l \) is:

\[ R_{IB}^{\text{lev}} (l) = 2f(l) \left[ \frac{R \cdot C}{(b_x \cdot b_y) \cdot 4} \right] \left[ \log_2 (2 \cdot r \cdot 2^{-l}) + 2^{l-1} + \log_2 r \right], \]

where \( f(l) = \{1:1 \leq l < k, \ 2:l = k\} \) 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:

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

By the comparison of the \( R_{SD} \) and the corresponding \( R_{IB}^{\text{tot}} (k) \) for various resolution levels \( k \), it can be seen that \( R_{IB}^{\text{tot}} (k) \leq 2R_{SD}, \forall k \). In addition, following the first approach of the previous section for the 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:

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

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, \( R_{IB}^{\text{res}} (l) \leq \frac{R_{SD}}{2} \) for \( l>1 \), i.e the motion vector bitrate for the lower-resolution decoders is at least half of the corresponding of the spatial-domain case. In this contribution we do not further extend the discussion on the possible coding modes for the multiresolution motion vectors. This is still an open field for the interframe wavelet video coding, hence many ideas are applicable [10] [11].

![Figure 4. Left: the ME process for a block of the LL subband. Right: The MC process with the motion vectors for an 1-level decomposition.](image-url)
**Experimental evaluation**

This contribution aims to provide some insight for the open problem of the relative tradeoffs of the application of the spatial transform before the temporal filtering in MCTF systems for full scalability. As seen until this point, during the course of our exploration we opted for the use of MCTF schemes without the update step, in order to have a fixed reference frame for the comparison of the two alternatives (spatial transform first or temporal transform first). In this way we can have an objective comparison for lower resolution and lower frame-rate decoding. In addition, as a side-effect of our exploration effort, we found a path for the merging of two different categories of scalable systems in one, i.e categories B and C [3] of the introduction section. However, we must stress that the purpose of this contribution is *not* to propose yet another coding system for interframe wavelet coding, but rather to provide a *first objective evaluation under the same conditions* of the in-band MCTF that utilizes the wavelet-domain ME/MC with the ODWT, versus the framework of spatial-domain MCTF.

To this extend, for the two different systems -SDMCTF and IBMCTF- we used a basic test-bench system that uses block-based, pixel-accurate ME/MC with a fixed block size and search range. Each of the two schemes compresses the produced H-frames with the same embedded intra-band wavelet coder [6]. To ensure resolution scalability, the coder is used separately per resolution level and a target mean-square error is sought in the transform representation of each resolution level. Hence the presented results stem from an implementation that is distortion controlled, rather than bitrate controlled. Nevertheless, it is rather straightforward to associate the distortion in the wavelet-domain with the produced bitrate [6] and hence make the current implementation operating under strict bitrate control. For both systems the same MSE-based control was used, both used four temporal decomposition levels and three levels for the spatial transform. In addition, for both schemes the bi-directional A-H-A temporal decomposition scheme from [2] was used. In fact, apart from the different processing order (spatial decomposition followed by temporal decomposition) 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 our preliminary test purposes we focused on relative high bitrates. This is done for three reasons:

- The possible optimizations in the motion-vector coding are not investigated in this contribution. Even if the same coding approach is used for both systems, the comparison is still inaccurate for low bitrate coding because there the motion-vector bitrate becomes significant and any potential gain of a method that exploits the multiresolution structure of the IBMCTF for the coding might change the outcome of the comparison. As a result, we simply calculated the uncoded motion-vector bits-per-frame for each method and assumed a typical compression ratio for both cases, i.e 20% for full resolution. In addition, for decoding at a lower resolution level \( l, 1 \leq l \leq k \), a compression ratio of \((20 \cdot 2^{l-1})\)% was assumed for the vectors of the in-band ME/MC.
• For low bitrate coding, one needs to resolve to high sub-pixel accuracy (1/4 or 1/8 pixel accuracy) in order to achieve acceptable visual quality. In addition, the use of variable block-sizes helps the motion estimation process in producing higher quality at lower bitrates.

• We are more interested in examining the asymptotic behavior of both schemes under comparison. In specific, 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 curves at all resolutions and frame-rates and whether they can converge to a very low distortion for high bitrate. In this manner we can establish a firm answer as to whether there are any fundamental problems with resolution scalability in the existing MCTF framework and whether more complex techniques, such as the IBMCTF should be pursued.

We present our PSNR results in the Figure 5-Figure 9, for four CIF sequences of first priority, as decoded for full resolution/full frame-rate, half resolution/half frame-rate, quarter resolution/quarter frame-rate. Figure 5 shows that although for the remaining A-frames of the last temporal level (level 4) both schemes achieve exactly the same distortion (since the same coder and distortion control is used), the prediction in the SDMCTF fails notably. Also, from Figure 6-Figure 9 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 for all the sequences tested. 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, which always 5-10 dB lower than the IBMCTF. Although the visual difference may not be so notable at the QCIF or Q-QCIF resolution, it is possible that this behavior will result in poor visual quality for the SDMCTF when HD material is used.

To assess better the visual quality of the smaller resolutions of this experiment, we extracted two examples from QCIF resolution (Y components) and interpolated them to 704x576 resolution using the same bicubic interpolation. The results are seen in Figure 10 and Figure 11. It can be seen that in these example the fast moving regions are visually represented better in the IBMCTF framework.

![Figure 5. PSNR comparison for QCIF decoding of the Stefan sequence at 590 Kbps (Y component, first 50 frames). IBMCTF: red dotted curve, SDMCTF: blue solid curve.](image-url)
Figure 6. Mobile sequence.
Figure 7. Foreman sequence.
Figure 8. Stefan sequence.
Figure 9. Coastguard sequence.
Figure 10. A visual example (zoom with interpolation) from the extracted QCIF stefan sequence at 600 Kbps. Top: IBMCTF, bottom: SDMCTF.
Figure 11. An example of the Coastguard sequence, interpolated from the QCIF decoding to 704x576 (top: IBMCTF, bot.:SDMCTF).
In Figure 12 we show a typical comparison between the visual output of the IBMCTF and the SDMCTF for three of the sequences used at full resolution (CIF). It can be seen that the wavelet-domain motion estimation does not produce the blocking artifacts that are typical for the low bitrate SDMCTF decoding of fast moving regions of video. Instead ringing is produced, however we found these effects much more acceptable as a visual distortion.

The explanation of this difference in the behavior of the two systems is simple; in the IBMCTF, after the motion compensation process, blocking artifacts exist in the wavelet-domain representation of the H-frames. As a result, after the decompression and the inverse transform, they become ringing due to the dispersion of the information from the wavelet filters. Although this is not a very significant advantage from the decoding point of view since deblocking methods can be applied at the decoder side for the SDMCTF, it can be considered significant for the encoding. In fact, we find that part of the coding gain seen in the IBMCTF scheme comes from the existence of the blocking artifacts in the wavelet domain; if a quadtree coding strategy is used for the error-frame coding, such as the intra-band coders of [6] or the EZBC coder [4], the adaptive block partitioning made by the coder during the bitplane parsing matches the motion compensated block size. Hence the “artificial” high-frequency content of the blocking artifacts is efficiently handled by the entropy coding.

On the other hand, in a system that does the temporal decomposition first by making the motion compensation in the spatial domain, after the application of the spatial decomposition the blocking artifacts become high-frequency content for which the utilized coder allocates a portion of the bitstream. This problem cannot be easily handled by using an adaptive deblocking strategy in the encoder (as it is done in fixed-rate coders) because in the interframe wavelet framework the deblocking cannot be adapted to a certain quantization step-size, since embedded coding that can compress up to near-lossless is used.

**Conclusions and future extensions**

We performed a first evaluation of the use of the spatial transform before the temporal decomposition under the framework of in-band ME/MC using the ODWT. The presented approach leads to an increase in the motion-vector bandwidth for the full-resolution due to the performance of multiple ME/MC, however the increase is limited. Additionally, the system complexity increases in both sides of the codec since the decoder of each resolution level independently performs a CODWT operation. However, it has been shown objectively and under the same experimental conditions that the presented IBMCTF framework can lead to improved performance for lower resolution video, while providing a scalable reduction in the motion information per resolution level.
Figure 12. Visual comparison of the decoded visual quality (CIF resolution) at the same bitrate (left: IBMCTF, right: SDMCTF).
References