1

State-of-the-art and trends in scalable video compression with wavelet based approaches

Nicola Adami, *Member, IEEE,* Alberto Signoroni* *Member, IEEE* and
Riccardo Leonardi, *Member, IEEE*

Abstract

This work describes the current state-of-the-art in scalable video coding (SVC), focusing on wavelet based motion-compensated approaches. After recalling the requirements imposed by multiple forms of video scalability (quality, picture size, frame rate) which typically exist jointly, it discusses individual components that have been designed to address the problem over the years. Therefore presentation shows how such components are typically combined to achieve meaningful architectures for video compression, which differ from the space-time order in which the wavelet transform operates, discussing strengths and weaknesses of the resulting implementations. The paper explains the Wavelet Video Reference architecture(s) studied by ISO/MPEG in its exploration on Wavelet Video Compression. It also attempts to draw a list of major differences between wavelet based solutions and the emerging SVC standard, jointly targeted by ITU and ISO/MPEG (JVT-SVC), based on MPEG-4 AVC technologies. A major emphasis is devoted to a WSVC solution, named STP-tool, which presents architectural similarities with respect to JVT-SVC. The presentation continues by providing performance comparisons between the different approaches, and draws some indications on the future trends being researched by the community to further improve current wavelet video codecs. Insights on application scenarios which could benefit from a wavelet based approach are provided.

Index Terms

Scalable video coding, wavelets, spatiotemporal multiresolution representations, motion compensated temporal filtering, entropy coding, video coding architectures, MPEG, video quality assessment

* Corresponding author. Email: alberto.signoroni@ing.unibs.it - Address: Università degli Studi di Brescia, Dipartimento di Elettronica per l'Automazione, via Branze, 38, I-25123, Brescia, Italy - Tel. +39 030 3715432, Fax. +39 030 380014.

All authors are with the Department of Electronic for Automation - Faculty of Engineering - University of Brescia - Brescia - Italy

I. INTRODUCTION

Traditional single operating point video coding systems can be surpassed using scalable video coding (SVC) architectures where, in a single bit-stream, a number of decodable streams can be extracted corresponding to various operating points in terms of spatial resolution, temporal frame rate or reconstruction accuracy. Different scalablity features can coexist in a single video coded bit-stream with coding performance approaching the state-of-art single point coding techniques. This has been more and more a trend in the last years [1] and it has become a reality thanks to the development of SVC systems derived either from hybrid schemes [2] (used in all MPEG-x or H.26x video coding standards) or from spatiotemporal wavelet technologies [3]. A great part of this development and relative debates regarding SVC requirements [4], applications and solutions has been carried out by people participating in the ISO/MPEG standardization.

This paper offers an overview of existing SVC architectures which are based on multi-resolution spatiotemporal representation of video sequences. In particular, wavelet-based SVC (WSVC) tools and systems will be classified and analyzed with the following objectives: a) to give a quite comprehensive tutorial reference to those interested in the field, b) to analyze strong and weak points of the main WSVC tools and architectures, c) to synthesize and account for the exploration activities made by the MPEG video group on WSVC and to describe the issued reference platform, d) to compare such platform with the ongoing JVT-SVC standardization effort in qualitative and quantitative terms, and e) to discuss and propose promising evolution paths and target applications for WSVC. For this purpose the presentation has been structured as follows. In Sec. II WSVC fundamentals as well as basic and advanced tools which enable temporal, spatial and quality salability are presented and discussed. Sec. III shows how such components are typically combined to achieve meaningful WSVC architectures, which typically differ from the space-time order in which the wavelet transform operates, discussing strengths and weaknesses of the resulting implementations. In the same section, an emphasis is placed on a promising architecture which presents some similarities to the JVT-SVC standard. The corresponding solution, named STP-tool, implements an inter-scale prediction mechanism thanks to a single stage spatial wavelet decomposition in the temporal transform domain. Subsequently, Sec.IV explains the Wavelet Video Reference architecture(s) studied by ISO/MPEG in its exploration on Wavelet Video Compression. The paper attempts as well (Sec. V) to draw a list of major differences between such architecture(s) and tools with respect to the JVT-SVC reference model, providing performance comparison in terms of coding efficiency and giving a critical analysis of the presented results. In Sec.VI further investigations

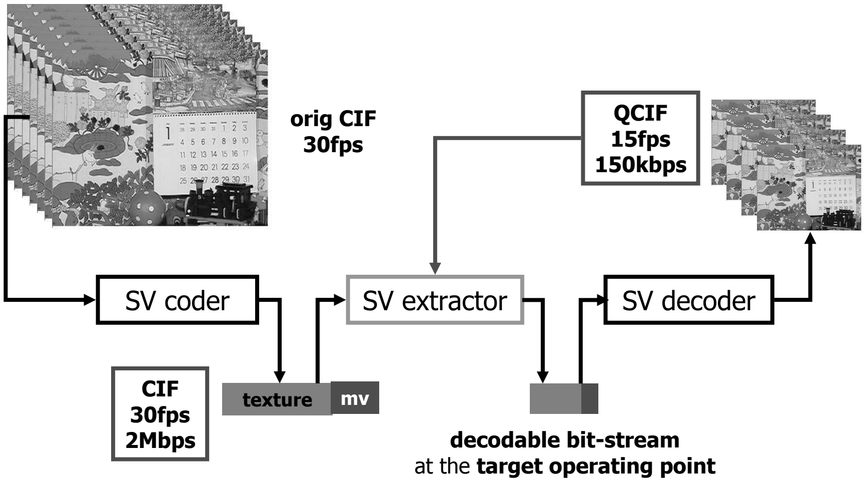


Fig. 1. Scalable Video Coding rationale. A unique bit-stream is produced by the Coder. Decodable sub-streams can be obtained with simple extractions.

on WSVC critical aspects are presented. A discussion on perspectives and applications that appear better aimed to wavelet based approach is made in Sec.VII. Finally, some conclusions are derived and an acknowledgment is addressed to all people that contributed to the WSVC state-of-the-art.

II. Basic principles of wavelet-based approaches to SVC

A. *WSVC fundamentals*

Fig.1, shows a typical SVC system which refers to the coding of a video signal at an original CIF resolution (288 height x 352 width) and a framerate of 30 fps. In the example, the highest operating point and decoding quality level correspond to a bit rate of 2Mbps associated with the data at the original spatiotemporal resolution. For a scaled decoding, in terms of spatial and/or temporal and/or quality resolution, the decoder only works on a portion of the originally coded bit stream according to the specification of a desired working point. Such a stream portion is extracted from the originally coded stream by a functional block called *extractor*. As shown in Fig.1 the extractor is arranged between the coder and the decoder. In any practical application, it can be implemented as an independent block or it can be an integral part of either the coder or the decoder. The extractor receives the information related to the desired working point – in the example of Fig.1, a lower spatial resolution QCIF (144 x 176), a lower frame rate (15 fps), and a lower bit-rate (quality) (150 kbps) – and extracts a decodable bit stream

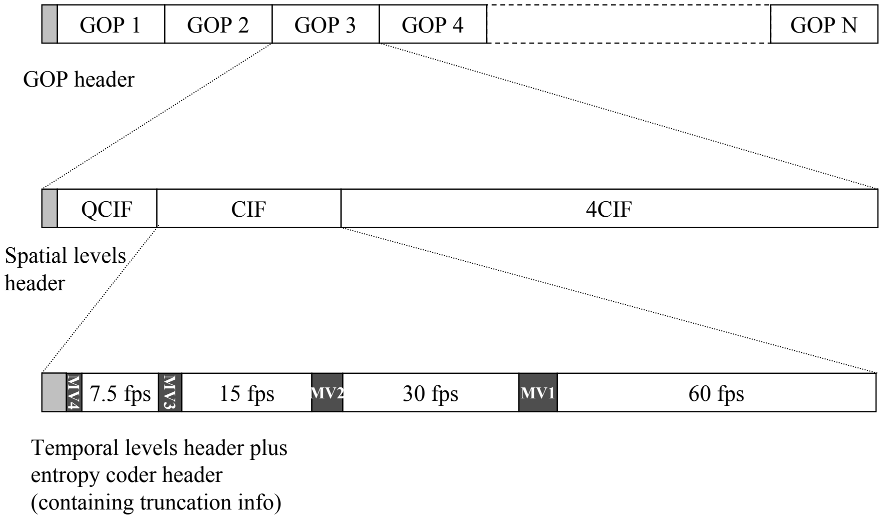


Fig. 2. Example of a scalable bit-stream structure.

matching or almost matching the working point specification. One of the main differences between an SVC system and a transcoding solution is the low complexity of the extractor which does not require coding/decoding operations and typically consists of simple parsing operations on the coded bit-stream.

For practical purposes, let us consider the following example where a 4CIF video (576x704) is coded allowing to extract three different spatial scalability layers and four embedded temporal resolutions. In this case, a possible structure of the bit-stream generated by an SVC system is shown in Fig.2. As indicated, the usual partition in independently coded groups of pictures (GOP) can be adopted for separating the various video frames. The GOP header contains the fundamental characteristics of the video and the pointers to the beginning of each independent GOP bit-stream. In this example, each GOP stream contains three bit-stream segments, one for each spatial level. Note that in an ideal scalable bit-stream, except for the lowest working point, all information is incremental. This means, for example, that in order to properly decode higher resolution (e.g. 4CIF), one needs to use the portions of the bit-stream which represent the lower spatial resolutions (e.g. QCIF and CIF). Given this bit-stream organization, for each spatial resolution level a temporal segmentation of the bit-stream can be further performed. Accordingly, the 7.5 fps portion contains data that allow to reconstruct the lowest temporal resolution of the considered spatial resolution reference video sequence. Similarly the 15fps, 30fps, 60fps parts refer to information used to refine the frame-rate until the allowed maximum. Since most successful video compression schemes require the use of motion compensation (MC), each temporal resolution level bit-stream portion should

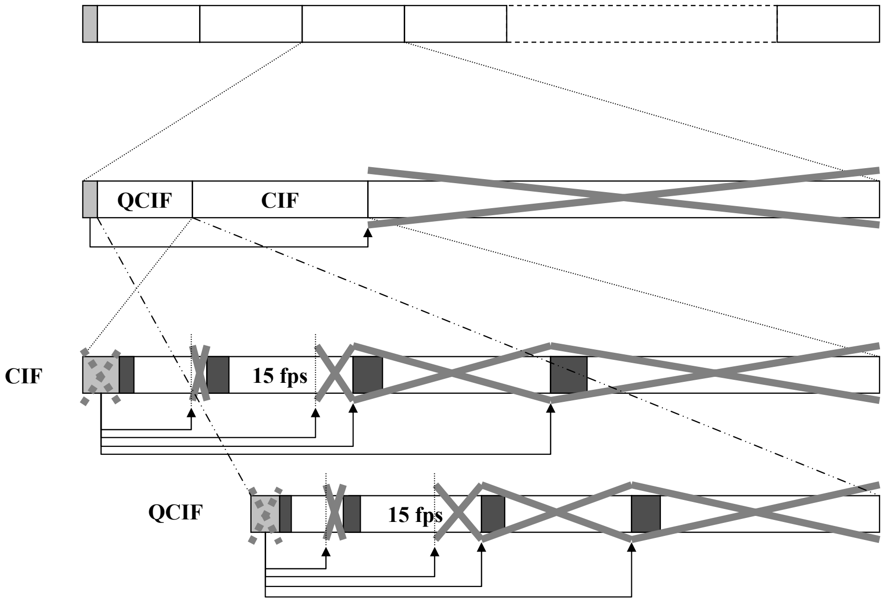


Fig. 3. Example of a sub-stream extraction: CIF at 15fps is extracted from a 4CIF sequence at 60fps.

include the relative motion information (usually consisting of coded motion vector fields), as shown in Fig.2. Further quality layers can be obtained, for example, by arranging for every temporal segment, the signal representation in a real or virtual bit-plane progressive fashion (see Sec.II-B.3). With such a bit-stream, every allowed combination in terms of spatial, temporal and quality scalability can be generally achieved and decided according to decoding capabilities. Fig.3 provides a graphical representation of a target bit-rate extraction of a CIF 15fps coded sequence, from the 4CIF 60fps bit-stream of the original video. Note that corresponding parts of QCIF and CIF bit-streams have been extracted to reconstruct the sequence at the desired working point.

A suitable property for concrete applications (e.g. for content distribution on inhomogeneous networks) is the possibility to adapt the bit-stream according to a multiple (chained) extraction path, from the highest towards lower operating points; this has been called bit-stream *multiple adaptation* [5]. Multiple adaptations should be allowed, of course, along every advisable path, and in general without any a-priori encoding settings. Multiple adaptation extraction differs from a single extraction only in one aspect; for multiple adaptations the bit-stream extraction information must be conserved (i.e. processed and re-inserted throwing out what become useless) from the source bit-stream. Clearly there is no need to re-insert (and spend bits for) the extraction information if no further extraction is expected from the latest extracted bit-stream.

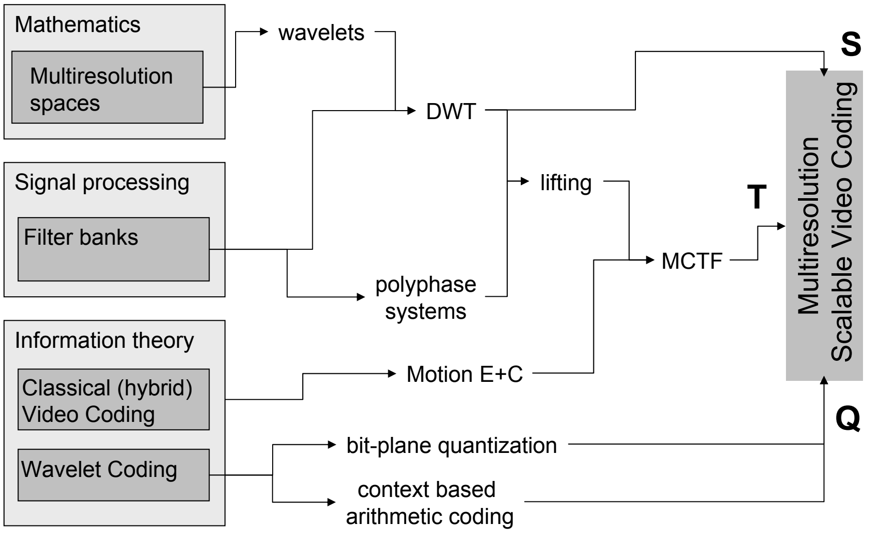


Fig. 4. WSVC theoretical context.

B. *Tools enabling scalability*

In the past two decades there has been a proliferation of activities focalized on theory and solutions for efficient (minimally redundant) multiresolution representation of signals. In general, a multiresolution signal decomposition (transform) inherently enables a corresponding low to high resolution scalability by the representation of the signal in the transformed domain. The tools that enable scalable multidimensional (space-time-quality) signal representation and coding are multidisciplinary. This fact is clearly depicted in Fig.4 that gives a quick overview of these tools and how they can be combined in order to achieve the desired scalability feature: Spatial, Temporal and Quality (S,T and Q). In the following, each scalability dimension, S,T and Q, will be considered and the most popular and effective solutions will be briefly described. In this context, Fig.4 can be helpful in keeping the big picture when analyzing a specific tool. Due to lack of space to describe each tool in detail, support references will be provided for the interested reader.

1) Spatial scalability: Tools producing a multiresolution representation of n-dimensional signals can be of different types. They can be linear or non linear, separable or not, redundant or critically sampled. Linear decompositions have been studied for a long time. The simplest mechanism to perform a two resolution representation of a signal is based on what here we call inter-scale prediction (ISP): a full

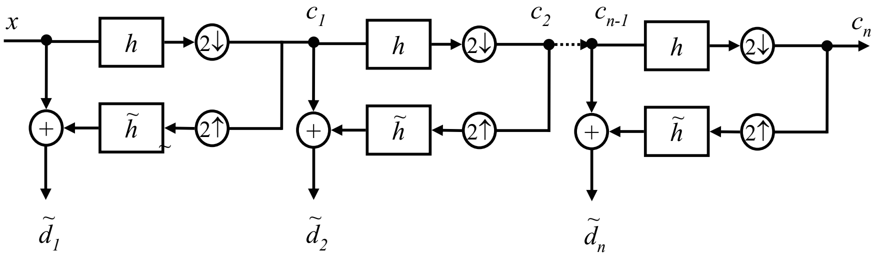


Fig. 5. A n-level Laplacian pyramid analysis filter bank.

resolution signal x can be seen as a *coarse* resolution signal c, obtained by a decimation of x (e.g., by filtering with a low-pass filter h and downsampling of factor 2), added to a *detail* signal $\tilde{d}$ calculated as the difference of the original signal and an interpolated version of c at full resolution, $\tilde{d} = x - \mathcal{I}(c)$ (throughout this document a $\sim$ symbol over a d signal means that the detail is not critically sampled). The Laplacian pyramid introduced by Burt and Adelson [6] is an iterated version of such an ISP mechanism resulting in a coarsest resolution signal c and a set of details $\tilde{d}(l)$ at decreasing levels of spatial resolution $l = 1 \ldots n$. Fig.5 shows a Laplacian decomposition of the one-dimensional signal x in n-stages (from full resolution 0 to the coarsest one n) where $\mathcal{I}(c_l)$ is implemented by upsampling by a factor of 2 and interpolating with $\tilde{h}$.

The discrete wavelet transform (DWT) is a linear operator which projects the original signal in a set of multiresolution subspaces [7] allowing a critically sampled (orthogonal or biorthogonal) representation of the signal in the transformed domain and guaranteeing perfect reconstruction synthesis. Similarly to the Laplacian pyramid, the DWT decomposition of x generates a coarse signal c and a series of detail signals $d(l)$ at various levels of spatial resolutions. By using the terminology of multirate filter banks [8], the transformed domain signals are also called subbands. In fact, the DWT can be implemented by means of a two-channel filter bank, iterated on a dyadic tree path, as shown in Fig.6 [9] [10]. Symbols $\tilde{h}$ and $\tilde{g}$ represent respectively the low-pass and high-pass analysis filters, whereas h and g are the synthesis for an orthogonal or biorthogonal wavelet decomposition. Symmetric filter responses are preferred for visual data coding applications due to their linear phase characteristics and can be obtained only with biorthogonal decompositions (except for length 2 filters, i.e. Haar basis). A popular filter pair for image coding purposes is the 9/7 one [10, pag. 279], which gives an analysis filter $\tilde{h}$ with 9 taps and a synthesis filter h with 7 taps.

For multidimensional signals, e.g. images, separate filtering on rows and columns are usually adopted

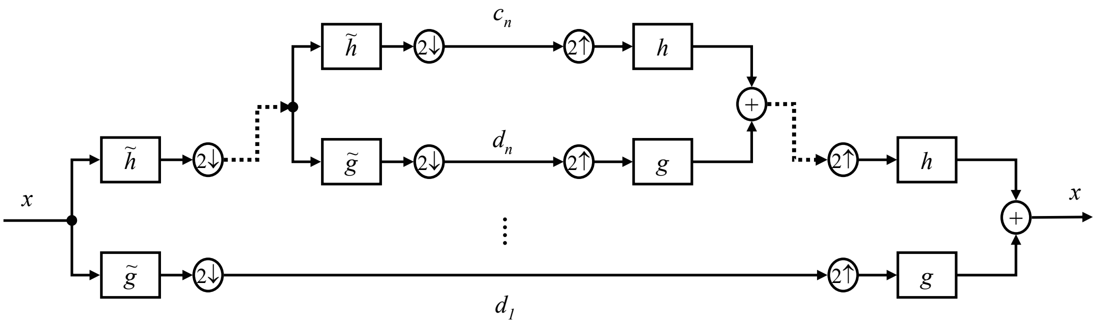


Fig. 6. A n-level (bi)orthogonal DWT analysis/sinthesis implemented with a dyadic tree filter bank.

to implement the so called *separable* pyramidal and DWT decompositions. In the case of separable 2D-DWT on images this generates, at each level of decomposition, one coarse subband c and three detail subband that we continue to indicate altogether as d.

Sweldens [11] introduced the *lifting scheme* as an alternative spatial domain processing to construct multiresolution signal representations. The lifting structure is depicted in Fig.7 and reveals a quite intuitive spatial domain processing of the original signal capable to generate a critically sampled (c, d) representation of signal x. According to this scheme, signal x is split in two polyphase components: the even and the odd samples x_{2i} and x_{2i+1}. As the two components (each one is half the original resolution) are correlated, a prediction P can be performed between the two; odd samples can be predicted with an operator which uses a neighborhood of even samples and that produces the residue

$$d_i = x_{2i+1} - P\{(x_{2i})_{i \in \mathbf{N}}\}. \tag{1}$$

The transform $x \rightarrow (x_{2i}, d_i)$ is not satisfactory from a multiresolution representation point of view as the subsampled signal x_{2i} could contain a lot of aliased components. As such, it should be *updated* to a filtered version c. The update step

$$c_i = x_{2i} + U\{(d_i)_{i \in \mathbf{N}}\} \tag{2}$$

can be constrained to do this job and it is structurally invertible as well as the preceding stages. Perfect reconstruction is simply guaranteed by

$$\begin{aligned} x_{2i} &= c_i - U\{(d_i)_{i \in \mathbf{N}}\} && (3) \\ x_{2i+1} &= d_i + P\{(x_{2i})_{i \in \mathbf{N}}\} && (4) \end{aligned}$$

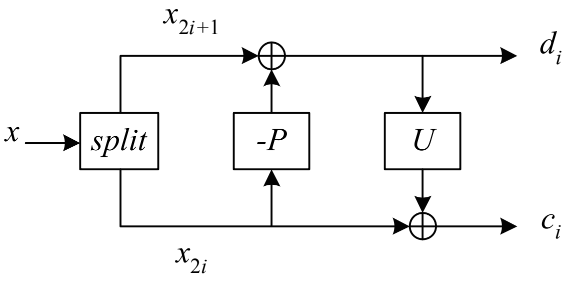


Fig. 7. Dyadic decomposition by lifting steps: split, prediction P and update U.

Daubechies and Sweldens [12] proved that every DWT can be factorized in a chain of lifting steps. The lifting structure is not limited to be an alternative and computationally convenient form to implement a DWT transform. Actually, it also allows non-linear multiresolution decompositions. For example, lifting steps implementing a DWT can be slightly modified by means of rounding operations in order to perform a so called integer to integer (non-linear) wavelet transform [13]. Such a transform is useful as a component of wavelet-based lossless coding systems. The kind and the degree of introduced non-linearity can depend on the structural or morphological characteristics of the signal that one may want to capture [14], keeping in mind that non linear systems can have unexpected behaviors in terms of error propagation (which may turn out to be a burden in presence of quantization of the transformed signal). Moreover, the lifting structure has a fundamental role for Motion Compensated Temporal Filtering (MCTF) as we will see in the next subsection.

2) Temporal scalability: A key tool which enables temporal scalability while exploiting temporal correlation is the MCTF. An introduction to the subject can be found in [15]. After the first work of Ohm proposing a motion compensated version of the Haar transform [16], other studies [17][18] began to show that video coding systems exploiting this tool could be competitive with respect to the classical and most successful hybrid ones (based on a combination of block-based spatial DCT transform coding and block-based temporal motion estimation and compensation). Other works (e.g. [19][20]) focus on the study of advanced MC lifting structures due to their versatility and strength in increasing the coding performance. Basically, an MCTF implementation by lifting steps can be represented by the already seen eqs.(1)-(4) where index i has now a temporal meaning and where the prediction and update operators P and U can be guided by motion information generated and coded after a motion estimation (ME) stage becoming MCP and MPU operators respectively. Fig.8 shows a conceptual block diagram of this general idea. ME/MC are implemented according to a certain motion model and in the case of spatiotemporal

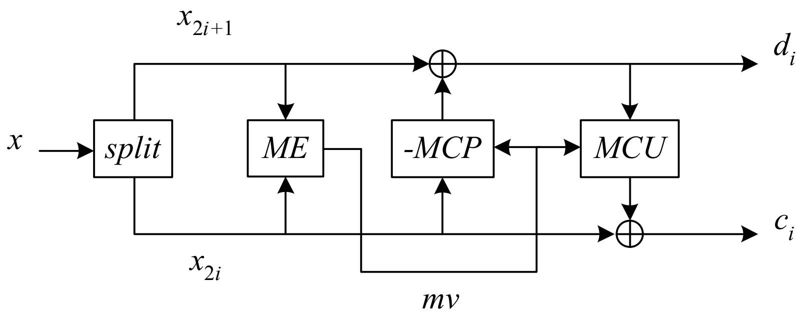


Fig. 8. The prediction and update lifting steps with explicit motion information conveyed to the P and U operators.

multi-resolution SVC and WSVC systems they usually generate a set of motion descriptions consisting of motion vector fields $mv(l, k)$. These are estimations of the trajectory of groups (usually rectangular blocks) of pixels between the temporal frames, at spatial level l, involved in the k^{th} MCTF temporal decomposition level.

Haar filters based [16][17][21] and longer filters MCTF solutions have been proposed [22][23][24]. For example MCTF implementations of the 5/3 transform exist [20]. 5/3 and in general longer filter solutions usually achieve better performance in terms of prediction efficiency and visual quality at lower temporal resolution frames, but make use of a bidirectional prediction and thus motion information is more bit-demanding than Haar based solutions. The two solutions recall unidirectional and bidirectional block mode prediction used in hybrid schemes. Similarly to advanced hybrid schemes, an adaptive selection among the two modes is also possible in the case of MCTF, leading to block-based adaptive MCTF solutions [25][26]. Alternatively, deformable motion models have also been explored with promising results [27].

By exploiting the local adaptability of the MCP and MCU operators and thanks to the transmitted (usually lossless coded) $mv(l, k)$ information, the MCTF can also be aware of and handle a series of issues affecting motion compensation: a) occlusion and uncovered area problems can be handled recognizing unconnected and multiple connected pixels and adapting the MCP and MCU structure accordingly [22][17][21]; b) in the case of block based ME/MC, blocking effects can be reduced in various ways by considering adjacent blocks (spatiotemporal lifting) realizing for example overlapped block-based motion compensation (OBMC) [21][26]; c) when MV's are provided with fractional pel precision the lifting structure can be modified to implement the necessary pixel interpolations [22][28][20] while preserving the temporal invertibility of the transform. It must also be noted that MCU makes use of a reversed motion field with respect to the one used by MCP. This may constitute a problem especially in those areas where inversion cannot be performed exactly, e.g. on occluded/uncovered areas, or does not fall on integer pixel positions because fractional pel accuracy is involved. In order not to duplicate the motion information and the associated bit-budget, mechanisms to derive or predict MCU fields from the MCP ones have been proposed [27][29][30] as well as methods that identify inversion problems and adopt interpolation based solutions (e.g. the Barbell lifting scheme [31]).

With lifting structures non dyadic temporal decomposition are also possible [32], leading to temporal scalability factors different from a power of two.

As already noticed, the update step could be omitted, without impairing perfect reconstruction, generating a so called *unconstrained* MCTF (UMCTF) [33]. This guarantees the direct use of the original frames for reduced frame-rates, then reduces the complexity of the transform stage and presents the benefit of eliminating ghosting artifacts. Ghosting artifacts can appear in some regions of the low pass temporal frames as a consequence of unavoidable local MC failures (associated to scene changes, unhandled object deformations, uncovered/occluded regions, incoming/outgoing objects). In such erroneously motion-compensated regions, the low pass temporal filtering updates the corresponding temporal subbands with "ghosts" (which "should not exist" in the current frame) coming from past or future frames. Depending on the entity of the ghosts, this can locally impair the visual appearance. On the other hand, for correctly motion compensated regions, the update step has a beneficial temporal smoothing effect that can improve visual perception at reduced frame-rates and also coding performance. Reduction of ghosting artifacts is then an important task and it has been obtained by the implementation of MCTF adaptive update steps [34][35] or by clipping the update signal [36].

Either MCTF or UMCTF generate a temporal subband hierarchy starting from higher temporal resolution to lower ones. Alternatively, the hierarchical B-frames decomposition (HBFD) [2] can be built enabling a similar temporal detail hierarchical structure but starting from the lower resolutions (predictions between more distant frames). This allows closed-loop implementations of the temporal prediction [37].

Efficient MCTF implementations have also been studied to reduce system complexity [38], memory usage [39][23] and coding delay [40].

3) Quality scalability: Among the best image compression schemes, wavelet-based ones currently provide for high rate-distortion (R-D) performance while preserving a limited computational complexity. They usually do not interfere with spatial scalability requirements and allow a high degree of quality scalability which, in many cases, consists in the possibility of optimally truncating the coded bit-stream at arbitrary points (bit-stream embedding). Most techniques that guarantee an optimal bit-stream embedding and consequently a bit-wise quality scalability are inspired from the zerotree idea first introduced by Shapiro's embedded zerotree wavelet (EZW) technique [41] and then reformulated with the set partitioning in hiearchical trees (SPIHT) algorithm by Said and Pearlman [42]. A higher performance zerotree inspired technique is the embedded zero-block coding (EZBC) algorithm [43], where quad-tree partitioning and context modeling of wavelet coefficients are well combined. The zerotree idea can also be reformulated in a dual way, allowing to directly build significant coefficients maps. To implement this idea, a morphological processing based on connectivity analysis has been used and also justified by the statistical evidence of the energy clustering in the wavelet subbands [44]. The EMDC technique [45][46] exploits principles of morphological coding and guarantees, for 2D and 3D data [47], coding performance comparable or superior to state-of-the-art codecs, progressive decoding and fractional bit-plane embedding for a highly optimized bit-stream.

Another popular technique, which does not use the zerotree hypothesis, is the embedded block coding with optimized truncation (EBCOT) algorithm [48], adopted in the JPEG2000 standard [49], which combines layered block coding, fractional bit-planes [50], block based R-D optimizations, and context-based arithmetic coding to obtain good (adjustable) scalability properties and high coding efficiency.

Proposed techniques for coding coefficients of spatiotemporal subbands generated by WSVC systems usually are extensions of image coding techniques. Obviously such 3D extensions shall not interfere with temporal scalability requirements as well, and can be achieved by considering a) single temporal frame coding with multiple frame statistical contexts updating, b) extended spatiotemporal non significance (or significance) prediction structures, c) spatiotemporal statistical contexts, or a combination of them. Temporal extensions of SPIHT [18], EZBC [51][52], EBCOT [36][24] and EMDC [53] have been proposed and used in WSVC video coding systems.

III. Coding architectures for WSVC systems

A. *WSVC notation*

In a multidimensional multilevel spatiotemporal decomposition there is a proliferation of signals and/or subbands. To support our discussion we have adopted a notation that guarantees easy referencing to the generated signals in a general decomposition structure. We consider a generic spatiotemporal signal $x = x(\underline{\theta}, t)$, with D spatial dimensions $\underline{\theta} = \{\theta_1, \ldots, \theta_D\}$. Signal x which undergoes an n-level multiresolution spatial transform $\mathcal{S}(n)$ is indicated with $x_{\mathcal{S}(n)}$. For non redundant transforms, such as the DWT, the spatially transformed signal actually consist of the subband set $x_{\mathcal{S}(n)} = \{x^{c}_{\mathcal{S}(n)}, x^{d(n)}_{\mathcal{S}(n)}, \ldots, x^{d(1)}_{\mathcal{S}(n)}\}$, where superscript c indicates the signal representation at the coarsest spatial resolution (low-pass) while, for each spatial level $\bar{n} \in [1 \ldots n]$, $d(\bar{n})$ indicates a critically sampled level of detail. In the case of separable transforms (such as the DWT) each $d(\bar{n})$ consists of $2^D - 1$ directional subbands. For redundant transforms (such as the Laplacian pyramid) the transformed signal consists of the signal set $x_{\mathcal{S}(n)} = \{x^c_{\mathcal{S}(n)}, x^{\tilde{d}(n)}_{\mathcal{S}(n)}, \ldots, x^{\tilde{d}(1)}_{\mathcal{S}(n)}\}$ where $\tilde{d}(\bar{n})$ is an oversampled detail at level $\bar{n}$. With a similar notation a spatiotemporal signal $x = x(\underline{\theta}, t)$ which undergoes a m-level multiresolution temporal transform $\mathcal{T}(m)$ is indicated as $x_{\mathcal{T}(m)}$. As before, superscripts c, $d(\cdot)$ and $\tilde{d}(\cdot)$ will be used referring to the temporal dimension to indicate coarsest temporal resolution, critically and not critically sampled temporal details respectively. In the figures of the present section we will graphically represent temporally transformed data subbands by associating a gray level to them: the white corresponds to c while grey-tones get darker as the detail level goes from the finest to the coarsest. Moreover, x and $\hat{x}$ indicate original and quantized (reduced coefficient representation quality) signals respectively. Symbol "*" will be used as wild card to refer to the whole set of spatiotemporal subbands of a given spatial or temporal transform.

A *decoded version* of the original signal x extracted and reconstructed at given temporal resolution $\bar{k}$, spatial resolution $\bar{l}$ and reduced quality rate, will be indicated as ${}^{\bar{l}}_{\bar{k}}\hat{x}$.

Concerning ME/MC, given a certain motion model, a description is produced consisting of a set of motion vector fields $mv(l,k)$, where, l and k refer to various spatial and temporal resolution levels respectively. As $mv(l,k)$ and $mv(l+1,k)$ are obviously not independent, mechanisms of up- or down-conversion, with or without refinement or approximation of the motion vector field, are usually implemented. Then, for example, $mv(l+1,k)$=$mv(l,k)_d$ indicates that motion vectors at a coarser resolution have been down-converted, by a suitable reduction of the block size and of the vector module, without approximation. Otherwise, in case of down-conversion with approximation the notation $mv(l,k)_{d,a}$ would have been used. Moreover, $mv(l,k)$=$mv(l+2,k)_{u(2),r}$ indicates that motion vectors at a 2-nd layer of finer resolution have been up-converted twice with a refinement of the block partition and/or of the motion vector value.

B. *Basic WSVC architectures*

In Figs.9 and 10 the main multi-resolution decomposition structures and corresponding WSVC architectures are shown. Each one will be presented in the following paragraphs.

1) t+2D: Perhaps the most intuitive way to build a WSVC system is to perform an MCTF in the original spatial domain followed by a spatial transform on each temporal subband. This is usually denoted as a t+2D scheme, it guarantees critically sampled subbands and it is represented in Fig.9(a). Earlier wavelet based coding systems were based on this scheme [16][17]. Many other wavelet based SVC systems are based on the t+2D spatiotemporal decomposition, among which we can cite

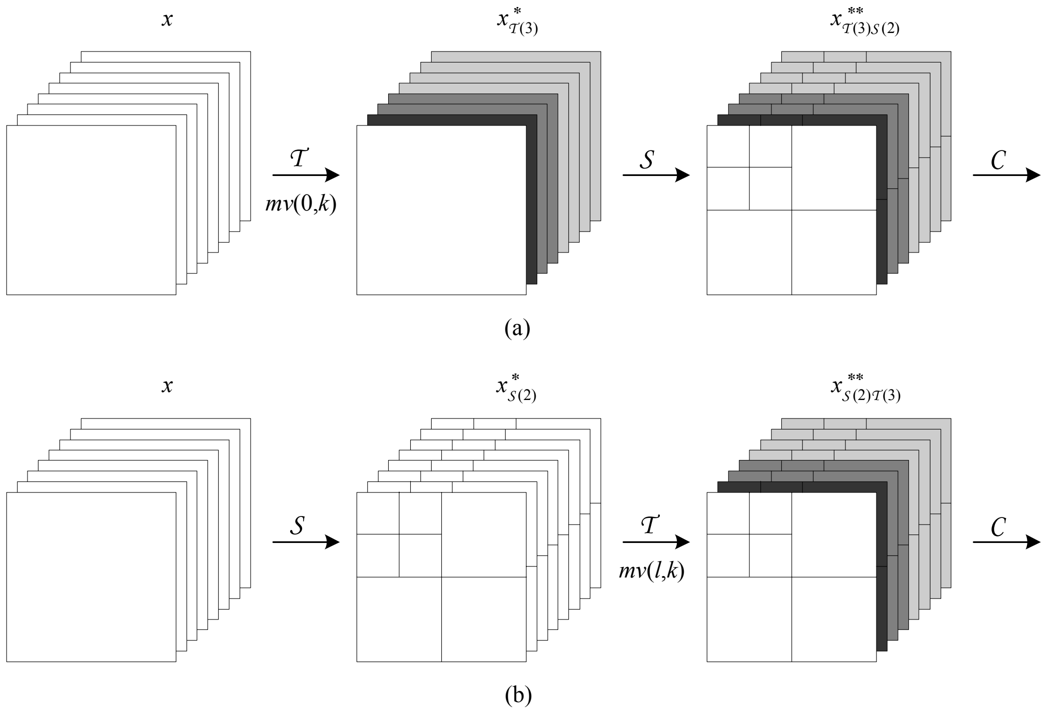

x
$\xrightarrow[mv(0,k)]{\mathcal{T}}$
$x^{*}_{\mathcal{T}(3)}$
$\xrightarrow{\mathcal{S}}$
$x^{**}_{\mathcal{T}(3)\mathcal{S}(2)}$
$\xrightarrow{\mathcal{C}}$
(a)
x
$\xrightarrow{\mathcal{S}}$
$x^{*}_{\mathcal{S}(2)}$
$\xrightarrow[mv(l,k)]{\mathcal{T}}$
$x^{**}_{\mathcal{S}(2)\mathcal{T}(3)}$
$\xrightarrow{\mathcal{C}}$
(b)

Fig. 9. Basic WSVC schemes in a signal perspective: t+2D (a), 2D+t (b).

[18][19][20][21][31][33][51][52][54][55][56][57][58].

Despite its conceptual simplicity, the t+2D solution presents some relevant drawbacks especially for spatial scalability performance. When only temporal and quality scalability (full spatial resolution decoding) are required, the process is reversed until reaching the desired frame-rate at resolution $\bar{k}$ (partial or complete MCTF inversion) and SNR quality: ${}_{\bar{k}}\hat{x} = x_{\mathcal{T}(m)\mathcal{S}(n)\mathcal{Q}\mathcal{S}^{-1}(n)\mathcal{T}^{-1}(m-\bar{k})}$ where both direct and inverse temporal transforms make use of the set of motion vectors $mv(0,k), k = 0 \ldots \bar{k}$. Instead, if a lower spatial resolution $\bar{l}$ version is needed the inversion process is incoherent with respect to the forward decomposition. For clarity, let us consider only spatial scalability, so that the spatially scaled signal is obtained by ${}^{\bar{l}}x = x_{\mathcal{T}(m)\mathcal{S}(n)\mathcal{S}^{-1}(n-\bar{l})\mathcal{T}^{-1}(m)}$. The problem is due to the fact that the inverse MCTF transform is obtained using motion vectors $mv(\bar{l},k)$ that are typically deduced from the estimated ones by $mv(\bar{l},k) = mv(0,k)_{d(\bar{l})}$. 1) This cannot guarantee perfect reconstruction because spatial and motion compensated temporal transforms cannot be inverted without most likely changing the result (due to the shift-variant nature of the spatial transform) [15], and 2) these motion vectors have not been estimated on data at the target resolution and therefore they cannot be considered optimal for use, when spatial subsampling has been made with non alias free wavelet filters. Moreover, the bit occupancy of full resolution motion vectors penalizes the performance of spatially scaled decoding especially at lower bit-rates.

The above issues can only be mitigated in such a critically sampled t+2D scheme but not structurally resolved. To reduce the MV bit load at full resolution scalable MV coding can be used [59][60], to reduce it at intermediate and lower resolutions approximated MV down-conversions $mv(\bar{l},k) = mv(0,k)_{d(\bar{l}),a}$ can be adopted (e.g. from half-pel at spatial resolution $\bar{l}$ again to half-pel at resolution $\bar{l}+1$ instead of the quarter-pel generated by the down-conversion). These two solutions have the disadvantage of introducing some imprecisions in the inverse temporal transform and their possible adoption should be balanced in a R-D sense. The problem of the non optimality of the down-conversion due to spatial aliasing, introduced by the decimation, can be reduced by using (possibly only for the needed resolutions) more selective wavelet filters [61] or locally adaptive spectral shaping by acting on the quantization parameters inside each spatial subband [62]. However, such approaches can determine coding performance loss at full resolution, because either the wavelet filters or coefficient quantization laws are moved away from ideal conditions for coding performance.

2) 2D+t: In order to solve the spatial scalability performance issues of t+2D schemes, a natural approach could be to consider a 2D+t scheme, where the spatial transform is applied before the temporal one as shown in Fig.9(b).

The resulting scheme is often called In-Band MCTF (IBMCTF) [63] because estimation of $mv(l,k)$ is made and applied independently on each spatial level in the spatial subband domain, leading to a structurally scalable motion representation and coding. With IBMCTF spatial and temporal scalability are more decoupled with respect to the t+2D case. Signal reconstruction after a combination of spatiotemporal and quality scalability is given by ${}^{\bar{l}}_{\bar{k}}\hat{x} = x_{\mathcal{S}(n)\mathcal{T}(m)\mathcal{Q}\mathcal{T}^{-1}(m-\bar{k})\mathcal{S}^{-1}(n-\bar{l})}$ where the (partial) inverse temporal transform makes use of $mv(l,k), k = 0 \ldots \bar{k}, l = 0 \ldots \bar{l}$, (a subset of) the same MVs used for the direct decomposition.

Despite its conceptual simplicity the 2D+t approach suffers as well from the shift-variant (periodically shift-invariant) nature of the DWT decomposition. In fact, motion estimation in the critically sampled detail subbands is not reliable because the coefficient patterns of moving parts usually change from one frame to consecutive or preceding ones. Solutions to this issue have been proposed [64][65]; they foresee a frame expansion formulation, at the expense of an increased complexity, to perform motion estimation and compensation in an overcomplete (shift-invariant) spatial domain, while transmitting only decimated samples.

Despite this solution, 2D+t schemes demonstrated lower coding efficiency compared to t+2D ones, especially at higher resolutions [66]. Motivations of this fact are probably due to the reduced efficiency of MCTF when applied in the high-pass spatial subband domain. In addition, spatial inversion of detail subbands which contain block-based MCTF coding artifacts visually smoothes blocking effects but probably causes an overall visual quality degradation.

Another possible explanation of a reduced coding efficiency resides in the fact that parent-child spatial relationships, typically exploited during embedded wavelet coding or context based arithmetic coding, are partially broken by independent temporal filtering on the spatial levels.

3) Adaptive architectures: Being characterized by different advantages and disadvantages, t+2D and 2D+t architectures seem inadequate to give a final answer to the composite issues arising in SVC applications. One possible approach comes from schemes which try to combine the positive aspects of each scheme and propose adaptive spatiotemporal decompositions optimized with respect to suitable criteria. In [67][68] authors suggest that content-adaptive 2D+t versus t+2D decomposition can improve coding performance. In particular, they show how various kind of disadvantages of fixed transform structures can be more or less preponderant depending on video data content, especially on the accuracy of the derived motion model. Therefore, local estimates of the motion accuracy allow to adapt the spatiotemporal data decomposition structure in order to preserve scalability while obtaining overall maximization of coding performance.

Recently, another adaptive solution, named aceSVC [69][70], has been proposed where the decomposition adaptation is driven by scalability requirements. This solution tries to achieve maximum bit-stream flexibility and adaptability to a selected spatiotemporal and quality working point configuration. This is obtained by arranging spatial and temporal decompositions following the principles of a generalised spatiotemporal scalability (GSTS) [71]. In aceSVC encoded bit-stream, thanks to the adapted decomposition structure, every scalability layer is efficiently used (without waste of bits) for the construction of higher ones according to the predefined spatial, temporal and quality working point configurations.

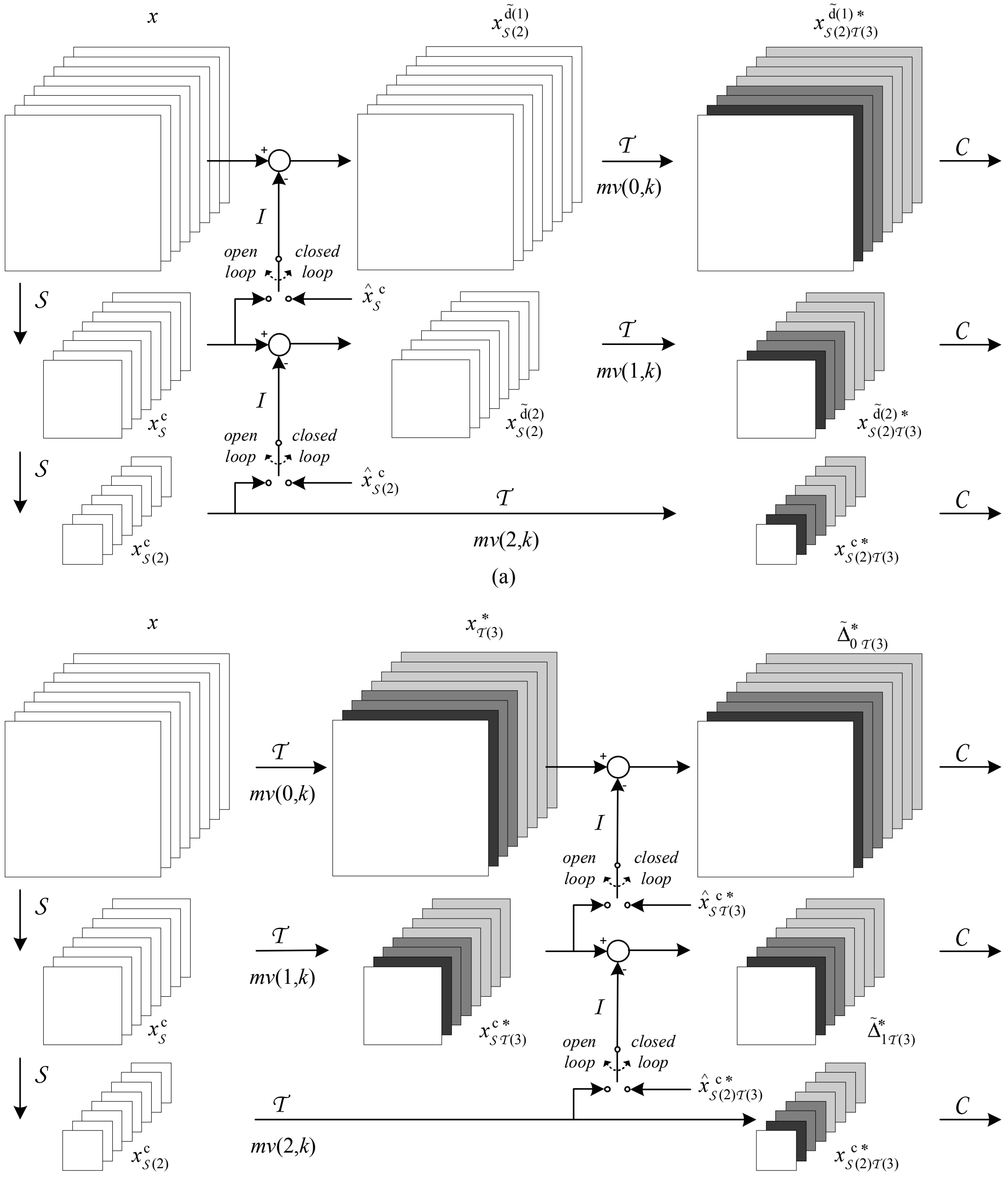
Another perspective to compensate for the t+2D vs 2D+t drawbacks can come from the pyramidal approaches, as discussed in the reminder of this section.

4) Multi-scale pyramids: From the above discussion it is clear that the spatial and temporal wavelet filtering cannot be decoupled because of the motion compensation taking place. As a consequence, it is not possible to encode different spatial resolution levels at once, with only one MCTF, thus both lower and higher resolution frames or subband sequences should be MCT-filtered. In this perspective, a possibility for obtaining good performance in terms of bit-rate and scalability is to use coding schemes which are usually referred as to 2D+t+2D or multi-scale pyramids [72]. In such schemes spatially scaled versions of video signals are obtained starting from the higher resolution, while the coding systems encompass Inter-Scale Prediction (ISP) mechanisms in order to exploit the multi-scale representation redundancy for coding efficiency purposes. These multi-resolution decompositions derive naturally from the first hierarchical representation technique introduced for images, namely the Laplacian pyramid [6] (see Sec. II-B). So, even if from an intuitive point of view 2D+t+2D schemes seems to be well motivated, they have the typical disadvantage of overcomplete transforms, namely those leading to a full size residual image, so that coding efficiency is harder to achieve. As in the case of the previously discussed t+2D and 2D+t schemes, two different multi-scale pyramid based schemes are conceivable, depending where the pyramidal residue generation is implemented, i.e. before or after the MCTF decomposition. These two schemes are shown in Fig.10. In both schemes the lowest spatial resolution temporal subbands $x^{c*}_{\mathcal{S}(n)\mathcal{T}(m)}$ are the same, while higher resolution residuals are generated differently.

The scheme of Fig.10(a) suffers from the problems of 2D+t schemes, because motion estimation is operated on spatial residuals, and to our knowledge did not lead to efficient SVC implementations.

In the scheme depicted in Fig.10(b) we first observe that the higher resolution residuals $\Delta^*_{l\,\mathcal{T}(m)}$, with $l = 0 \ldots n-1$, cannot be called $x^{*\,\tilde{d}(l+1)}_{\mathcal{T}(m)\mathcal{S}(n)}$ since the spatial transform is not exactly a Laplacian one because the temporal transform stages, which are placed between the spatial down-sampling and the prediction with interpolation stages, operate independently (on separate originals). This fact can cause prediction inefficiencies which may be due to motion model discrepancies among different resolutions or simply to the non linear and shift variant nature of the motion compensation even in the presence of similar motion models. It is worth noting that the effect of these inefficiencies affects the whole residue and then it worsens the problem of the increased number of coefficients to code. Strategies to reduce these possible inefficiencies are possible and we will shortly return on these issues in the following subsection. Note that the coding scheme shown in Fig.10(b) can be used to represent the main coding principles underlying the JVT-SVC system [2].

It is also important to mention that some recent works demonstrate that using the frame theory [7] it is possible to improve the coding performance associated to the Laplacian pyramid representation [73][74] and also to reduce its redundancy [75]. In the future these works may play a role in improving the spatial scalability performance of the coding schemes described in this paragraph.


x
$x^{\tilde{d}(1)}_{\mathcal{S}(2)}$
$x^{\tilde{d}(1)*}_{\mathcal{S}(2)\mathcal{T}(3)}$
+
-
$\mathcal{T}$
$mv(0,k)$
$\mathcal{C}$
I
open loop
closed loop
$\hat{x}^{c}_{\mathcal{S}}$
$\mathcal{S}$
+
-
$\mathcal{T}$
$mv(1,k)$
$\mathcal{C}$
$x^{c}_{\mathcal{S}}$
$x^{\tilde{d}(2)}_{\mathcal{S}(2)}$
$x^{\tilde{d}(2)*}_{\mathcal{S}(2)\mathcal{T}(3)}$
I
open loop
closed loop
$\mathcal{S}$
$\hat{x}^{c}_{\mathcal{S}(2)}$
$\mathcal{T}$
$\mathcal{C}$
$mv(2,k)$
$x^{c}_{\mathcal{S}(2)}$
$x^{c*}_{\mathcal{S}(2)\mathcal{T}(3)}$
x
$x^{*}_{\mathcal{T}(3)}$
$\tilde{\Delta}^{*}_{0\,\mathcal{T}(3)}$
$\mathcal{T}$
$mv(0,k)$
+
-
$\mathcal{C}$
I
open loop
closed loop
$\mathcal{S}$
$\hat{x}^{c*}_{\mathcal{S}\mathcal{T}(3)}$
$\mathcal{T}$
$mv(1,k)$
+
-
$\mathcal{C}$
$x^{c}_{\mathcal{S}}$
$x^{c*}_{\mathcal{S}\mathcal{T}(3)}$
I
$\tilde{\Delta}^{*}_{1\mathcal{T}(3)}$
open loop
closed loop
$\mathcal{S}$
$\hat{x}^{c*}_{\mathcal{S}(2)\mathcal{T}(3)}$
$\mathcal{T}$
$\mathcal{C}$
$mv(2,k)$
$x^{c}_{\mathcal{S}(2)}$
$x^{c*}_{\mathcal{S}(2)\mathcal{T}(3)}$

(a)

(b)

Fig. 10. Pyramidal WSVC schemes with the pyramidal decomposition put (a) before and (b) after the temporal MCTF.

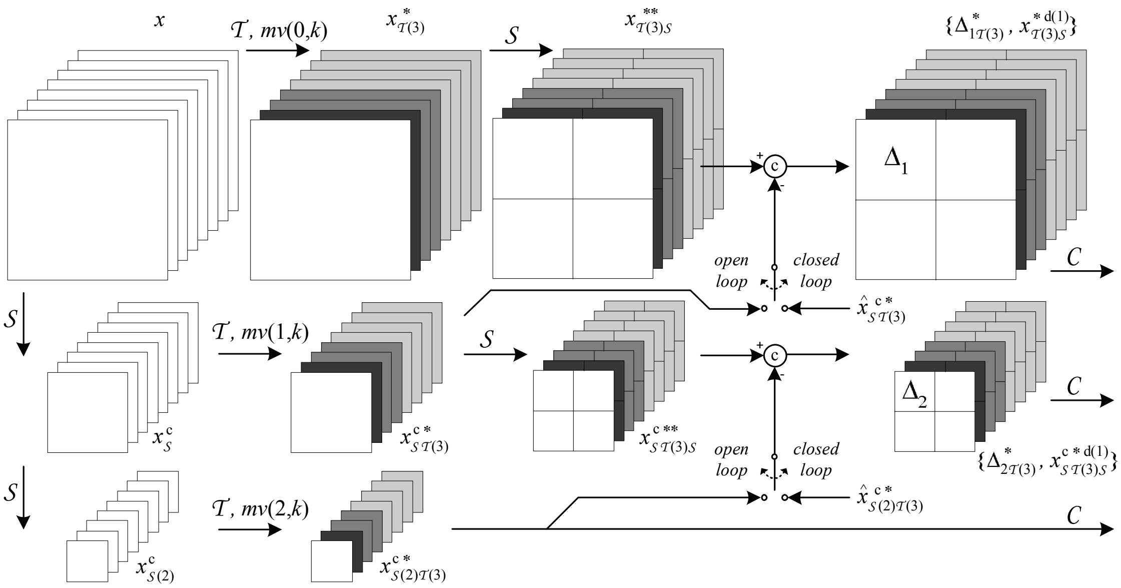

x
$\mathcal{T}$, $mv(0,k)$
$x^*_{\mathcal{T}(3)}$
$\mathcal{S}$
$x^{**}_{\mathcal{T}(3)\mathcal{S}}$
$\{\Delta^*_{1\mathcal{T}(3)}, x^{*\,\mathrm{d}(1)}_{\mathcal{T}(3)\mathcal{S}}\}$
Δ_1
c
+
-
C
open loop
closed loop
$\hat{x}^{\mathrm{c}*}_{\mathcal{S}\mathcal{T}(3)}$
$\mathcal{S}$
$\mathcal{T}$, $mv(1,k)$
$x^{\mathrm{c}}_{\mathcal{S}}$
$x^{\mathrm{c}*}_{\mathcal{S}\mathcal{T}(3)}$
$\mathcal{S}$
$x^{\mathrm{c}**}_{\mathcal{S}\mathcal{T}(3)\mathcal{S}}$
c
+
-
Δ_2
C
open loop
closed loop
$\{\Delta^*_{2\mathcal{T}(3)}, x^{\mathrm{c}*\,\mathrm{d}(1)}_{\mathcal{S}\mathcal{T}(3)\mathcal{S}}\}$
$\hat{x}^{\mathrm{c}*}_{\mathcal{S}(2)\mathcal{T}(3)}$
$\mathcal{S}$
$\mathcal{T}$, $mv(2,k)$
$x^{\mathrm{c}}_{\mathcal{S}(2)}$
$x^{\mathrm{c}*}_{\mathcal{S}(2)\mathcal{T}(3)}$
C

Fig. 11. STP-tool scheme in a generated signals/subbands perspective.

C. *The STP-tool scheme*

We here present a WSVC architecture that has been developed [76] and also adopted as a possible configuration of the MPEG VidWav reference software [36] (see Sec.IV-B). This scheme is based on a multi-scale pyramid and differs from the one of Fig.10(b) in the ISP mechanism. The solution, depicted in Fig.11, is called STP-tool as it can be interpreted as a prediction tool (ISP) for spatiotemporal subbands. After the representation of the video at different scales, these undergo a normal MCTF. The resulting signals are then passed to a further spatial transform stage (typically one level of a wavelet transform) before the ISP. This way, it is possible to implement a prediction between two signals which are likely to bear similar patterns in the spatiotemporal domain without the need to perform any interpolation. In particular, instead of the full resolution residuals $\Delta^*_{l\mathcal{T}(m)}$ of Fig.10(b) the spatiotemporal subbands and residues $\{\Delta^*_{l+1\,\mathcal{T}(m)}, x^{c*d(l+1)}_{\mathcal{S}(l)\mathcal{T}(m)\mathcal{S}}\}$ are produced for resolutions $l = 0 \dots n-1$ where $\Delta^*_{l+1\,\mathcal{T}(m)}$ are produced by the coarse resolution adder (shown in Fig. 11) according to: a) $\Delta^*_{l+1\,\mathcal{T}(m)} = x^{c*c}_{\mathcal{S}(l)\mathcal{T}(m)\mathcal{S}} - x^{c*}_{\mathcal{S}(l+1)\mathcal{T}(m)}$ in a spatial open loop architecture or b) $\hat{\Delta}^*_{l+1\,\mathcal{T}(m)} = x^{c*c}_{\mathcal{S}(l)\mathcal{T}(m)\mathcal{S}} - \hat{x}^{c*}_{\mathcal{S}(l+1)\mathcal{T}(m)}$ in a spatial closed loop architecture.

In an open-loop scheme and without the temporal transform, the Δ residual would be zero (if the same pre- and post-spatial transforms were used). In this case, the scheme would be reduced to a critically sampled one. With the presence of the temporal transform, things are more complicated due to the datadependent structure of the motion model and to the non linear and time variant behavior of the MCTF. In general, it is not possible to move the spatial transform stages forward and backward with respect to the MCTF without changing the result. If the spatial transform used to produce spatial subbands are the same, the main factor that produces differences between the $x^{c*c}_{\mathcal{S}(l)\mathcal{T}(m)\mathcal{S}}$ and the $x^{c*}_{\mathcal{S}(l+1)\mathcal{T}(m)}$ signals can be identified in the incoherences between the motion models at different spatial resolutions. On one hand, independent motion estimation at various spatial resolution could lead to intra-layer optimized motion compensation. However, a motion model that captures the physical motion in the scene is certainly effective for the motion compensation task and reasonably should be self similar across scales. Therefore, despite a certain need to differentiate the motion model at different scales, a desirable objective is to have a hierarchical motion estimation where high-resolution motion fields are obtained by refinement of corresponding lower resolution ones (or viceversa). Indeed this guarantees a certain degree of inter-scale motion field coherence which makes the STP-tool prediction more effective. A hierarchical motion model is a desired feature here and in addition it helps to minimize the motion vector coding rate.

The STP-tool architecture presents another relevant and peculiar aspect. Although the number of subband coefficients to code is not reduced with respect to a pyramidal approach, different (prediction) error sources are separated into different critically sampled signals. This does not happen in a pyramidal schemes where only full resolution residuals $\Delta^*_{l\,\mathcal{T}(m)}$ are produced (see Fig.10(b)).

To better understand these facts let us analyze separately each residual or detail component in the set $\{\Delta^*_{l+1\,\mathcal{T}(m)}, x^{c*d(l+1)}_{\mathcal{S}(l)\mathcal{T}(m)\mathcal{S}}\}$. In an open loop architecture,

- the residue signals $\Delta^*_{l+1\,\mathcal{T}(m)}$ account for the above mentioned *inter-scale* motion model discrepancies; while
- the temporal high-pass, spatial high-pass subbands $x^{c\,d(k)\,d(l+1)}_{\mathcal{S}(l)\mathcal{T}(m)\mathcal{S}}$ account for the *intra-scale* motion model failure to completely absorb and represent the real motion (occluded and uncovered areas and motion feature not handled by the motion model); finally
- the temporal low-pass, spatial high-pass subbands $x^{c\,c\,d(l+1)}_{\mathcal{S}(l)\mathcal{T}(m)\mathcal{S}}$ contain the detail information to increase the resolution of low pass temporal frames.

Thus, maximizing the inter-scale motion coherence implies minimizing the energy (and the coding cost) of $\Delta^*_{l+1\,\mathcal{T}(m)}$. Note that even if a perfect motion down-scaling is adopted, localized residual components would be due to the shift variant properties of the motion compensation especially where uncovered areas and occlusion handling must be adopted. In other words, the coefficient increase, determined by the presence of the residues $\Delta^*_{l+1\,\mathcal{T}(m)}$, with respect to the critically sampled case is the price to pay for allowing a suitable motion estimation and compensation across different scales.

In a closed-loop ISP architecture we also have the great advantage that the quantization errors related to resolution l are confined to signals $\varepsilon^{c*}_{l+1\,\mathcal{T}(m)} = x^{c*}_{\mathcal{S}(l+1)\mathcal{T}(m)} - \hat{x}^{c*}_{\mathcal{S}(l+1)\mathcal{T}(m)}$ of resolution $l+1$ that contribute in a more or less predominant way to the residues to code, that is $\hat{\Delta}^{*}_{l+1\,\mathcal{T}(m)} = \Delta^{*}_{l+1\,\mathcal{T}(m)} + \varepsilon^{c*}_{l+1\,\mathcal{T}(m)}$. This is of great advantage for the reduced number of coefficients on which the closed-loop error is spread upon. Besides, the $\varepsilon^{c*}_{l+1\,\mathcal{T}(m)}$ signals are less structured, and thus more difficult to code, with respect to $x^{c*\,d(l+1)}_{\mathcal{S}(l)\mathcal{T}(m)\mathcal{S}}$ and $\Delta^{*}_{l+1\,\mathcal{T}(m)}$. Therefore the STP-tool scheme allows a useful separation of the above signals, with the advantage that suitable coding strategies can be designed for each source.

As for t+2D and 2D+t schemes, the STP-tool solution may suffer from the presence of spatial aliasing on reduced spatial resolutions. If not controlled this may also create some inter-scale discrepancies between the motion models. Again, a partial solution can be found with the use of more selective decimation wavelet filters, at least for decomposition levels corresponding to decodable spatial resolutions.

The STP-tool solution has been tested on different platforms with satisfactory performance. In particular, it was inserted as a possible configuration in the MPEG reference software platform for WSVC called VidWav. In the next section, we present an overview of such VidWav reference software configurations and components.

IV. WSVC REFERENCE PLATFORM IN MPEG

In the past years, the ISO/MPEG standardization group initiated an exploration activity to develop scalable video coding (SVC) tools and systems mainly oriented to obtain spatiotemporal scalability. One of the key issues of the conducted work was not to sacrifice coding performance with respect to the state-of-the-art reference video coding standard MPEG4-AVC.

In 2004, at Palma meeting, the ISO/MPEG group set up a formal evaluation of scalable video coding technologies. Several solutions [1][66][77][3], either based on wavelet spatiotemporal decomposition or MPEG4-AVC technologies, had been proposed and compared. The MPEG visual testing indicated that performance of an MPEG4-AVC pyramid [25] appeared the most competitive and it was selected as a starting point for the new standard. At the next meeting, Hong Kong, MPEG, jointly with the IEC/ITU-T video group, started a joint standardization initiative accordingly. A scalable reference model and software platform named JSVM derived from MPEG4-AVC technologies was adopted [2]. During the above mentioned competitive phases of the SVC process, among other wavelet based solutions, the platform submitted by Microsoft Research Asia (MSRA) was selected as the reference to continue experiments on this technology. The MPEG WSVC reference model and software [78], hereinafter will be indicated with

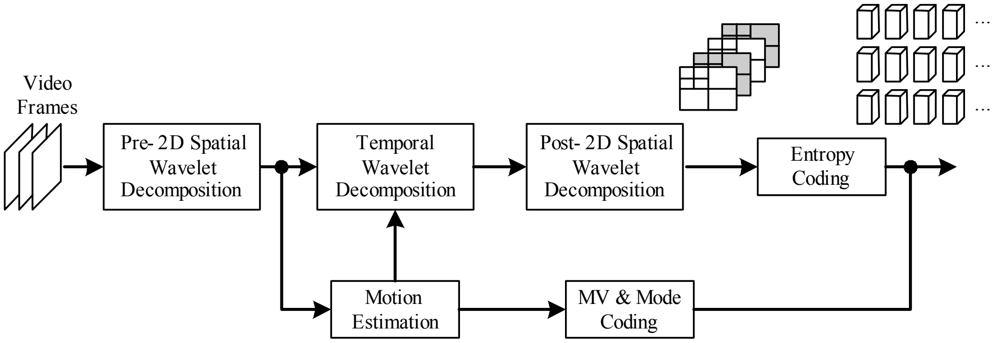

Video
Frames
Pre- 2D Spatial Wavelet Decomposition
Temporal Wavelet Decomposition
Post- 2D Spatial Wavelet Decomposition
Entropy Coding
Motion Estimation
MV & Mode Coding
...
...
...

Fig. 12. VidWav general framework.

the acronym VidWav (Video Wavelet). The VidWav Reference Model and Software (RM/RS) has evolved seeing various components integrated, provided they would give sufficient performance improvement after appropriate core experiments.

In the paragraphs of the present section we mainly recall the characteristics of the VidWav RM. Document [78] should be taken as a reference to go into further details. In the next section, we will discuss the main differences with respect to the JVT-SVC standard. We will also provide some comparison results between the two approaches.

A. VidWav: general framework and main modules

The general framework supported by the VidWav RM is shown in Fig.12. With Pre- and Post- spatial decomposition different WSVC configurations can be implemented: t+2D, 2D+t and STP-tool. The main modules of the manageable architectures are briefly considered in the following.

Motion estimation and coding: A macroblock based motion model was adopted with MPEG4-AVC like partition patterns. To support spatial scalability, macroblock size is scaled according to the decimation ratio. Interpolation accuracy is also scaled. Therefore, full pel estimation on 64x64 macroblocks at 4CIF resolution can be used at quarter-pel precision on a 16x16 macroblock basis for QCIF reconstructions. Hence, motion is estimated at full resolution and kept coherent across scales by simply scaling the motion field partition and the motion vector values. This poses the limitations already discussed in Sec.III-B for the t+2D architecture.

For each block there is the possibility to select a forward, backward or bidirectional motion model. This requires a mode selection information. Mode selection also determines the block aspect ratio and informs about the possibility of determining current MV values from already estimated ones. Motion vector modes and values are estimated using rate-constrained lagrangian optimizations and coded with variable length and predictive coding, similarly to what happens in MPEG4-AVC. For motion compensation filtering a connected /disconnected flag is included to the motion information on units of 4x4 pixels. For 4:2:0 YUV sequences, motion vectors for chroma components are half those of the luminance.

Temporal Transform: Temporal transform modules implement a framewise motion compensated (or motion aligned) wavelet transform on a lifting structure. The block based temporal model is adopted for temporal filter alignment. Motion vectors are directly used for the motion aligned prediction (MAP) lifting step while the motion field should be inverted for motion aligned update (MAU). To this end, Barbell update lifting [31] is adopted which directly uses original motion vectors and bilinear interpolation to find values of update signals from non integer pixel values. In order to further reduce blocking artifacts, overlapped block motion compensation (OBMC) can be used during the MAP. This is accomplished by multiplying the reference signals by a pyramidal window with support larger than 4×4 before performing the temporal filtering. The update signal is also clipped according to a suitable threshold in order to limit the visibility of possible ghosting artifacts. Moreover, the implemented temporal transform is generally locally and adaptively a Haar or a 5/3-taps depending whether motion block modes are unidirectional (forward or backward) or bidirectional.

Spatial transform: A simple syntax allows to configure Pre- and Post- spatial transforms on each frame or previously transformed subband. This allows to implement both t+2D and 2D+t schemes. Different spatial decompositions are possible for different temporal subbands for possible adaptation to the signal properties of the temporal subbands.

Entropy Coding: After the spatiotemporal modules the coefficients are coded with a 3D (spatiotemporal) extension of the EBCOT algorithm, called 3D EBCOT. Each spatiotemporal subband is divided into 3D blocks which are coded independently. For each block, fractional bit-plane coding and spatiotemporal context based arithmetic coding are used.

B. VidWav STP-tool configuration

The VidWav reference software can be configured to follow the STP-tool architecture of Fig.11. A corresponding functional block diagram of the resulting system is shown in Fig.13. The coding process starts from the lower resolution which acts as a base layer for ISP. In particular, a closed loop ISP is implemented as coded temporal subbands are saved and used for STP-tool prediction for the higher resolution. STP-tool prediction can be applied on all spatiotemporal subbands or only to a subset of them. Decoded temporal subbands at a suitable bit-rate and temporal resolution are saved and used for ISP. At present, no particular optimization exists in selecting the prediction point. Obviously, it should be not too far from the maximum point for a certain temporal resolution at lower spatial resolution in order to avoid bit-stream spoilage (lower resolution bit-stream portions are useless for higher resolution prediction) and prediction inefficiency (too large prediction residual). Spatial transforms on the levels interested by STP-tool ISP can be implemented either by using the three lifting steps (3LS) [61] or the 9x7 filters. In order to better exploit inter scale redundancy these filters should be of the same type of those used for the generation of the reference videos for each of the considered spatial resolutions. In this implementation, some important limitations remain. These are mainly related to software integration problems due to the data structures used by the original MSRA software. Therefore, coherent and refinable bottom up motion estimation and coding are not present in the current VidWav software. When STP-tool is used, three independent motion fields are estimated independently. This introduces some undesired redundancy in the bit-stream which has a major impact for the lowest quality point at higher spatial resolutions. In addition, it does not provide the motion field coherence across scales which was seen as a desirable feature for good STP-tool prediction.

C. VidWav additional modules

Other modules have been incorporated to the baseline VidWav reference platform after careful review of performance improvement by appropriate testing conditions. Such tools can be turned on or off through configuration parameters. More detailed description of these tools can be found in [78].

Base layer: The minimum spatial, temporal and quality resolution working point define a Base Layer on which additional Enhancement Layers are built to provide scalability up to a desired maximum spatial, temporal and quality resolution. This tool enables the use of a Base Layer, in 2D+t subband scalable video coding, which has been previously compressed with a different technique. In particular, the MPEG4-AVC stream can be used as Base Layer by implementing a hierarchical B-frames decomposition (HBFD) structure. This allows each B picture of the Base Layer to temporally match corresponding high-pass subbands in the Enhancement Layers. It therefore enables the frames to share similar MCTF structure decomposition and this correlation can be efficiently exploited. Motion information derived from the Base Layer codec can also be used as additional candidate predictors for estimating motion information of the blocks in the corresponding high-pass temporal subbands of the Enhancement Layers. Clearly, the Base Layer tool can significantly improve the coding performances, at least of the lowest working point

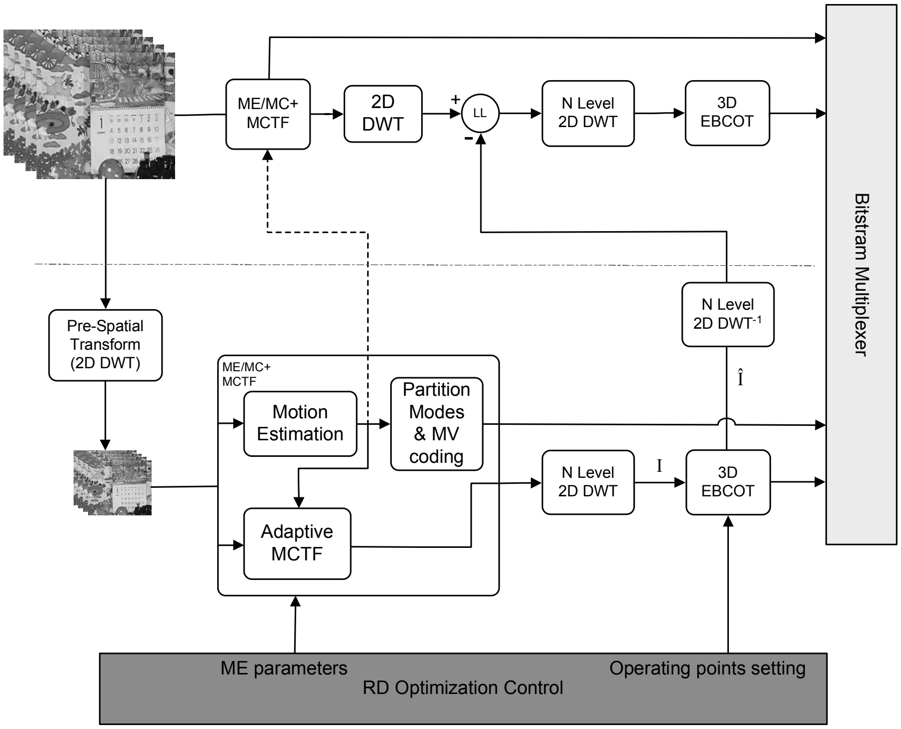


Fig. 13. VidWaw framework in STP-tool configuration. The lowest spatial resolution layer is equivalent to a t+2D scheme. Higher spatial layers implement STP-tool inter scale prediction.

which is now encoded with a non scalable method. A second but not less relevant provided benefit, is the backward compatibility with preexisting non scalable MPEG4-AVC decoder.

In-band 2D+t: The 2D+t scheme is activated when a non identity pre-spatial transform in Fig.12 is set. In such a configuration the MCTF is applied to each spatial subband generated by the pre-spatial transform. In Section III-B.2 it has been already evidenced that in-band scheme suffers a loss in coding performance at full resolution when the MCTF is performed independently on each subband. Conversely, if high resolution signals are involved in the prediction for low resolution, drifting occurs when decoding low resolution video, because in that case high resolution signals are not available at the decoder. The VidWav RM adopts a technique [79] based on the use on additional macroblock inter coding modes in the MCTF of the spatial low-pass band. With respect to the pure 2D+t scheme, this method provides a better trade-off between the lower resolution drifting error and the high resolution coding efficiency and also improves the global coding performances.

Wavelet ringing reduction: In Wavelet based video compression, ringing caused by the loss of details in the high pass spatial subbands is a major concern for the resulting visual quality [80]. Ringing is generated during inverse DWT because quantization-noise of lossily-coded wavelet coefficients is spread by the reconstruction filters of the IDWT. The oscillating pattern of these filter impulse responses at different reconstruction scales becomes objectionable especially in flat luma and chroma regions. The ringing magnitude is related to the quantization step-size. In order to reduce these ringing patterns, a 2D bilateral filter applied to each component of the output video is used.

Intra mode prediction: The intra-prediction mode exploits the spatial correlation within the frame in order to predict intra areas before entropy coding. Similarly to the MPEG4-AVC approach, a macroblock can be predicted form the values of its neighbors, i.e. by using spatial predictors.

V. WSVC AND SVC STANDARDIZATION REFERENCE PLATFORMS: A COMPARISON

A. Architecture and tools: similarities and differences

In this paragraph some differences and similarities between JSVM and VidWav RM (see Fig. 13) frameworks will be highlighted and discussed.

1) Single layer coding tools: We will start to analyze the differences between the VidWav RM [36] and the JSVM [2] when they are used to encode a single operating point, i.e. no scalability features being considered. Broadly speaking this can be seen as a comparison between a t+2D WSVC (which will remain fully scalable at least for temporal and quality resolution) and something similar to MPEG4-AVC (possibly supporting temporal scalability thanks to HBFD).

VidWav uses a block based motion model. Block mode types are similar to those of JSVM with the exception of the Intra-mode which is not supported by VidWav. This is due to the fact that the Intra mode determines blocks rewriting from the original and prediction signal (high-pass temporal subbands) producing a mixed (original+residue) source which is detrimental for subsequent whole frame spatial transforms. This is also one of the main differences for spatiotemporal redundancy reduction between the two approaches. JSVM (as well as all MPEG4-AVC and most of the best performing hybrid video coding methods) operates in a local manner: single frames are divided into macro-blocks which are treated separately in all the coding phases (spatial transform included). Instead, the VidWav (as well as WSVC approaches) operates with a global approach since the spatiotemporal transform is directly applied to a group of frames. The fact that in the VidWav framework a block based motion model has been adopted is only due to efficiency and the lack of effective implementations of alternative solutions.

However, this cannot be considered an optimal approach. In fact, the presence of blockiness in the high-pass temporal subbands due to block based predictions is inconsistent and causes inefficiencies with respect to the subsequent spatial transform. The energy spreading around block boundaries, generates artificial high frequency components and disturbs both zerotree and even block based entropy coding mechanisms which act on spatiotemporal subbands. On the other hand blockiness is not a main issue when block-based transforms, such as the DCT, are instead applied, as in the case of JSVM. OBMC techniques can reduce the blockiness in the high-pass temporal subbands but do not remove it.
Contrarily to JSVM, the single layer (t+2D) VidWav framework only supports open loop encoding/decoding. Thanks to the closed loop encoding, JSVM can exploit an in-loop de-blocking filter, which makes a similar job with respect to the post-processing de-blocking(JSVM)/de-ringing(VidWav) filters with the additional benefit of improving the quality of the reconstructed signals used in closed-loop predictions.
Closed-loop configurations within WSVC systems would be possible by adopting an HBFD like decomposition with the aim of incorporating in-loop de-ringing strategies as well.

2) Scalable coding tools: We now consider spatial scalability activated in both solutions, looking in particular to JSVM and VidWav in STP-tool configuration. The latter can be considered as to the WSVC implementation closest to the JSVM one. The main difference between the two systems in terms of a spatial scalability mechanism is again related to the block-based vs frame-based dichotomy. In JSVM, a decision on each single macro-block determines whether the macroblock is encoded by using the information of its homologous at a lower spatial resolution or not. In the latter case, one of the compatible modality available for a single layer encoding will be used. The decision, among all possible modes is globally regulated by R-D optimization. On the contrary, ISP inside STP-tool architecture is forced to be made at a frame level (with possible decisions on which frames to involve).
Similarly to JSVM, STP-tool can use both closed and open loop interlayer encoding while this is not true for t+2D or 2D+t configurations which can only operate in open loop mode.

B. *Objective and visual result comparisons*

The comparison between DCT based SVC and WSVC systems is far from being an obvious task. Objective and subjective comparisons are critical even when the most ”similar” architectures are considered (e.g. JSVM and STP-tool), especially when comparisons are performed using decoded video sequences at reduced spatial resolutions. In this case, since in general different codecs use different down-sampling filters, the coding-decoding processes will have different reconstruction reference sequences, i.e. different down-sampled versions of the same original. JSVM adopt the highly selective half-band MPEG down-sampling filter, while WSVC systems are more or less forced to use wavelet filter banks which in general offer less selective half-band response and smoother transition bands. In particular, in t+2D architectures wavelet down-sampling is a requisite while in STP-tool ones (see Sec.III-C) one could adopt a more free choice, at the expense of a less efficient ISP. Visually the reference sequences generated by wavelet filters are in general more detailed but sometimes, depending on data or visualization conditions, some spatial aliasing effects could be seen. Therefore visual benefits in using wavelet filters is controversial while schemes adopting more selective filters can take advantage, in terms of coding efficiency, of the lower data entropy. Unfortunately the MPEG down-sampling filter is not a good low-pass wavelet filter because the corresponding high-pass filter have an undesired behavior in the stop-band [81]. Attempts to design more selective wavelet filters have been made [61] [81] with expected benefits in terms of visual and coding performance for WSVC schemes.

Anyway, depending on the particular filter used for spatial down-sampling, reduced spatial resolution decoded sequences will differ even at full quality, impairing fair objective comparisons. For the above reasons PSNR curves at intermediate spatiotemporal resolution have been mainly used to evaluate the coding performance of a single family of architectures while for a direct comparison between two different systems more time-demanding and costly visual comparisons have been preferred. In the following we summarize the latest objective and subjective performance comparison issued by the work on SVC systems as conducted in April 2006 by ISO/MPEG. Finally, we will also return on the possibility of increasing objective comparisons fairness by proposing a common reference solution.

1) Objective comparison results: Experiments have been conducted under testing conditions described in document [82]. Three classes of experiments have been completed with the intention of evaluating coding performance of different systems and tool settings. The first class of experiments, named *combined scalability*, aimed to test the combined use of spatial, temporal and quality scalabilities according to shared scalability requirements [4]. The comparison has been performed using JSVM 4.0 and the VidWav RS 2.0 in STP-tool configuration. Fig.14 shows an example of the R-D curves where Y PSNR are only shown. The example considers different sequences, quality points and spatiotemporal resolutions, i.e. 4CIF-60Hz, CIF-7.5Hz, QCIF-15Hz (original resolution is 4CIF-60Hz in all cases). The complete set of results can be found in [3, Annex1].

As mentioned in Section IV-B, the VidWav STP-tool configuration implements an inefficient motion estimation and coding that partially justifies the performance differences visible in the above mentioned PSNR plots. The other two groups of experiments, *spatial scalability* and *SNR scalability*, aimed to test the use and the R-D performance of single scalability tools when other forms of scalability are not

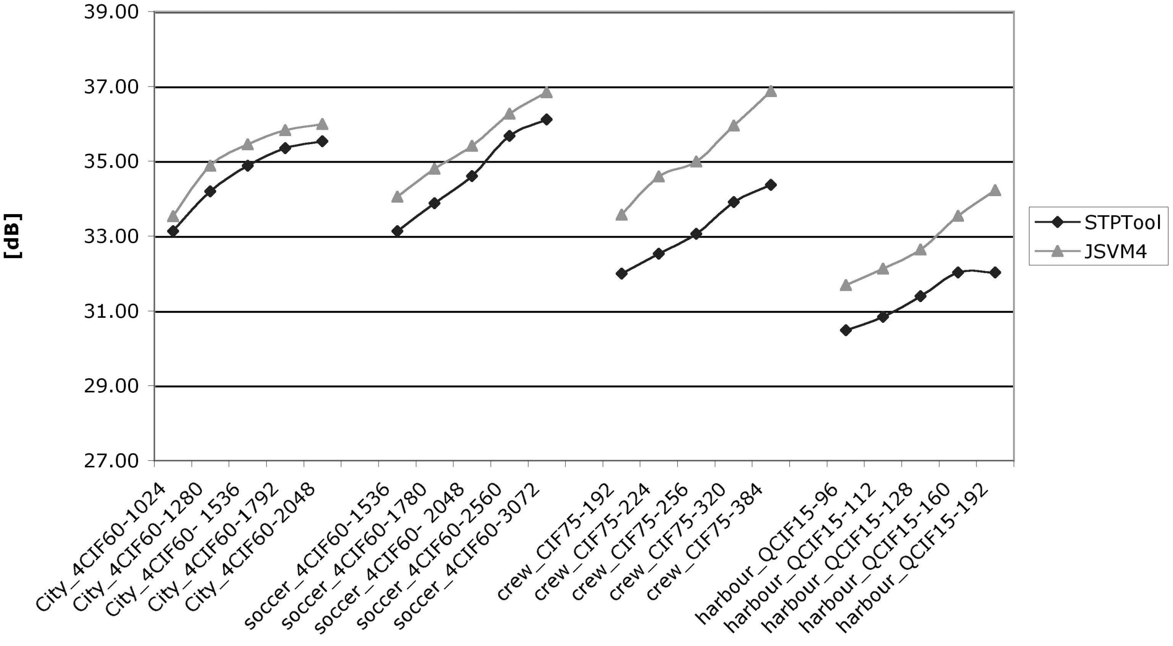


Fig. 14. Combined Scalability example results on 4 test sequences: City, Soccer, Crew and Harbour, at different spatiotemporal resolutions and bitrates using encapsulated bit-stream decoding as described in [82]

requested. VidWav in STP-tool and in t+2D configuration modes were tested. For the interested reader complete results are reported in [3, Annex2 and Annex3].

Where PSNR comparisons did not suffer from the different original reference signal problems, mixed performance were observed with a predominance of higher values for sequences obtained using JSVM.

2) Subjective comparison results: Visual tests conducted by ISO/MPEG included 12 expert viewers and have been performed according to guidelines specified in [83]. On the basis of these comparisons between VidWav RS 2.0 and JSVM 4.0 appear on average superior for JSVM 4.0, with marginal gains in SNR conditions, and superior gains in combined scalability settings. A full description of such test can be found in [3, Annex 4]. It should be said that, from a qualitative point of view, a skilled viewer could infer from which system the decoded image is generated. In fact, for motivations ascribable to spatial filtering aspects, VidWav decoded frames are in general more detailed with respect to JSVM ones, but bits not spent in background details can be used by JSVM to better patch motion model failures and then to reduce the impact of more objectionable motion artifacts for the used test material.

3) Objective measures with a common reference: A reasonable way to approach a fair objective comparison, between two systems that adopt different reference video sequences V_1 and V_2, generated

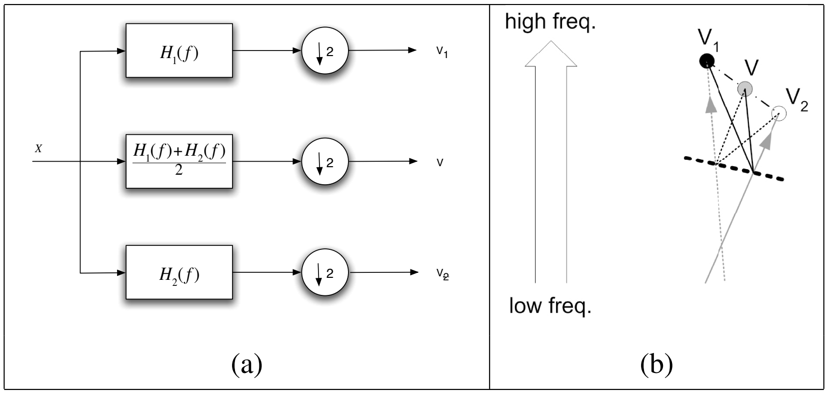


Fig. 15. Common reference generation (a) and representation (b).

from a same original sequence X, is to create a common weighted sequence $V = \alpha_1 V_1 + \alpha_2 V_2$. In particular, by selecting $\alpha_1 = \alpha_2 = 1/2$ it can be easily verified that PSNR$(V, V1)$ = PSNR$(V, V2)$. This means that V_1 and V_2 are both equally disadvantaged by the creation of the mean reference V. Fig.15(a) shows the generation of the above three signals using two different low pass-filters $H_1(f)$ and $H_2(f)$ which are assumed here to be a low-pass wavelet filter and the MPEG down-sampling filter respectively. As already discussed the output V_2 will have a more confined low-frequency content with respect to V_1. Fig.15(b) is a *projection* of the decoding process which shows vertically the video frame frequency content and horizontally other kind of differences (accounting for filter response discrepancies and spatial aliasing). Such differences can be considered small since V_1 and V_2 are generated starting from the same reference X while targeting the same down-sampling objective. Thus V can be considered halfway on the above projection being a linear average. Now, as transform based coding systems tend to reconstruct first the lower spectral components of the signal, it is plausible that, at a certain bit rate (represented in Fig.15(b) by the dashed bold line), the VidWav reconstructed signal lies nearer to the JSVM reference than to its own reference. The converse is not possible and this gives to V the capability to compensate for this disparity. Therefore, signal V can reasonably be used as a common reference for PSNR comparisons of decoded sequences generated by different SVC systems. These conjectures find a clear confirmation in experimental measures as it can be seen e.g. in Fig.16.

VI. Further investigations on WSVC schemes

In order to better identify the aspects or tools that could significantly improve WSVC performances and to drive future investigations and improvements more precisely, a series of basic experiments has been conducted comparing JSVM and its most similar WSVC codec, namely STP-tool.

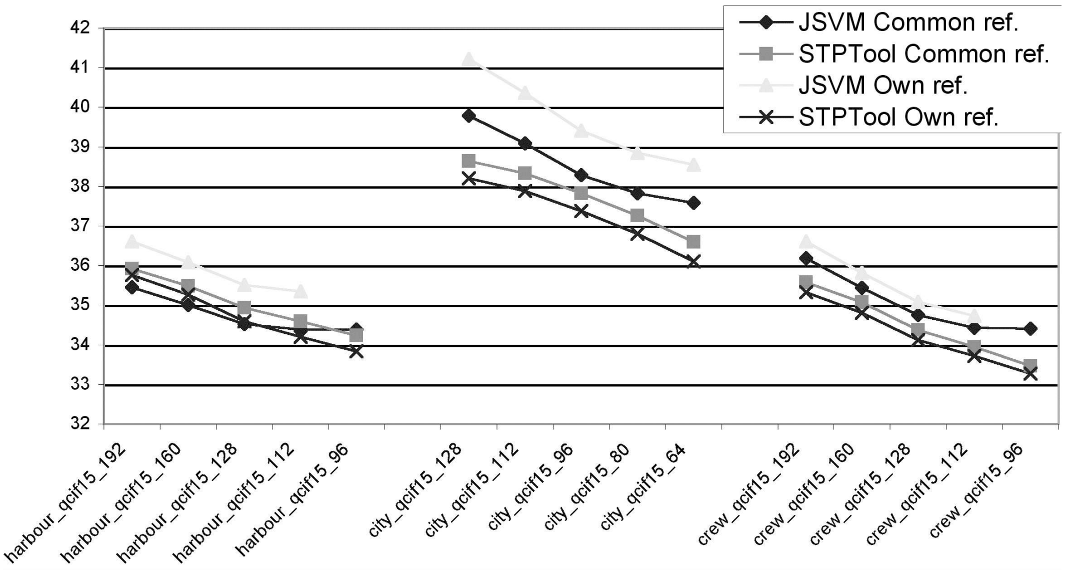


Fig. 16. PSNR comparison at QCIF resolution with own and common reference usage

In a first experiment a set of generic video sequences and still images were selected with the intent to compare the two codecs only in terms of intra coding performance. In other words, we tested the ability of independently compressing a still image that can be either a still image or frame randomly selected from a video sequence. Each test image was coded using both codecs in a single layer configuration repeating the procedure for a range of different rate values. These values have been selected in a way to have a reconstruction quality ranging from 30 to 40 dB. On the basis of the quality of decoded images the test set would be divided into two groups. The first group included all the images for which JSVM gives the best average coding performance while the second includes those where PSNR is in favor of the STP-tool codec. Not surprisingly, the first group contains only frame of video sequences such as those used for the MPEG comparisons and sequences acquired with low resolution cameras. Conversely, the second group was characterized by the presence of all the tested still images and high definition (HD) video sequences acquired with devices featuring the latest digital technologies. Thus, most performing DCT-based technologies appear to outperform wavelet-based ones for relatively smooth signals and vice versa. This indicates that eligible applications for WSVC are those that produce or use high-definition/high-resolution content. The fact that wavelet-based image (intra) coding can offer comparable if not better performance when applied to HD content is compatible with the results reported by other independent studies, e.g., [84][85]. In [86] the STP-tool has been used to test the performances

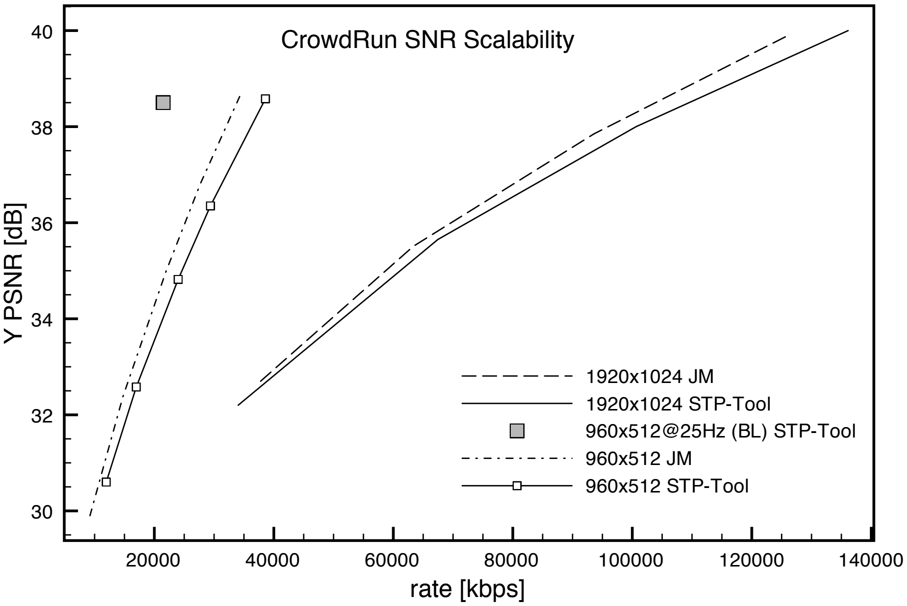


Fig. 17. SNR Scalability for the HD video sequence CrowdRun

of WSVC approaches for in home distribution of HD video sequence. As it can be noticed from Fig. 17 the SNR performance of the proposed codec are comparable with those offered by the state of the art of single point encoding techniques, e.g. MPEG-4 AVC.

Another set of tests has been conducted with the intent of analyzing the coding efficiency of the multi-resolution to compress high pass MC temporal subbands (or hierarchical B-frames residues). In this case coding efficiency was evaluated by comparing the performance only on the residual image by using the same predictor for both encoders and by changing the filters of the spatial wavelet transform. The main consideration arising from the obtained results is that commonly used filters, such as the 9x7, 5x3 and 2x2, are not as effective on high pass temporal subbands (or B residues) as much as they are on low pass temporal subbands (or frames). This is mainly because of the nature of the signal being considered which contains the residue information produced by a local adaptation procedure while the wavelet filter operates as a global transformation of the considered temporal subband. This means, for example, that if we want to encode a frame containing reconstruction errors only for one macro-block after the spatial transformation these errors will be spread across all the spatial subbands and the relative error support will also be extended because of filter lengths. A more detailed analysis of the energy of spatial subbands revealed that, contrarily to what happens for low pass temporal frames, most of the energy is located in the highest spatial resolution subbands. This is in accordance with the multi-resolution interpretation for the used spatial transform but because of their size these subbands cannot be efficiently encoded. Additionally, this prevents inter-band prediction mechanisms to correctly work.

VII. Perspectives

Although considerable efforts have been made to improve WSVC solutions there are some reasons to say that we could be at the beginning of a prolific research field with many possible applications. Present coding performances are not too far from the state-of-the-art JSVM which implements the technology defined in the forthcoming SVC standard. Hybrid video coding (especially based on block-based spatial DCT and block-based temporal motion estimation and compensation) has experimented a strong increase in performance as a consequence of the introduction of many optimized tools and configuration parameters. While some of these solutions have not yet been tested on a common reference architecture such as VidWav, other were forcely discarded since not adequate to the whole frame nature of the spatial wavelet transform. For example advanced block-based predictions with multiple block-mode selection, advanced block-based motion estimation, compensation and coding can be fully adopted with block-based transforms while they are not natural with full frame transforms. An example for which an attempt has been made is represented by the "intra-prediction" mode. This tool works remarkably well for JSVM but if simply imported on a VidWav architecture, it does not give the expected results. Wavelet based video coding systems would require alternative prediction modes and motion models. Existing solutions appear still preliminary and, in order to maximize their benefit, they need to be effectively integrated in the complete system. When compared to JSVM, existing wavelet codecs lack several performance optimization tools. In particular R-D optimized solutions are hardly implementing those tools which are not fully compatible with the image based approach used in WSVC systems. Therefore, two research paths seem to be important for the improvement and development of WSVC solutions: a) activities directed on finding new tools and solutions, b) activities directed on shaping and optimizing the use of existing tools.

New application potentials

There are a number of functionalities and applications which can be effectively targeted by WSVC technology and seem to prospect comparable if not more natural solutions with respect to their hybrid

DCT+MEC SVC counterparts. A brief list is reported here, while a more complete overview can be found in [3].

- HD material storage and distribution. Thanks to the built-in scalability brought by wavelets, it is possible to encode a given content with non-predefined scalability range and at a quality up to near lossless. At the same time, a very low definition decoding is possible allowing, a quick preview of the content, for example.
- The possibility of using non-dyadic wavelet decompositions [87], which enables the support of multiple HD formats is an interesting feature not only for the aforementioned application but also for video surveillance and for mobile video applications.
- WSVC with temporal filtering can also be adapted so that some of the allowable operating points can be decoded by J2K and MJ2K (MJ2K is only "intra" coded video using J2K; if J2K compatibility is preserved MJ2K compatibility will also remain).
- Enabling efficient similarity search in large video databases. Different methods based on wavelets exist. Instead of searching full resolution videos, one can search low quality videos (spatially, temporally, and reduced SNR), to accelerate the search operation. Starting from salient points, on low quality videos, similarities can be refined in space and time.
- Multiple Description Coding which would lead to better error-resilience. By using the lifting scheme, it is easy to replicate video data to transmit two similar bit streams. Using intelligent splitting, one can decode independently or jointly the two bit streams (spatially, temporally and/or SNR reduced).
- Space variant resolution adaptive decoding. When encoding the video material, it is possible to decode at a high spatial resolution only a certain region while keeping at a lower resolution the surrounding areas.

VIII. Conclusion

This paper has provided a picture of various tools that have been designed in recent years for scalable video compression. It has focused on multi-resolution representations with the use of the Wavelet Transform and its extensions to handle motion in image sequences. A presentation of benefits of WT based approaches has also been suggested for various application perspectives. The paper has also shown how it is possible to organize various multi-resolutions tools into different architecture families depending on the (space/time) order in which the signal transformation is operated. All such architectures make it possible to generate embedded bit-streams that can be parsed to handle combined quality, spatial and temporal scalability requirements. This has been extensively simulated during the exploration activity on wavelet video coding as carried out effort by ISO/MPEG, demonstrating that credible performance results could be achieved. Nevertheless, when compared to the current SVC standardization effort, which provides a pyramidal extension of MPEG-4/AVC, it appears that though close, the performance of wavelet based video codecs remains inferior. This seems partly due to:

- the level of fine (perceptual and/or rate-distortion theoretic) optimization that can be achieved for hybrid (DCT+MEC) codecs because of the deep understanding on how all its components interact;
- the spectral characteristics of standard resolution video material on which standard multiresolution (wavelet based) representation already exhibit lower performance even in intra-mode (some conducted preliminary experiments on high resolution material are showing increasing performance for WT approaches) ;
- the inability of the experimented multiresolution (WT) approaches to effectively represent the information locality in image sequences (block based transform can be instead more easily adapted to handle the non stationarity of visual content, in particular by managing uncovered/covered areas through a local effective representation of intracoded blocks).

This last argument deserves probably some further considerations. For image compression, WT based approaches are showing quite competitive performance due to the energy compaction ability of the WT to handle piecewise polynomials that are known to well describe many natural images. In video sequences, the adequacy of such model falls apart unless a precise alignment of moving object trajectories can be achieved. This might remain only a challenge, since as for any segmentation problem, it is difficult to achieve it in a robust fashion, due to the complex information modeling which is often necessary. As an alternative, new transform families [88] could also be studied so as to model the specificity of spatiotemporal visual information.

IX. Acknowledgments

This paper is an attempt to summarize the current understanding in wavelet video coding as carried out collectively in the framework of the ISO/MPEG standardization. While it has not yet been possible to establish wavelet video coding as an alternative video coding standard for SVC, its performance appears on the rise when compared to previous attempts to establish credibly competitive video coding solutions with respect to hybrid motion-compensated transform based approaches. We hope that the picture offered will help many future researchers to undertake work in the field and shape the future of wavelet video compression. The authors would like to acknowledge the efforts of colleagues and friends who have helped to establish the state-of-the-art in the field as described in this work. Particular thanks are meant for all those institutions (academia, funding organizations, research centers and industries) which continue to support work in this field and in particular the people who have participated in the ISO/MPEG wavelet video coding exploration activity during several years, e.g. Christine Guillemot, Woo-Jin Han, Jens Ohm, Stéphane Pateux, Beatrice Pesquest-Popescu, Julien Reichel, Mihaela Van der Shaar, David Taubmann, John Woods, Ji-Zheng Xu, just to name a few.

References

[1] ISO/MPEG Video, "Registered responses to the call for proposals on scalable video coding," ISO/IEC JTC1/SC29/WG11, 68th MPEG Meeting, Munich, Germany, Tech. Rep. M10569/S01-S24, Mar. 2004.

[2] ——, "Joint scalable video model (JSVM) 6.0," ISO/IEC JTC1/SC29/WG11, 76th MPEG Meeting, Montreux, Switzerland, Tech. Rep. N8015, Apr. 2006.

[3] R. Leonardi, T. Oelbaum, and J.-R. Ohm, "Status report on wavelet video coding exploration," ISO/IEC JTC1/SC29/WG11, 76th MPEG Meeting, Montreux, Switzerland, Tech. Rep. N8043, Apr. 2006.

[4] ISO/MPEG Requirements, "Applications and requirements for scalable video coding," ISO/IEC JTC1/SC29/WG11, 71st MPEG Meeting, Hong Kong, China, Tech. Rep. N6880, Jan. 2005.

[5] ISO/MPEG Video, "Description of core experiments in MPEG-21 scalable video coding," ISO/IEC JTC1/SC29/WG11, 69th MPEG Meeting, Redmond, WA, USA, Tech. Rep. N6521, July 2004.

[6] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, pp. 532–540, Apr. 1983.

[7] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA, USA: Academic Press, 1998.

[8] P. Vaidyanathan, *Multirate systems and filter banks*. Prentice Hall PTR, 1992.

[9] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*. Englewood Cliffs, NJ: Prentice-Hall, 1995.

[10] I. Daubechies, *Ten lectures on wavelets*. Philadelphia, PA, USA: SIAM, 1992.

[11] W. Sweldens, "The lifting scheme: a construction of second generation wavelets," *SIAM J. Math. Anal.*, vol. 29, no. 2, pp. 511–546, 1997.

[12] I. Daubechies and W. Sweldens, "Factoring wavelet transforms into lifting steps," *J. Fourier Anal. Appl.*, vol. 4, no. 3, pp. 247–269, 1998.

[13] A. Calderbank, I. Daubechies, W. Sweldensand, and B.-L. Yeo, "Wavelet transforms that map integers to integers," *Appl. Comput. Harmon. Anal.*, vol. 5, no. 3, pp. 332–369, 1998.

[14] H. Heijmans and J. Goutsias, "Multiresolution signal decomposition schemes. part 2: Morphological wavelets," CWI, Amsterdam, The Nederlands, Tech. Rep. PNA-R9905, June 1999.

[15] J.-R. Ohm, *Multimedia Communication Technology*. Springer Verlag, 2003.

[16] J. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Trans. Image Processing*, vol. 3, pp. 559–571, Sept. 1994.

[17] S.-J. Choi and J. Woods, "Motion-compensated 3-d subband coding of video," *IEEE Trans. Image Processing*, vol. 8, no. 2, pp. 155–167, Feb. 1999.

[18] B.-J. Kim, Z. Xiong, and W. Pearlman, "Low bit–rate scalable video coding with 3D set partitioning in hierarchical trees (3D SPIHT)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, pp. 1374–1387, Dec. 2000.

[19] A. Secker and D. Taubman, “Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression,” *IEEE Trans. Image Processing*, vol. 12, no. 12, pp. 1530–1542, Dec. 2003.

[20] G. Pau, C. Tillier, B. Pesquet-Popescu, and H. Heijmans, “Motion compensation and scalability in lifting–based video coding,” *Signal Processing: Image Communication*, vol. 19, no. 7, pp. 577–600, Aug. 2004.

[21] B. Pesquet-Popescu and V. Bottreau, “Three dimensional lifting schemes for motion compensated video compression,” in *IEEE International Conference on Acoustic Speech and Signal Processing (ICASSP’01)*, vol. 3, Salt Lake City, USA, May 2001, pp. 1793–1796.

[22] J. Ohm, “Motion-compensated wavelet lifting filters with flexible adaptation,” in *Thyrrenian Int. Workshop on Digital Communications, IWDC 2002*, Capri, Italy, Sept. 2002.

[23] Y. Zhan, M. Picard, B. Pesquet-Popescu, and H. Heijmans, “Long temporal filters in lifting schemes for scalable video coding,” ISO/IEC JTC1/SC29/WG11, 61th MPEG Meeting, Klagenfurt, Austria, Tech. Rep. M8680, July 2002.

[24] J. Xu, Z. Xiong, S. Li, and Y.-Q. Zhang, “Three-dimensional embedded subband coding with optimal truncation (3D ESCOT),” *Appl. Comput. Harmon. Anal.*, vol. 10, pp. 290–315, 2001.

[25] H. Schwarz, D. Marpe, and T. Wiegand, “Scalable extension of h.264/avc,” ISO/IEC JTC1/SC29/WG11, 68th MPEG Meeting, Munich, Germany, Tech. Rep. M10569/S03, Mar. 2004.

[26] R. Xiong, F. Wu, S. Li, Z. Xiong, and Y.-Q. Zhang, “Exploiting temporal correlation with block-size adaptive motion alignment for 3D wavelet coding,” in *Visual Communications and Image Processing (VCIP’04)*, SPIE, Ed., vol. 5308, San Jose, CA, Jan. 2004, pp. 144–155.

[27] A. Secker and D. Taubman, “Highly scalable video compression using a lifting-based 3D wavelet transform with deformable mesh motion compensation,” in *IEEE International Conference on Image Processing (ICIP’02)*, Sept. 2002, pp. 749–752.

[28] L. Luo, F. Wu, S. Li, Z. Xiong, and Z. Zhuang, “Advanced motion threading for 3D wavelet video coding,” *Signal Processing: Image Communication*, vol. 19, no. 7, pp. 601–616, 2004.

[29] D. Turaga, M. van der Schaar, and B. Pesquet-Popescu, “Temporal prediction and differential coding of motion vectors in the mctf framework,” in *IEEE International Conference on Image Processing (ICIP’03)*, Barcelona, Spain, Sept. 2003.

[30] V. Valéntin, M. Cagnazzo, M. Antonini, and M. Barlaud, “Scalable context-based motion vector coding for video compression,” in *Picture Coding Symposium, PCS’03*, Saint-Malo, F, Apr. 2003, pp. 63–68.

[31] J. Xu, R. Xiong, B. Feng, G. Sullivan, M.-C. Lee, F. Wu, and S. Li, “3D subband video coding using barbell lifting,” ISO/IEC JTC1/SC29/WG11, 68th MPEG Meeting, Munich, Germany, Tech. Rep. M10569/S05, Mar. 2004.

[32] C. Tillier, B. Pesquet-Popescu, and M. van der Schaar, “Highly scalable video coding by bidirectional predict-update 3-band schemes,” in *ICASSP’04*, Montreal, Canada, May 2004.

[33] M. V. der Schaar and D. Turaga, “Unconstrained motion compensated temporal filtering (UMCTF) framework for wavelet video coding,” in *IEEE International Conference on Acoustic Speech and Signal Processing (ICASSP’03)*, vol. 3, Hong-Kong, China, Apr. 2003, pp. 81–84.

[34] N. Mehrseresht and D. Taubman, “Adaptively weighted update steps in motion compensated lifting based on scalable video compression,” in *IEEE International Conference on Image Processing (ICIP’03)*, vol. 3, Barcelona, Spain, Sept. 2003, pp. 771–774.

[35] D. Maestroni, M. Tagliasacchi, and S. Tubaro, “In–band adaptive update step based on local content activity,” in *Visual Communications and Image Processing (VCIP’05)*, SPIE, Ed., vol. 5960, Beijing, China, July 2005, pp. 169–176.

[36] ISO/MPEG Video, “Wavelet codec reference document and software manual,” ISO/IEC JTC1/SC29/WG11, 73th MPEG Meeting, Poznan, Poland, Tech. Rep. N7334, July 2005.

[37] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF," in *International Conference on Multimedia & Expo, ICME'06*, Toronto, Canada, July 2006.

[38] D. Turaga, M. van der Schaar, and B. Pesquet-Popescu, "Complexity scalable motion compensated wavelet video encoding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 8, pp. 982–993, Aug. 2005.

[39] C. Parisot, M. Antonini, and M. Barlaud, "Motion-conpensated scan based wavelet transform for video coding," in *Thyrrenian Int. Workshop on Digital Communications, IWDC 2002*, Capri, Italy, Sept. 2002.

[40] G. Pau, B. Pesquet-Popescu, M. van der Schaar, and J. Viéron, "Delay-performance trade-offs in motion-compensated scalable subband video compression," in *ACIVS'04*, Brussels, Belgium, Sept. 2004.

[41] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.

[42] A. Said and W. A. Pearlman, "A new fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 243–250, June 1996.

[43] S.-T. Hsiang and J. W. Woods, "Embedded image coding using zeroblocks of subband/wavelet coefficients and context modeling," in *MPEG-4 Workshop and Exhibition at ISCAS 2000*, Geneva, Switzerland, May 2000.

[44] S. D. Servetto, K. Ramchandran, and M. T. Orchard, "Image coding based on a morphological representation of wavelet data," *IEEE Trans. Image Processing*, vol. 8, pp. 1161–1174, Sept. 1999.

[45] F. Lazzaroni, R. Leonardi, and A. Signoroni, "High-performance embedded morphological wavelet coding," *IEEE Signal Processing Lett.*, vol. 10, no. 10, pp. 293–295, Oct. 2003.

[46] ——, "High-performance embedded morphological wavelet coding," in *Thyrrenian Int. Workshop on Digital Communications, IWDC 2002*, Capri, Italy, Sep. 2002, pp. 319–326.

[47] F. Lazzaroni, A. Signoroni, and R. Leonardi, "Embedded morphological dilation coding for 2D and 3D images," in *Visual Communcations Image Processing 2002*, vol. SPIE-4671, San José, California, Jan. 2002, pp. 923–934.

[48] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Trans. Image Processing*, vol. 9, pp. 1158–1170, July 2000.

[49] D. Taubman, E. Ordentlich, M. Weinberger, and G. Seroussi, "Embedded block coding in JPEG 2000," *Signal Proc.: Image Comm.*, vol. 17, pp. 49–72, Jan. 2002.

[50] J. Li and S. Lei, "Rate-distortion optimized embedding," in *Picture Coding Symp.*, Sept. 1997, pp. 201–206.

[51] S.-T. Hsiang and J. Woods, "Embedded video coding using invertible motion compensated 3-d subband/wavelet filter bank," *Signal Processing: Image Communication*, vol. 16, no. 8, pp. 705–724, May 2001.

[52] P. Chen and J. W. Woods, "Bidirectional MC-EZBC with lifting implementation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, pp. 1183–1194, Oct. 2004.

[53] N. Adami, M. Brescianini, M. Dalai, R. Leonardi, and A. Signoroni, "A fully scalable video coder with inter-scale wavelet prediction and morphological coding," in *Visual Communications and Image Processing (VCIP'05*, SPIE, Ed., vol. 5960, Beijing, China, July 2005, pp. 535–546.

[54] V. Bottreau, M. Benetiere, B. Felts, and B. Pesquet-Popescu, "A fully scalable 3D subband video codec," in *IEEE International Conference on Image Processing (ICIP'01)*, vol. 2, Thessaloniki, Greece, Oct. 2001, pp. 1017–1020.

[55] K. Ho and D. Lun, "Efficient wavelet based temporally scalable video coding," in *IEEE International Conference on Image Processing (ICIP'02)*, vol. 1, New York, USA, Aug. 2002, pp. 881–884.

[56] N. Cammas, "Codage video scalable par maillages et ondelettes t+2d," Ph.D. dissertation, Univ. of Rennes 1, IRISA, Rennes, France, Nov. 2004. [Online]. Available: http://www.inria.fr/rrrt/tu-1122.html

[57] G. Pau, "Ondlettes et décompositions spatio–temporelles avancées; application au codage vidéo scalable," Ph.D. dissertation, École nationale supérieure de télecomunications, Paris, France, 2006.

[58] M. Flierl and B. Girod, "Video coding with motion-compensated lifted wavelet transforms," *Signal Processing: Image Communication*, vol. 19, no. 7, pp. 561–575, Aug. 2004.

[59] J. Barbarien, A. Munteanu, F. Verdicchio, Y. Andreopoulos, J. Cornelis, and P. Schelkens, "Motion and texture rate–allocation for prediction–based scalable motion–vector coding," *Signal Processing: Image Communication*, vol. 20, no. 4, pp. 315–342, Apr. 2005.

[60] M. Mrak, N. Sprljan, G. Abhayaratne, and E. Izquierdo, "Scalable generation and coding of motion vectors for highly scalable video coding," in *Picture Coding Symposium, PCS'04*, San Francisco, CA, USA, 2004.

[61] V. Bottreau, C. Guillemot, R. Ansari, and E. Francois, "Svc ce5: spatial transform using three lifting steps filters," ISO/IEC JTC1/SC29/WG11, 70th MPEG Meeting, Palma de Mallorca, Spain, Tech. Rep. M11328, Oct. 2004.

[62] Y. Wu and J. Woods, "Aliasing reduction for scalable subband/wavelet video coding," ISO/IEC JTC1/SC29/WG11, 73rd MPEG Meeting, Poznan, Poland, Tech. Rep. M12376, July 2005.

[63] Y. Andreopoulos, A. Munteanu, J. Barbarien, M. van der Schaar, J. Cornelis, and P. Schelkens, "Inband motion compensated temporal filtering," *Signal Processing: Image Communication*, vol. 19, no. 7, pp. 653–673, Aug. 2004.

[64] 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, pp. 577–587, 2000.

[65] Y. Andreopoulos, A. Munteanu, G. V. D. Auwera, J. Cornelis, and P. Schelkens, "Complete-to-overcomplete discrete wavelet transforms: theory and applications," *IEEE Trans. Signal Processing*, vol. 53, no. 4, pp. 1398–1412, Apr. 2005.

[66] ISO/MPEG Video, "Subjective test results for the CfP on scalable video coding technology," ISO/IEC JTC1/SC29/WG11, 68th MPEG Meeting, Munich, Germany, Tech. Rep. N6383, Mar. 2004.

[67] N. Mehrseresht and D. Taubman, "An efficient content-adaptive MC 3D-DWT with enhanced spatial and temporal scalability," in *IEEE International Conference on Image Processing (ICIP'04)*, Oct. 2004.

[68] D. Taubman and N. Mehrseresht and R. Leung, "SVC technical contribution: Overview of recent technology developments at UNSW," ISO/IEC JTC1/SC29/WG11, 69th MPEG Meeting, Redmond, WA, USA, Tech. Rep. M10868, July 2004.

[69] N. Sprljan, M. Mrak, T. Zgaljic, and E. Izquierdo, "Software proposal for wavelet video coding exploration group," ISO/IEC JTC1/SC29/WG11, 75th MPEG Meeting, Bangkok, Thailand, Tech. Rep. M12941, Jan. 2006.

[70] M. Mrak, N. Sprljan, T. Zgaljic, N. Ramzan, S. Wan, and E. Izquierdo, "Performance evidence of software proposal for wavelet video coding exploration group," ISO/IEC JTC1/SC29/WG11, 6th MPEG Meeting, Montreux, Switzerland, Tech. Rep. M13146, Apr. 2006.

[71] C. Ong, S. Shen, M. Lee, , and Y. Honda, "Wavelet video coding - generalized spatial temporal scalability (GSTS)," ISO/IEC JTC1/SC29/WG11, 72nd MPEG Meeting, Busan, Korea, Tech. Rep. M11952, Apr. 2005.

[72] ISO/MPEG Video, "Scalable video model v 2.0," ISO/IEC JTC1/SC29/WG11, 69th MPEG Meeting, Redmond, WA, USA, Tech. Rep. N6520, July 2004.

[73] M. N. Do and M. Vetterli, "Framing pyramids," *IEEE Trans. Signal Processing*, vol. 51, no. 9, pp. 2329–2342, Sept. 2003.

[74] M. Flierl and P. Vandergheynst, "An improved pyramid for spatially scalable video coding," in *IEEE International Conference on Image Processing (ICIP'05)*, Genova, Italy, Sept. 2005.

[75] G. Rath and C. Guillemot, "Representing laplacian pyramids with varying amount of redundancy," in *EUSIPCO'06*, Florence, Italy, Sept. 2006.

[76] N. Adami, M. Brescianini, R. Leonardi, and A. Signoroni, “Svc ce1: Stool - a native spatially scalable approach to svc,” ISO/IEC JTC1/SC29/WG11, 70th MPEG Meeting, Palma de Mallorca, Spain, Tech. Rep. M11368, Oct. 2004.

[77] ISO/MPEG Video, “Report of the subjective quality evaluation for svc ce1,” ISO/IEC JTC1/SC29/WG11, 70th MPEG Meeting, Palma de Mallorca, Spain, Tech. Rep. N6736, Oct. 2004.

[78] R. Leonardi and S. Brangoulo, “Wavelet codec reference document and software manual v2.0,” ISO/IEC JTC1/SC29/WG11, 74th MPEG Meeting, Nice, France, Tech. Rep. N7573, Oct. 2005.

[79] D. Zhang, S. Jiao, J. Xu, F. Wu, W. Zhang, and H. Xiong, “Mode-based temporal filtering for in-band wavelet video coding with spatial scalability,” in *Visual Communications and Image Processing (VCIP’05*, SPIE, Ed., vol. 5960, Beijing, China, July 2005, pp. 355–363.

[80] M. Beermann and M. Wien, “Wavelet video coding, ee4: Joint reduction of ringing and blocking,” ISO/IEC JTC1/SC29/WG11, 74th MPEG Meeting, Nice, France, Tech. Rep. M12640, Oct. 2005.

[81] M. Li and T. Nguyen, “Optimal wavelet filter design in scalable video coding,” in *IEEE International Conference on Image Processing (ICIP’05)*, Genova, Italy, Sept. 2005.

[82] ISO/MPEG Video, “Description of testing in wavelet video coding,” ISO/IEC JTC1/SC29/WG11, 75th MPEG Meeting, Bangkok, Thailand, Tech. Rep. N7823, Jan. 2006.

[83] C. Fenimore, V. Baroncini, T. Oelbaum, and T. K. Tan, “Subjective testing methodology in MPEG video verification,” in *Applications of Digital Image Processing XXVII. Edited by Tescher, Andrew G. Proceedings of the SPIE, Volume 5558, pp. 503-511 (2004).*, ser. Presented at the Society of Photo-Optical Instrumentation Engineers (SPIE) Conference, A. G. Tescher, Ed., vol. 5558, Nov. 2004, pp. 503–511.

[84] M. Ouaret, F. Dufaux, and T. Ebrahimi, “On comparing JPEG2000 and intraframe AVC,” in *International Society for Optical Engineering SPIE Optics & Photonics*, vol. 1, Genova, Italy, Aug. 2006.

[85] D. Marpe, S. Gordon, and T. Wiegand, “H.264/MPEG4-AVC fidelity range extensions: Tools, profiles, performance, and application areas,” in *IEEE International Conference on Image Processing (ICIP’05)*, vol. 1, Genova, Italy, Sept. 2005, pp. 593–596.

[86] L. Lima, F. Manerba, N. Adami, A. Signoroni, and R. Leonardi, “Wavelet-based encoding for hd applications,” in *International Conference on Multimedia & Expo, ICME’07*, Beijing, China, July 2007.

[87] G. Pau and B. Pesquet-Popescu, “Image coding with rational spatial scalability,” in *EUSIPCO’06*, Florence, Italy, Sept. 2006.

[88] B. Wang, Y. Wang, I. Selesnick, and A. Vetro, “An investigation of 3d dual-tree wavelet transform for video coding,” in *IEEE International Conference on Image Processing (ICIP’04)*, Singapore, Oct. 2004.