

REMOVING REDUNDANCY IN MULTI-RESOLUTION SCALABLE VIDEO CODING SCHEMES

Guillaume Boisson, Edouard François

THOMSON R&D France,
1 avenue de Bellefontaine, CS 17616,
F-35576 Cesson-Sévigné, France

ABSTRACT

Nowadays standard technologies for spatially scalable video coding use Gaussian pyramidal approaches, that naturally lead to redundant descriptions after the temporal analysis. However solutions have been proposed to preserve the critical sampling criterion within a multi-resolution framework. That research area is highly worth the interest since the redundancy suppression potentially improves compression efficiency. We focus here on two different solutions, with a special focus on the spectral composition of the transmitted information. Last some results are presented and discussed.

Index Terms— Video coding, Wavelet transforms

1. INTRODUCTION

A number of video coding schemes that use (t+2D) sub-band approach have been proposed so far [1, 2, 3, 4, 5]. These schemes present state-of-the-art compression performances, but suffer from strong limitations concerning spatial scalability. Indeed spatio-temporal low-frequency frames present poor quality, even at maximal bit-rate.

Besides, numerous other schemes [6, 7, 8] have explored a different track : (2D+t) wavelet coding. As far as scalability is concerned, that architecture is very well designed, but unfortunately, compression expectations are not met, especially at high resolution.

Considering the relative failure of these wavelet-based solutions to supply a spatially scalable bit-stream on e. g. three scales of resolution while keeping competitive, the video compression community turned its efforts toward multi-resolution techniques (Cf. [9]). These simulcast-like solutions take as source's input Gaussian pyramids, and deliver therefore redundant descriptions. The only feature that distinguishes them from pure-simulcast techniques is the possibility to use inter-layer prediction. Yet the resulting bit-stream remains redundant, harming the potential compression efficiency.

In this paper we investigate the issue of redundancy and critical sampling within that particular framework of multi-resolution scalable video coding. Solutions have already been proposed to design critically-sampled video codecs while keeping multi-resolution approaches advantages. After having defined some notations, we'll focus in section 3 on the LBC (*Low-Band Correction*) method, proposed by Han in [10], then in section 4 on the (2D+t+2D) scheme of Mehrseresht & Taubman [11, 12], pointing out differences and similarities. Last results are shown in comparison with state-of-the-art video coding schemes.

2. NOTATIONS

Information relative to spatial resolutions will figure in superscript, whereas information relative to temporal levels and instants will fig-

ure in subscript. Conventionally, the zero index refers to the original source's characteristics. Last bold face will denote sequences : $\mathbf{f}_t^p = \{f_{t,k}^p\}_{k \in K}$.

Temporal analysis and synthesis will be denoted $\mathbf{a} = \begin{pmatrix} \mathbf{a}_t \\ \mathbf{a}_b \end{pmatrix}$ and $\mathbf{s} = \begin{pmatrix} \mathbf{s}_t \\ \mathbf{s}_b \end{pmatrix}$. At a given temporal level t , the \mathbf{f}_t^p sequence is transformed into its high- and low-frequency half-signals : $\mathbf{h}_t^p = \mathbf{a}_b(\mathbf{f}_t^p)$ and $\mathbf{l}_t^p = \mathbf{a}_t(\mathbf{f}_t^p)$. The iterative filtering process goes further on, considering $\mathbf{f}_{t+1}^p = \mathbf{l}_t^p$.

Regarding motion compensation, the index of the current frame (the predicted frame) is pointed by the reference frame's (the warped one's) index, respecting chronological order. E.g. in the well-known 5/3 MCTF framework, each high-frequency frame h_k is obtained by subtracting from the f_{2k+1} frame the half-sum of $\mathcal{W}_{2k \rightarrow 2k+1}(f_{2k})$ and $\mathcal{W}_{2k+1 \rightarrow 2k+2}(f_{2k+2})$.

Last, let $\mathcal{A} = \begin{pmatrix} \mathcal{A}_L \\ \mathcal{A}_H \end{pmatrix}$ and $\mathcal{S} = \begin{pmatrix} \mathcal{S}_L \\ \mathcal{S}_H \end{pmatrix}$ respectively stand for the spatial analysis and synthesis. We'll denote too $\mathcal{L}\mathbf{f}^p = \mathcal{A}_L(\mathbf{f}^p)$ and $\mathcal{H}\mathbf{f}^p = \mathcal{A}_H(\mathbf{f}^p)$ the low and high frequencies of sequence \mathbf{f}^p .

We can trivially state that :

$$\begin{aligned} \mathcal{S} \circ \mathcal{A} &= \begin{pmatrix} \mathcal{S}_L & \mathcal{S}_H \end{pmatrix} \cdot \begin{pmatrix} \mathcal{A}_L \\ \mathcal{A}_H \end{pmatrix} \\ &= (\mathcal{S}_L \circ \mathcal{A}_L) + (\mathcal{S}_H \circ \mathcal{A}_H) = \mathcal{I}d_{\mathcal{P}ixel} \end{aligned}$$

And :

$$\begin{aligned} \mathcal{A} \circ \mathcal{S} &= \begin{pmatrix} \mathcal{A}_L \circ \mathcal{S}_L & \mathcal{A}_L \circ \mathcal{S}_H \\ \mathcal{A}_H \circ \mathcal{S}_L & \mathcal{A}_H \circ \mathcal{S}_H \end{pmatrix} \\ &= \begin{pmatrix} \mathcal{I}d_{\mathcal{B}\mathcal{F}} & 0 \\ 0 & \mathcal{I}d_{\mathcal{H}\mathcal{F}} \end{pmatrix} = \mathcal{I}d_{\mathcal{F}req} \end{aligned}$$

3. EMBEDDED SIMULCAST CODING WITH THE LBC METHOD

3.1. Encoding and embedding

The LBC (*Low-Band Correction*) method has been presented in [10] and derives from an idea proposed in [13]. It resorts on the use of two linear operators of down- and up-sampling (denoted \mathcal{D} and \mathcal{U} in [10]), verifying $\mathcal{D} \circ \mathcal{U} = \mathcal{I}d$. Translated to the present notations, these operators become \mathcal{A}_L and \mathcal{S}_L .

In the following we limit our explanations to a group of two frames at two different resolutions, but the scheme can be trivially extended to more levels and frames.

Let f_{2k}^0 and f_{2k+1}^0 be two consecutive frames at original resolution, and f_{2k}^1 and f_{2k+1}^1 their sub-sampled versions (Cf. Fig. 1).

LBC's multi-resolution temporal analysis is purely predictive and does not comprise any update step. At every resolution, high frequency frames are computed by forward prediction of odd frames :

$$\{ h_k^i = f_{2k+1}^i - \mathcal{W}_{2k \rightarrow 2k+1}^i (f_{2k}^i) \} \quad (1)$$

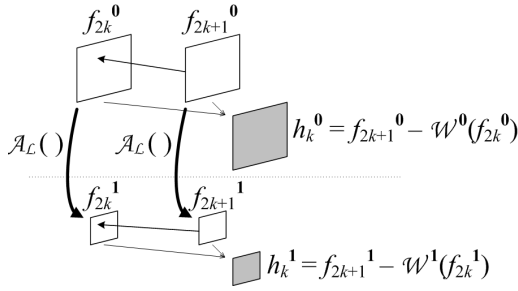


Fig. 1. Multi-Resolution Analysis.

Let's focus now on the embedding process. Trivially the reference frame contains its sub-sampled versions $\{f_{2k}^i = \mathcal{A}_L^i(f_{2k}^0)\}_{i \geq 1}$; therefore it is transmitted at original resolution. Besides, prediction error frames are merged, putting h_k^1 in place of spatial low-frequencies of h_k^0 . Written in the manner of [10], the transmitted pieces of information are :

$$\begin{cases} l_k = f_{2k}^0 \\ h_k = h_k^0 - (\mathcal{S}_L \circ \mathcal{A}_L)(h_k^0) + \mathcal{S}_L(h_k^1) \end{cases}$$

And in terms of spatio-temporal sub-bands (Cf. Fig. 2) :

$$\begin{cases} f_{2k}^1 \text{ and } \mathcal{H}f_{2k}^0 \\ h_k^1 \text{ and } \mathcal{H}h_k^0 \end{cases} \quad (2)$$

Yet that expression is foolish, because, unlike for the intra-coded frame, for temporal high-frequencies, we have :

$$h_k^1 \neq \mathcal{A}_L(h_k^0) \quad \text{and consequently} \quad h_k^0 \neq \mathcal{S} \begin{pmatrix} h_k^1 \\ \mathcal{H}h_k^0 \end{pmatrix}.$$

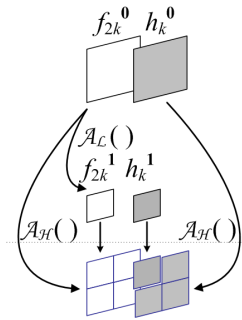


Fig. 2. Multi-Resolution Embedding Process.

3.2. Spectral composition of transmitted information

A key-point of scalable multi-resolution and (t+2D) wavelet-based schemes lies in the way spatial and temporal transforms are associated. Therefore it is highly worth studying the spectral composition of transmitted information, especially concerning the spatial and temporal highest frequencies (i.e. (2) bottom-right) :

$$\mathcal{H}h_k^0 = \mathcal{H}f_{2k+1}^0 - (\mathcal{A}_H \circ \mathcal{W}_{2k \rightarrow 2k+1}^0)(f_{2k}^0) \quad (3)$$

3.3. Multi-resolution decoding

Lowest resolution decoding process is obvious. Reference frame f_{2k}^1 is decoded, then f_{2k+1}^1 is reconstructed from h_k^1 and $\mathcal{W}_{2k \rightarrow 2k+1}^1(f_{2k}^1)$:

$$\begin{cases} f_{2k}^1 = \mathcal{A}_L(l_k) \\ f_{2k+1}^1 = \mathcal{W}_{2k \rightarrow 2k+1}^1(f_{2k}^1) + \mathcal{A}_L(h_k) \end{cases}$$

Upper-resolution decoding process is less trivial. First, the whole original-resolution reference frame is reconstructed from f_{2k}^1 and $\mathcal{H}f_{2k}^0$. Then the point is to recover the whole prediction error h_k^0 , from which only the spatial high-frequencies have explicitly been transmitted (Cf. (2)). Fortunately it is possible to extract missing h_k^0 's low-frequencies from lowest-resolution frame f_{2k+1}^1 together with the motion field related to $\mathcal{W}_{2k \rightarrow 2k+1}^0$ and the reconstructed frame f_{2k}^0 . Indeed, according to (1) :

$$\begin{aligned} \mathcal{L}h_k^0 &= \mathcal{A}_L(f_{2k+1}^0 - \mathcal{W}_{2k \rightarrow 2k+1}^0(f_{2k}^0)) \\ &= f_{2k+1}^1 - (\mathcal{A}_L \circ \mathcal{W}_{2k \rightarrow 2k+1}^0)(f_{2k}^0) \end{aligned} \quad (4)$$

Upper-resolution synthesis equations are thus (Cf. Fig. 3) :

$$\begin{cases} f_{2k}^0 = \mathcal{S} \begin{pmatrix} f_{2k}^1 \\ \mathcal{H}f_{2k}^0 \end{pmatrix} \\ f_{2k+1}^0 = \mathcal{W}_{2k \rightarrow 2k+1}^0(f_{2k}^0) + \mathcal{S} \begin{pmatrix} f_{2k+1}^1 - (\mathcal{A}_L \circ \mathcal{W}_{2k \rightarrow 2k+1}^0)(f_{2k}^0) \\ \mathcal{H}h_k^0 \end{pmatrix} \end{cases}$$

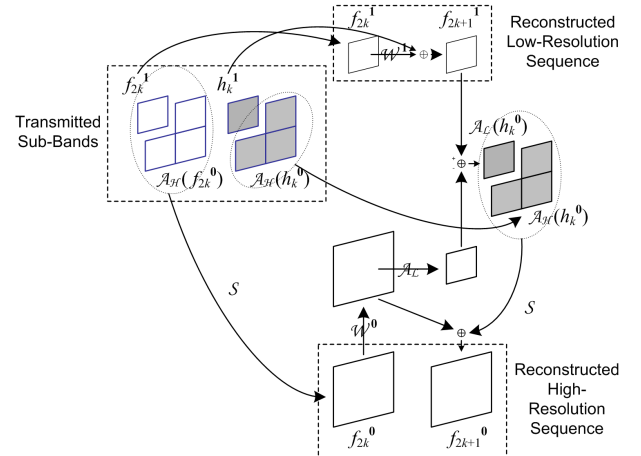


Fig. 3. Multi-Resolution Synthesis.

Note that this decoding policy prevents from introducing an update lifting step without jeopardizing the reconstruction of the missing spatial low-frequencies (Cf. (4)).

4. THE (2D+T+2D) METHOD

4.1. Principle and encoding architecture

Let's consider the simplest configuration, with one spatial level before and after temporal analysis. On Figure 4, SST and MC-TST refer to *Spatial Sub-band Transform* and *Motion-Compensated Temporal Sub-band Transform*.

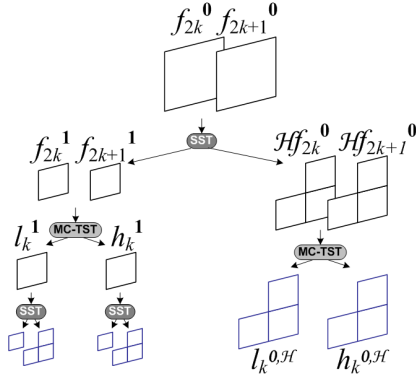


Fig. 4. Spatio-temporal analysis process.

The lowest resolution MC-TST process is classically performed, with a predict then an update step :

$$\begin{cases} h_k^1 &= f_{2k+1}^1 - \mathcal{W}_{2k \rightarrow 2k+1}^1 (f_{2k}^1) \\ l_k^1 &= f_{2k}^1 + \mathcal{W}_{2k \leftarrow 2k+1}^1 (h_k^1) \end{cases}$$

Regarding upper resolution, note that the filtering process affects high-frequencies $\mathcal{H}f^0$ only. Yet motion compensation is performed in pixel-domain to circumvent wavelet's shift-variance. This is done by modifying the classical warping stage into a new operator $\mathcal{W}^{\mathcal{H}}$:

$$\mathcal{W}^{p,\mathcal{H}} = \mathcal{A}_{\mathcal{H}} \circ \mathcal{W}^p \circ \mathcal{S}_{\mathcal{H}} \quad (5)$$

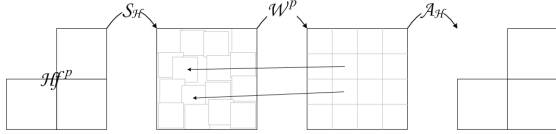


Fig. 5. Motion Compensation for high-frequency bands.

4.2. Motion Compensation and Inter-band “leakages”

Actually, the intra-band compensation (5) does not efficiently decorrelate the signal along time. Indeed block-based motion compensation is highly non-linear and induces information “leakages” between neighboring bands. Mehrseresht & Taubman respectively denoted them “type I” and “type II” according they affect higher or lower frequencies. Of course these phenomena, depicted on Figure 6, dramatically reduce the compression efficiency.

Type II leakages can not be compensated without altering the encoding’s reversibility at lower resolutions, because high frequency data used at analysis can not be available at decoder side.

Also for perfect reconstruction purpose, we focus here on type I leakages compensation. In [11] authors propose to modify the warping operator $\mathcal{W}^{\mathcal{H}}$ to take into account the information contained in every lower band (Cf. Fig. 7) :

$$\widetilde{\mathcal{W}}^{p,\mathcal{H}}(\cdot) = (\mathcal{A}_{\mathcal{H}} \circ \mathcal{W}^p \circ \mathcal{S}) \left(f^{p+1} \right) \quad (6)$$

Obviously that technique can only be introduced in predict step :

$$\begin{cases} h_k^{0,\mathcal{H}} &= \mathcal{H}f_{2k+1}^0 - \widetilde{\mathcal{W}}_{2k \rightarrow 2k+1}^{0,\mathcal{H}} (\mathcal{H}f_{2k}^0) \\ l_k^{0,\mathcal{H}} &= \mathcal{H}f_{2k}^0 + \mathcal{W}_{2k \leftarrow 2k+1}^{0,\mathcal{H}} (h_k^{0,\mathcal{H}}) \end{cases}$$

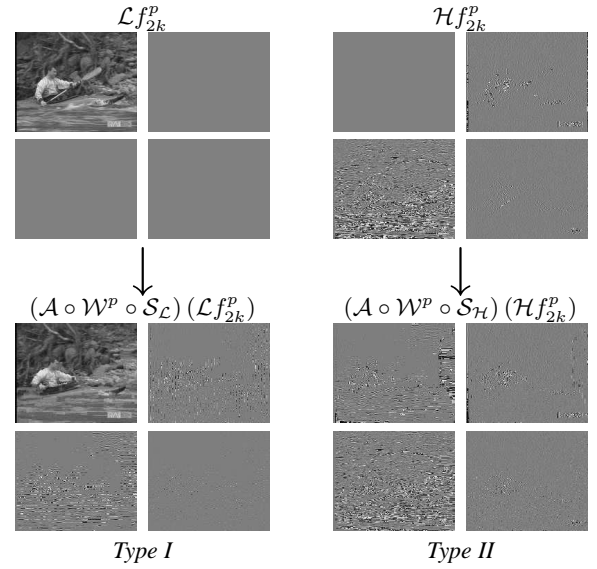


Fig. 6. Appearing energy due to “leakages”.

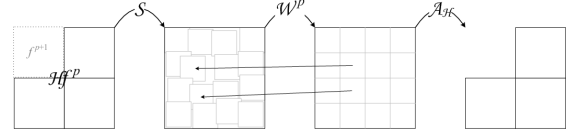


Fig. 7. Motion Compensation accounting for type I leakages.

4.3. Synthesis equations

Whatever the resolution, the sequence is reconstructed by inverting the lifting steps in order to recover desired spatial frequencies :

$$\begin{cases} f_{2k}^1 &= l_k^1 - \mathcal{W}_{2k \leftarrow 2k+1}^1 (h_k^1) \\ f_{2k+1}^1 &= \mathcal{W}_{2k \rightarrow 2k+1}^1 (f_{2k}^1) + h_k^1 \\ f_{2k}^0 &= \mathcal{S} \left(l_k^{0,\mathcal{H}} - \mathcal{W}_{2k \leftarrow 2k+1}^{0,\mathcal{H}} (h_k^{0,\mathcal{H}}) \right) \\ f_{2k+1}^0 &= \mathcal{S} \left(\widetilde{\mathcal{W}}_{2k \rightarrow 2k+1}^{0,\mathcal{H}} (f_{2k}^0) + h_k^{0,\mathcal{H}} \right) \end{cases}$$

4.4. Spectral composition of transmitted information

In the (2D+t+2D) scheme, the transmitted information is :

$$\begin{cases} h_k^1 &\text{and } h_k^{0,\mathcal{H}} \\ l_k^1 &\text{and } l_k^{0,\mathcal{H}} \end{cases}$$

When type I leakages are compensated, we have :

$$\begin{aligned} h_k^{0,\mathcal{H}} &= \mathcal{H}f_{2k+1}^0 - (\mathcal{A}_{\mathcal{H}} \circ \mathcal{W}_{2k \rightarrow 2k+1}^0 \circ \mathcal{S}) \left(\begin{matrix} f_{2k}^1 \\ \mathcal{H}f_{2k}^0 \end{matrix} \right) \\ &= \mathcal{H}f_{2k+1}^0 - (\mathcal{A}_{\mathcal{H}} \circ \mathcal{W}_{2k \rightarrow 2k+1}^0) (f_{2k}^0) \end{aligned} \quad (7)$$

In light with (3), we see here that (2D+T+2D)’s sub-band $h_k^{0,\mathcal{H}}$ and LBC’s sub-band $\mathcal{H}h_k^0$ have the same spectrum.

That means that in spite of very different approaches, Han’s LBC and Mehrseresht & Taubman’s (2D+T+2D) do remain the same compression technique (apart from the presence of an update step). Second, this proves it’s possible to implement true orthogonal MCTF (with update) within an embedded multi-resolution framework.

5. RESULTS

5.1. Discussion

Regarding encoding, LBC and (2D+t+2D)'s main difference is that at each resolution the latter affects only spatial high-frequencies with temporal filtering. But the embedding stage and the $\tilde{W}^{\mathcal{H}}$ operator overpass that difference and :

$$\begin{array}{ccc} \mathcal{H}h_k^0 & = & h_k^{0,\mathcal{H}} \\ \parallel & & \parallel \\ (\mathcal{A}_{\mathcal{H}} \circ \mathbf{a}_h) (\mathbf{f}^0) & & (\tilde{\mathbf{a}}_h \circ \mathcal{A}_{\mathcal{H}}) (\mathbf{f}^0) \end{array}$$

Concerning decoding, one could say that LBC aims at “temporally” reconstructing f_{2k+1}^0 from f_{2k}^0 and h_k^0 – at the expense of an update-uncompliant “trick” to recover missing information. Besides, (2D+T+2D) scheme consists in “spatially” reconstructing each resolution, from which frequencies $h_k^{p,\mathcal{H}}$ and $l_k^{p,\mathcal{H}}$ were explicitly transmitted. In other words, with the same sub-band, LBC performs $(\mathfrak{s} \circ S)$ whereas (2D+T+2D) performs $(S \circ \mathfrak{s})$.

Last it is worth noting that inter-resolution prediction, that significantly improves LBC's performances when the source lacks temporal correlation, is strictly equivalent to pure intra-coding during predict-step of (2D+t+2D).

5.2. Performance

Embedded multi-resolution solutions significantly challenge classical (t+2D) approaches such as [5] (5/3 MCTF + JPEG-2000) (Cf. Fig. 8). Compared to previous sub-band approaches, the gap with highly-mature redundant multi-resolution techniques, such as the JSVM [9], has been noticeably reduced. Considering that critically-sampled solutions present potential improvements in comparison with pyramidal solutions, they can be expected to reach equivalent performances with a significant complexity reduction, especially at the encoder side.

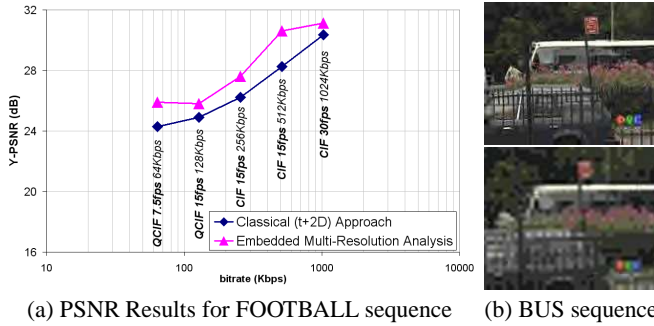


Fig. 8. Embedded MR-Analysis vs Classical (t+2D) approach: objective results and visual quality.

6. CONCLUSION

We described and compared two spatially scalable video coding solutions. They're both based on multi-resolution analysis approach, however they present critical sampling, unlike most contemporary pyramidal techniques. Though their principles are dual, we showed that they lead to the same sub-band decomposition, the latter allowing to introduce an update step in the temporal analysis process. The

described methods present also high compression efficiency together with a wide range of resolution scalability.

7. REFERENCES

- [1] J.-R. Ohm, “Three-dimensional subband coding with motion compensation,” *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 559–571, Sept. 1994.
- [2] S.-T. Hsiang and J.W. Woods, “Invertible three-dimensional analysis-synthesis system for video coding with half-pixel-accurate motion compensation,” in *SPIE Conference on Visual Communication and Image Processing*, San Jose, California, 1999, pp. 537–546.
- [3] J. Xu, Z. Xiong, S. Li, and Y.-Q. Zhang, “Three-dimensional embedded subband coding with optimal truncation (3-D ES-COT),” *Journal Applied and Computational Harmonic Analysis, (Special Issue On “Wavelet applications in Engineering”)*, vol. 10, pp. 290–315, May 2001.
- [4] B. Pesquet-Popescu and V. Bottreau, “Three-dimensional lifting schemes for motion compensated video compression,” in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP'01*, Salt Lake City, Utah, May 2001.
- [5] G. Boisson, E. François, and C. Guillemot, “Accuracy scalable motion coding for efficient scalable video compression,” in *Proceedings of 11th IEEE International Conference on Image Processing, ICIP'2004*, Singapore, Oct. 2004.
- [6] 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, Apr. 2000.
- [7] Y. Andreopoulos, A. Munteanu, G. Van der Auwera, J. Barbarien, P. Schelkens, and J. Cornelis, “Wavelet-based fine granularity scalable video coding with in-band prediction,” in *ISO/IEC JTC1/SC29/WG11 MPEG2002/M7906*, Jeju Island, Korea, Mar. 2002.
- [8] G. Boisson, E. François, D. Thoreau, and C. Guillemot, “Motion-compensated spatio-temporal context-based arithmetic coding for full-scalable video compression,” in *Proceedings of Picture Coding Symposium, PCS 2003*, Saint Malo, France, Apr. 2003.
- [9] H. Schwarz, D. Marpe, and T. Wiegand, “MCTF and scalability extension of H264/AVC,” in *Proceedings of Picture Coding Symposium, PCS 2004*, San Francisco, California, Dec. 2004.
- [10] W.-J. Han, “Responses of Call-for-Proposal for Scalable Video Coding,” in *ISO/IEC JTC1/SC29/WG11 MPEG2004/M10569/S17*, Muenchen, Germany, Mar. 2004.
- [11] N. Mehrseresht and D. Taubman, “Spatial scalability and compression efficiency within a flexible motion compensated 3D-DWT,” in *Proceedings of 11th IEEE International Conference on Image Processing, ICIP'2004*, Singapore, Oct. 2004.
- [12] N. Mehrseresht and D. Taubman, “An efficient content-adaptive MC 3D-DWT with enhanced spatial and temporal scalability,” in *Proceedings of 11th IEEE International Conference on Image Processing, ICIP'2004*, Singapore, Oct. 2004.
- [13] T. Kimoto and Y. Miyamoto, “Multi-Resolution MCTF for 3D Wavelet Transformation in Highly Scalable Video,” in *ISO/IEC JTC1/SC29/WG11 MPEG03/M9770*, Trondheim, Norway, July 2003.