A FULLY SCALABLE WAVELET VIDEO CODING SCHEME WITH HOMOLOGOUS INTER-SCALE PREDICTION

Nicola Adami, Michele Brescianini,
Riccardo Leonardi, Alberto Signoroni

Università degli Studi di Brescia

In this paper, we present a fully scalable wavelet-based video coding architecture called STP-Tool, in which motion-compensated temporal-filtered subbands of spatially scaled versions of a video sequence can be used as a base layer for inter-scale predictions. These predictions take place in a pyramidal closed-loop structure between homologous resolution data, i.e., without the need of spatial interpolation. The presented implementation of the STP-Tool architecture is based on the reference software of the Wavelet Video Coding MPEG Ad-Hoc Group. The STP-Tool architecture makes it possible to compensate for some of the typical drawbacks of current wavelet-based scalable video coding architectures and shows interesting objective and visual results even when compared with other wavelet-based or MPEG-4 AVC/H.264-based scalable video coding systems.

1. INTRODUCTION

In present and emerging video coding applications, there is an increasing demand for scalability. This is mainly due to the heterogeneous characteristics of devices and communication infrastructures which share the same source of video data, support different coding technologies and standards, and, in the case of an adequate degree of scalability, use the same compressed bit-stream. Scalability is related to the possibility (in any time and system configuration) of having direct access to the right amount of coded information (i.e., avoiding over-transmission or data format conversion or transcoding) in order to access, communicate, and use the desired video content with respect to the allowable transmission throughput and receiving device features in an optimal manner. Academic and industrial communities are more and more convinced that a combination of different scalability attributes (here and quite commonly referred to as *full scalability*) can be achieved without sacrificing coding performance. Full scalability in terms of reconstruction quality (SNR) and spatio-temporal resolution is usually required to adapt optimally and dynamically to the size of displaying terminals, to the related frame-rate reproduction capabilities and/or power saving (temporary or structural) needs, and to the available throughput on communication networks channels and nodes. Fig. 1 shows a typical SVC system referring to the coding of a video

signal at an original resolution CIF (288 height x 352 width) and a framerate of 30 fps. In the example, the higher operating point and decoded quality corresponds to a bit-rate of 2Mbps with respect to the original spatial and temporal resolutions. For a scaled decoding in terms of spatial and/or temporal and/or quality resolution, the decoder only works on a portion of the original coded bit stream according to the indication of the desired working point. Such stream portion is extracted from the originally coded stream by a block called "extractor."

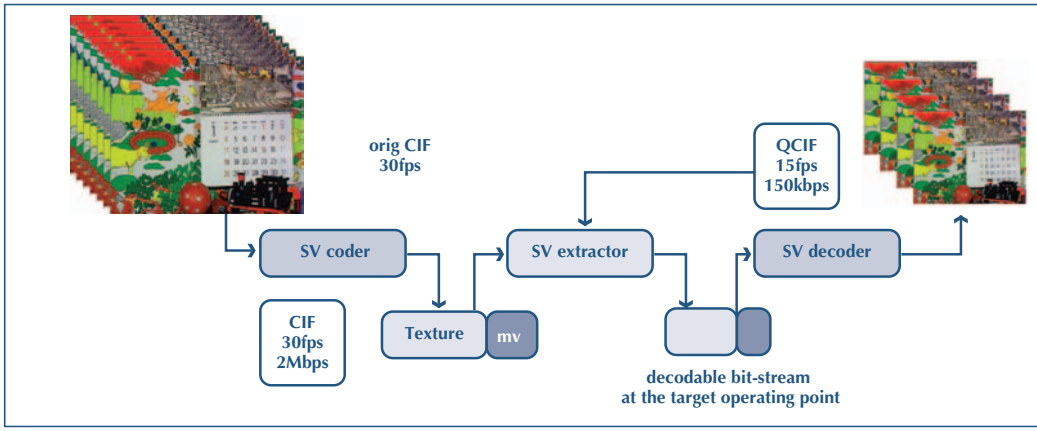

FIGURE 1: SVC RATIONALE.

In Fig. 1, it is shown to be arranged between coder and decoder. According to the application field, it can be realized as an independent block, or it can be an integral part of the coder or decoder. The extractor receives the information relating 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 matching or almost matching the specifications of the indicated working point. One of the main differences between an SVC system and a transcoding system is the low complexity of the extraction block that does not require coding/decoding operations and typically consists of simple "cut and paste" operations on the coded bit-stream.

Scalable Wavelet Video Coding (SVC) is an expanding research field always needing technologies which enable and support scalability in various dimensions: spatial and temporal resolution, SNR and/or visual quality, complexity, and sometimes others [1]. In the last years, stimulated by the advances in the field of wavelet video coding and by the work of a series of MEPG Ad-Hoc Groups on the subject, competitive (in terms of coding performance) fully Scalable Video Coders (SVC) have begun to appear. These are based on core technologies inspired to both highly optimized single-rate hybrid coding schemes (mainly MPEG-4 AVC/H.264) and newer schemes based on the spatio-temporal wavelet transform. In particular, some MPEG-4 AVC-based and wavelet-based SVC schemes have been perceptually compared [2] at the 70th MPEG meeting (Palma de Mallorca, October 2004), and the former solutions, already optimized in most aspects, have performed better on the assigned testing conditions, while some of the latter (including our STP-Tool system [3, 4]) demonstrated similar performances manifesting their limits/potentials and their need of further maturation. Hence, in MPEG, the decision to start a new SVC standard [5] in conjunction with ITU-T (JVT-SVC) was made, taking into consideration wavelet-based solutions for longer-term objectives and applications [6-8]. After this meaningful, but in some aspects unnatural, comparison with respect to today's predominant video coding technologies, wavelet-based SVC research can resume on new and less constrained paths.

Let us recall why the discrete wavelet transform (DWT) is a congenial tool to be used in a SVC perspective. A digital video

can be decomposed according to a compound of spatial DWT and wavelet-based motion-compensated temporal filtering (MCTF) [9]. In general, wavelet-based SVC systems can be conceived according to different kinds of spatio-temporal decomposition structures designed to produce a multiresolution spatio-temporal subband hierarchy, and then coded with a progressive or quality scalable coding technique [10-14]. A classification of SVC architectures has been suggested by the MPEG Ad-Hoc Group on SVC [15]. The so called t+2D schemes (one example is [11]) first performs an MCTF, producing temporal subband frames; then the spatial DWT is applied on each one of these frames. Alternatively, in a 2D+t scheme (one example is [16]), a spatial DWT is applied first to each video frame, and then MCTF is made on spatial subbands. A third approach named 2D+t+2D uses a first stage DWT to produce reference video sequences at various resolutions; t+2D transforms are then performed on each resolution level of the obtained spatial pyramid.

As already stated, each scheme has evidenced its pros and cons [17, 2] in terms of coding performance. From a theoretical point of view, the critical aspects of the above SVC schemes mainly reside:

- in the coherence and trustworthiness of the motion estimation at various scales (especially for t+2D schemes);
- in the difficulties of compensating for the shift-variant nature of the wavelet transform (especially for 2D+t schemes); and
- in the performance of inter-scale prediction (ISP) mechanisms (especially for 2D+t+2D schemes).

Our STP-Tool, which belongs to the 2D+t+2D class, was initially presented in [3]. In this paper, we better explain and justify the STP-Tool solution (Section 2) and compare the related features with respect to other wavelet-based SVC schemes. In Section 3, we present some recent architectural advancements and the first implementation of the STP-Tool scheme on the Wavelet Video Coding MPEG reference software. In Section 4, we present some experimental results with particular emphasis on visual performance comparison and on a fair PSNR comparison among SVC reference systems.

2. STP-TOOL PRINCIPLES

Spatial scalability can be obtained in coding schemes by using the lower spatial resolution information (at spatial level *s*) as a base-layer from which the finer resolution (at spatial level *s*+1) can be predicted. According to a common pyramidal approach [15,18], the inter-scale prediction (here abbreviated with ISP) is obtained by means of data interpolation from level s to level s+1. Our STP-Tool idea [3] consists of performing an ISP where, by means of proper (e.g., reversible) spatial transforms, reference and predicted information are always compared at the same spatial resolution (possibly after being subjected to the same kind of spatio-temporal transformations). From this principle, we can derive different STP-Tool architectures which are typically of the 2D+t+2D kind but where ISP predictions take place without the need of data interpolation. STP-Tool architectures can be configured to be fully space-time-quality scalable [1], and multiple adaptation capabilities [19] can be also designed without sacrificing coding performance. As we will show, STP-Tool architectures solve or reduce the impact of some critical issues that afflict t+2D and 2D+t schemes.

One main way to subdivide 2D+t+2D architectures is between open-loop ISP (the prediction signal is obtained from the original information) and closed-loop ISP solutions (the prediction signal is obtained from the decoded information). In a purely closed-loop ISP scheme, the prediction signal used at a spatial level *s*+1 must collect all the decoded information coming from the previously coded prediction and residue signals. In a purely open loop scheme, the signal at spatial resolution *s* is directly taken as the prediction signal; then, prediction at spatial level *s*+1 only depends on spatial level *s*. However, ISP open loop schemes, especially at low bit-rates, undergo drift problems because part of the information used for prediction would not be available to the decoder. For this reason, open loop ISP schemes are no longer considered here.

2.1 Closed-loop ISP STP-Tool architecture

The closed-loop ISP STP-Tool architecture is presented in Fig. 2 for a 4CIF-CIF-QCIF spatial resolutions implementation. A main

characteristic of this fully scalable (SNR, spatial, and temporal resolution) scheme is its native dyadic spatial scalability. In fact, in the example of Fig. 2 (MEC stands for motion estimation and coding, T stands for spatial transform and EC stands for entropy coding, with coefficient quantization included), three different coding chains are performed. Each chain operates at a different spatial level and presents temporal and SNR scalability. Because of the information interdependencies at different scale layers, it is possible to re-use a suitable quality decoded (closed loop implementation) information of a coarser spatial resolution (e.g., spatial level s) in order to predict a finer spatial resolution level s+1. This is a requisite for every 2D+t+2D (spatially predictive) scheme. Our closed-loop STP-Tool approach differs from a classical 2D+t+2D approach mainly in two aspects:

1. the prediction is not performed in the data domain but between MCTF temporal subbands at spatial level s+1, named f_{s+1}, starting from the decoded MCTF subbands at spatial level s, dec(f_s);
2. rather than interpolating the decoded subbands, a single level spatial wavelet decomposition is applied to the portion of temporal subband frames f_{s+1} we want to predict. The prediction is then applied between dec(f_s) (closed-loop STP-Tool) and the low-pass (LL) component of the spatial wavelet decomposition, namely $DWT_L(f_{s+1})$. This has the advantage of feeding the quantization errors of dec(f_s) only into such low-pass components, which represent, at most, ¼ of the number of coefficients of the s+1 resolution level.

By adopting such a strategy, the predicted subbands $DWT_L(f_{s+1})$ and the predicting ones dec(f_s) have undergone the same number and type of spatio-temporal transformations, but in a different order (a temporal decomposition followed by a spatial one (t+2D) in the first case and a spatial decomposition followed by a temporal one in the second case (2D+t)). For the s+1 resolution, the prediction error $\Delta f_s = DWT_L(f_{s+1}) - \text{dec}(f_s)$ is further coded instead of $DWT_L(f_{s+1})$ (see the related detail in Fig. 3).

The question of whether and how the above predicted and reference subbands actually resemble each other cannot be taken for granted in a general framework. In fact, it strongly depends on the exact type of spatio-temporal transforms and on the way the motion is estimated and compensated for the various spatial levels. In order to achieve a reduction of the prediction error energy of Δf_s, the same type of transforms should be applied, and a certain degree of coherence between the structure and precision of the motion fields across the different resolution layers should be guaranteed.

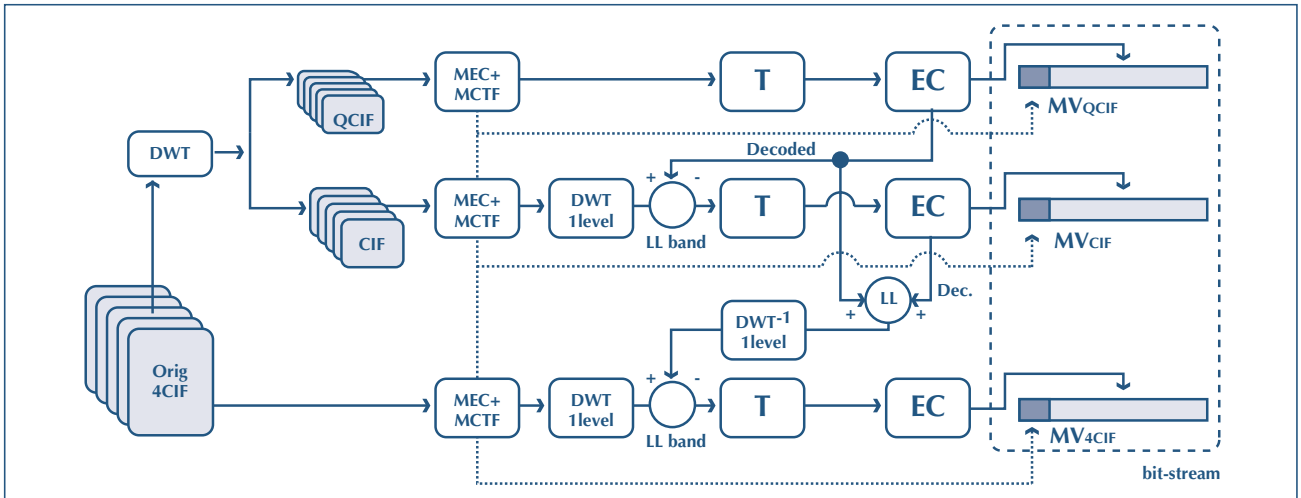

FIGURE 2: STP-TOOL CODING ARCHITECTURE.

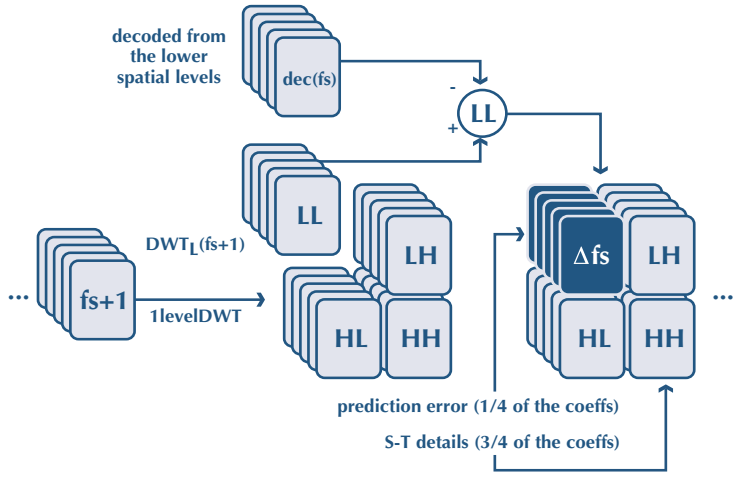

FIGURE 3: STP-TOOL PREDICTION DETAIL.

2.2 STP-Tool and other SVC architectures

We now aim at giving some insight about the differences between the proposed method and other existing techniques for hierarchical representation of video sequences. As explained in the previous section, the proposed method is essentially based on predicting the spatial low pass bands $DWT_L(f_{s+1})$ of the temporal subbands of a higher resolution level from the decoded temporal subbands $dec(f_s)$ of the lower resolution one. This method leads to a scheme that is quite different from previous wavelet-based SVC systems. An important thing to note is that the predicting coefficients and the predicted ones have been obtained by applying the same spatial filtering procedure to the video sequence but at different points with respect to the temporal filtering process. This implies that, even without quantizing and by perfectly scaling-down the motion field from high to low resolution, these coefficients are generally different. In fact, temporal and spatial transform would be interchangeable only if the spatio-temporal transform were separable, and this happens only when filtering and downsampling operations in spatial and temporal dimensions do not interfere each other. Now, due to motion compensation, the temporal filtering is, in general, misaligned with respect to the spatio-temporal downsampling grid. Thus, the motion-compensated spatio-temporal transform is not space-time separable.

Consequently, the prediction error contains not only the noise due to quantization of the low resolution sequence but also the effects of applying the spatial transform, in one case before, and in the other case after, the temporal decomposition. We note that this fact and its handling are crucial in wavelet-based video coding schemes, and differences between the $dec(f_s)$ and $DWT_L(f_{s+1})$ are responsible for a loss in performance in the t+2D schemes as explained hereafter.

2.2.1 t+2D

The following considerations about the differences between our STP-Tool scheme and the t+2D one reveal several advantages of the former one. A t+2D scheme acts on the original video sequence (at full spatial resolution) by applying a temporal MCTF decomposition followed by a spatial DWT; in other words, a spatial DWT transform is applied on each *temporal subband frame* issued by the MCTF. When full spatial resolution decoding is required, the process is reversed until the desired frame-rate (partial versus complete MCTF inversion) and SNR quality are reached; instead, if a lower spatial resolution version is needed, the inversion process discloses an incoherence with respect to the forward decomposition. The problem consists of the fact that the inverse MCTF transform is performed on the lower spatial resolution (obtained by the partial inversion

of the spatial DWT) of the temporal subband frames, and inverse motion compensation uses the same (scaled) motion field estimated for the higher resolution sequence analysis. Because of the non-ideal decimation performed by the low-pass wavelet decomposition (which generates spatial aliasing), a simply scaled motion field is, in general, not optimal to invert the temporal transform at lower resolution level. This problem can be reduced for intermediate and lower resolutions by using (for that resolution) more selective wavelet filters [20] or locally adaptive spectral shaping acting on the quantization parameters inside each spatial subband [21]. However, such approaches can determine coding performance loss at full resolution (because either wavelet filters or coefficient quantization laws are moved from coding performance *ideal* conditions).

Another relevant problem is represented by the ghosting artifacts that appear on the low pass temporal subbands when MC is not applied or when it fails due to unreliable motion vectors or to inadequate motion models. Such ghosting artifacts become visible when high pass subbands are discarded, that is, when reduced framerate decoding is performed. A solution to this issue has been proposed under the framework of *unconstrained* MCTF (UMCTF) [22], which basically consists of omitting the "update" lifting step so that only the "prediction" is performed in the lifting implementation of a MCTF. This solution does not take into account that the temporal update step is beneficial because it creates low-pass temporal subband frames, thus reducing temporal aliasing. Omitting it can cause visual coding performance worsening on reduced frame rate decoding, while, as stated above, keeping it causes ghosting artifacts where the MC model fails. A solution that tries to weight the update step adaptively according to a motion field reliability model parameter has been proposed in [23, 24].

In the common motion-compensated temporal filtering cases (e.g., with Haar or 5/3 kernels), an UMCTF approach actually leads to temporal open-loop versions of classical motion-compensated (uni- or bi-directional, respectively) temporal prediction schemes with eventually multiple reference frames, as supported in AVC. UMCTF is also used for low-delay and/or low-complexity wavelet-based SVC configurations (see e.g., [25]). A closed-loop version of UMCTF has been recalled "hierarchical B-frames prediction" and adopted in the latest version of JSVM [26].

2.2.2 2D+t

An alternative approach is the 2D+t configuration, where the spatial transform is applied before the temporal ones, which are then made on spatial subband group of frames (in-band MCTF). As in the t+2D case, MVs should be coded in a spatially scalable way, but MV scaling does not represent an issue here because each MCTF inversion is made with the original estimated MV's. Visual artifacts due to MC failures are also mitigated by the spatial transform.

Unfortunately, the 2D+t approach suffers from the shift-variant nature of the wavelet decomposition, leading to inefficiencies in the motion estimation and compensation of the spatial subbands. This problem has found a solution in schemes where motion estimation and compensation take place in an overcomplete (shift-invariant) wavelet domain [16], bringing texture coding back in the critically sampled wavelet domain. Inter-scale motion compensation coherence and increased computational complexity are among the residual problems of the 2D+t approach.

2.2.3 Pyramidal 2D+t+2D

From the above discussion, it becomes clear that the spatial and temporal wavelet filtering cannot be decoupled because of the motion compensation. As a consequence, it is not possible to encode different spatial resolution levels at once with only one MCTF; thus, both higher and lower resolution sequences must be MCTF-filtered. In this perspective, a possibility of obtaining good coding and scalability performance is to use ISP. What has been proposed to this end in the video coding literature is to use prediction between the lower and the higher resolution before applying the spatio-temporal transform. The low resolution

sequence is interpolated and used as prediction for the high resolution sequence. The residual is then filtered both temporally and spatially. Fig. 4 shows such an interpolation-based inter-scale prediction scheme. 2D+t+2D architectures have got their basis in the first hierarchical representation technique introduced for images, namely the Laplacian pyramid [27]. Thus, even if from an intuitive point of view the scheme seems to be well-motivated, it has the typical disadvantage of overcomplete representations, namely that of leading to a full-size residual image. This way, the detail (or refinement) information to be encoded comes spread on a high number of coefficients, and efficient encoding is hardly achievable. In the case of image coding, this drawback favored the research on the critically sampled wavelet transform as an efficient approach to image coding. In the case of video sequences, however, the corresponding counterpart would be a 2D+t scheme that we have already shown to be problematic due to the relative inefficiency of motion estimation and compensation across the spatial subbands.

The reference model considered for MPEG standardization [18] falls in this pyramidal family in that prediction is made just after the temporal transform but only on intra (not temporally transformed) blocks.

2.2.4 STP-Tool 2D+t+2D

Looking at the above issues, the STP-Tool idea leads to valid alternative approaches. It efficiently introduces the idea of prediction between different spatial resolution levels within the framework of spatio-temporal wavelet transforms. Compared with the previous schemes, it has several advantages. First of all, different spatial resolution levels undergo a MCTF, thus preventing incoherency of t+2D schemes from the MC inversion. Motion vectors can always be estimated hierarchically and coded in a scalable way, but MV optimization and reliability issues can be addressed at each spatial level in a more flexible way. Our experiments show that good coding performance can be obtained even with MV fields estimated and coded independently at each spatial level. Spatial aliasing reduction strategies and/or UMCTF solutions, as described for t+2D schemes, can be adopted as well.

In addition, in STP-Tool schemes, the MCTF's are applied before spatial DWT (they are not applied on high pass temporal subbands), thereby bypassing the 2D+t scheme issues.

Furthermore, contrary to what happens in pyramidal 2D+t+2D schemes, the prediction is restricted to a subset of the coefficients of the predicted signal, which is of the same size of

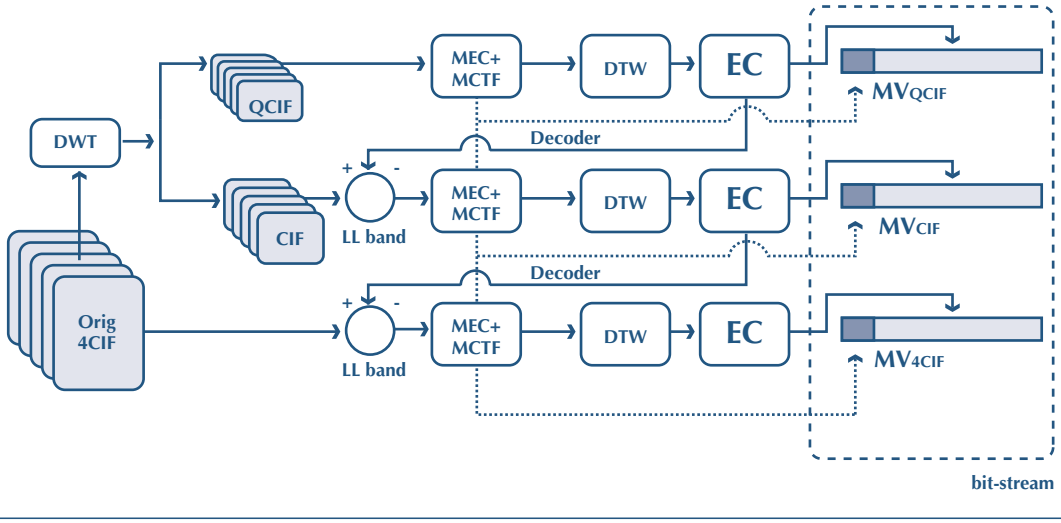

FIGURE 4: 2D+t+2D PYRAMIDAL SCHEME: PREDICTION WITH INTERPOLATION.

the prediction signal at the lower resolution. Therefore, there is a clear distinction between the coefficients that are interested in the prediction and the coefficients that are associated with higher spatio-temporal resolution details. This constitutes an advantage between the prediction schemes based on interpolation in the original sequence domain in that the subsequent coding can be adapted to the characteristics of the different sources.

From the above observations, we can remark that STP-Tool architecture is highly flexible in that it allows for several adaptations and additional features which in turn make it possible to preserve full scalability and improve coding performance.

3. ADVANCEMENTS AND IMPLEMENTATION OF THE STP-TOOL SCHEME

The results presented in this paper refer to an implementation of our scheme based on modifications of the MPEG Wavelet Video Coding (WVC) Reference Software [5], based on the *barbell lifting* MCTF [14], which accommodates the 2D+t+2D STP-Tool interband prediction mechanism, as described in Document [28]. In the following paragraphs, the most relevant implementation solutions which exploit the STP-Tool adaptability in order to improve coding performance are presented.

3.1 AVC base-layer

The considered STP-Tool codec is compatible with the use of an external base-layer bit-stream. We used the AVC base-layer functionality of the WVC reference software in our experiments. Visual results at various resolution levels take advantage of this choice because of the smoothing characteristics of AVC, which actually produce a good prediction signal even if not generated by means of a DWT as the predicted one. This fact also tells us that the STP-Tool principle is somehow "robust" in the sense that it could be used also in a not strictly wavelet-based coding environment.

3.2 STP-Tool prediction on a subset of temporal frames

Explanation of the present and following paragraphs is supported by the Fig. 5 which shows, for two different levels of spatial resolution (for example, CIF and QCIF), the result of 3 levels of MCTF. Eight video frames (V) at an initial temporal resolution of M fps are decomposed, resulting in three temporal decomposition levels (FTi). After the first decomposition, eight temporal frames FT3 are obtained, subdivided in four temporal subband frames at a temporal resolution M/2 and four temporal subband details (level of detail 3). The temporal decomposition is iterated only on the four frames at resolution M/2 to obtain four temporal subbands FT2, i.e., two frames at a temporal resolution M/4 and two details at level 2; similarly, an iteration is carried out on the two frames at a resolution M/4 to obtain one frame at a resolution M/8, called level 0, and one detail at level 1.
The frame at level 0 and the details at level 1, 2 and 3 together can allow for a perfect reconstruction of V [28]. To better understand the temporal references of the single transformed frames, on each frame, a black rectangle representing a moving object is shown. Such an object is found on all frames at a low resolution and the corresponding detail information (white rectangles) on all detail frames.

The STP-Tool prediction can be limited on a subset of the MCTF subbands, while the remaining subbands are directly coded. Fig. 5 indicates, by the rectangle dashed in line-dot, that only the (0,1) subbands could be involved in STP-Tool prediction instead of the whole group (3,2,1,0). The selection of which subbands should be involved in STP-Tool prediction can be empirical or computational (data content based or R-D based). It is also important to note that this degree of freedom is not allowed for data domain prediction schemes, like the scheme of Fig. 4. At the time, we have explored empirical solutions and remarked that a non-negligible coding gain can be obtained by using this degree of freedom.

3.3 STP-Tool prediction on an adapted temporal decomposition depth

Another degree of freedom that can be used within the STP-Tool prediction mechanism consists of adapting the temporal

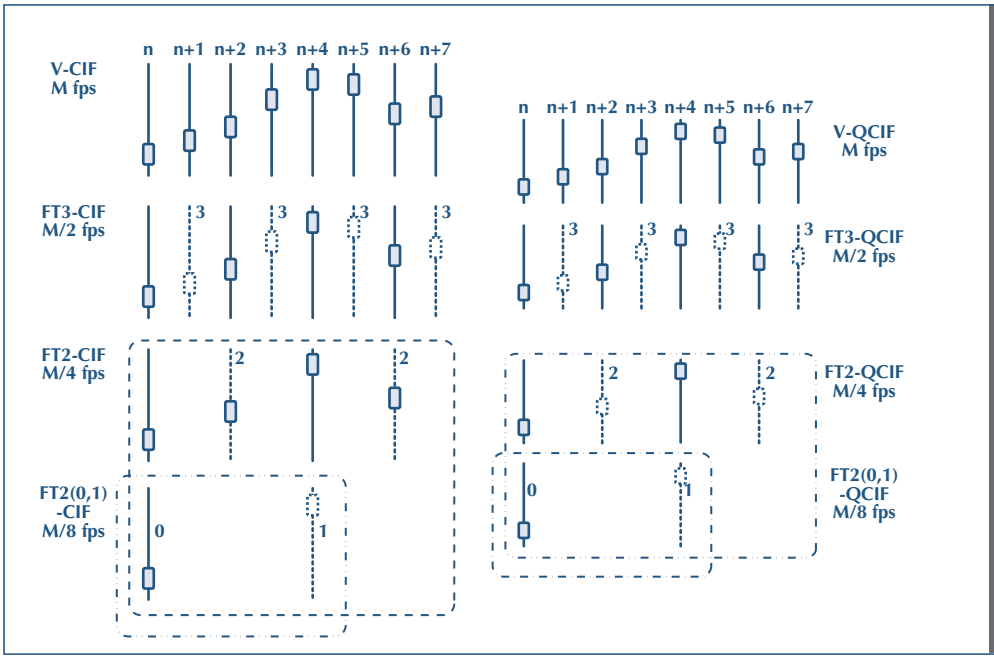

FIGURE 5: POSSIBILE VARIATIONS OF THE STP-TOOL PREDICTION ON THE MCTF SUBBAND HIERARCHY.

decomposition level at which the prediction takes place, according to the target temporal resolution. This is because, in many likely applications, the temporal decomposition depth (or the target frame rate) could not be the same for all spatial resolutions in a prediction pyramid; for example, lower resolutions could also involve lower frame rate reproductions for certain video streaming scenarios. In the example of Fig. 5, two temporal decompositions are shown, one for the CIF and the other for the QCIF resolution level, starting from reference videos at M fps. Let us suppose that the maximum expected target rate for CIF resolution is M/2 fps, while it goes down to M/8 fps for the QCIF resolution. In this case, we can perform the STP-Tool prediction in two somehow opposite ways (and other intermediate ones):

a. by executing a 3-level temporal decomposition for the CIF video in order to be able to perform a STP-Tool prediction adapted to the needed temporal decomposition depth at QCIF level (in Fig. 5, the dashed rectangle contains the additional subbands); and
b. by temporally transforming the reference videos according to their needed levels (e.g., one level for CIF and three levels for QCIF) and, in order to perform the STP-Tool prediction, by inverting the supernumerary levels (in Fig. 5, the rectangle dashed in line-double dot contains the inverted subbands in our example).

We tested both solutions a and b and remarked that the second one usually performs better in that prediction of low-pass temporal subbands is more effective than prediction of details.

3.4 Asymmetric closed-loop ISP

Another degree of freedom that we have in implementing a STP-Tool SVC architecture is the possibility of using an asymmetric closed-loop prediction.

This gave us sensible coding performance improvements in extracting critical operating points, especially when using a multiple and adaptive extraction path (see next section). The idea is depicted in Fig. 6, where for clearness only two spatial levels are considered. The coded base layer bit-stream can be entirely used (until it reaches the maximum of its assigned dimension, D_{max}) for base-layer video reconstruction.

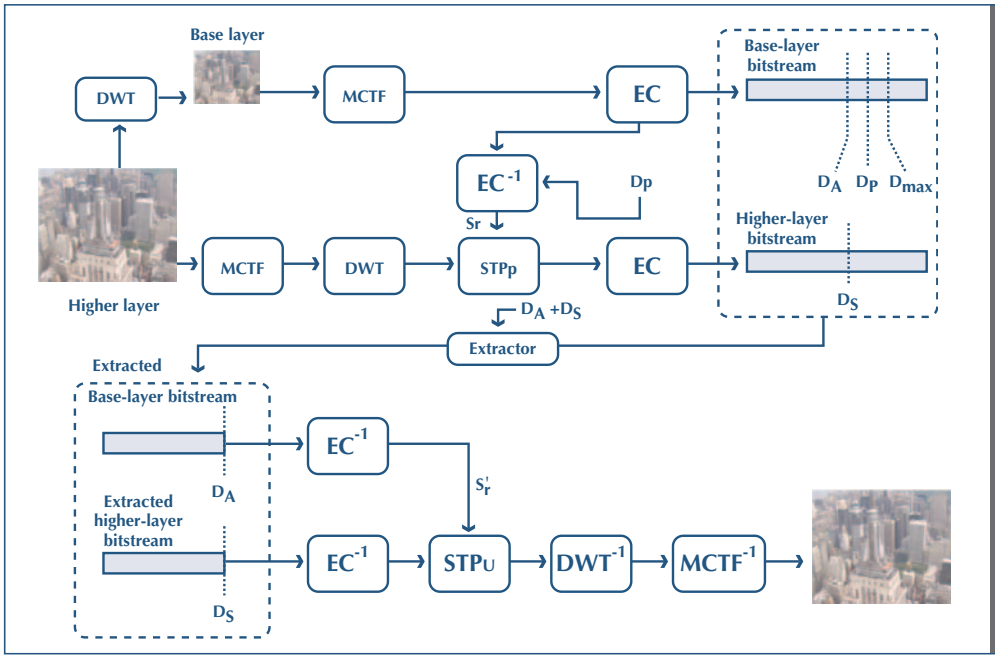

FIGURE 6: ASYMMETRIC CLOSED LOOP STP-TOOL PREDICTION.

An ordinary closed-loop STP-Tool mechanism consists of using, at the encoder side, a bit-stream portion D_P (measured in bps), corresponding to a suitable quality of the reconstructed signal s_r, in order to predict the higher spatial level. The same portion D_P should be normally extracted and used at the decoder side in order to update the prediction error. Instead, an asymmetry in this mechanism actually allows the use of a sub-portion D_A of portion D_P for updating the prediction (causing s'_r to differ form s_r). Keeping the (D_P-D_A) spread within certain limits, and considering the fact that a target extraction of a higher resolution operating point undergoes a $D=D_A+D_S$ target dimension, a coding gain can be achieved by exploiting this asymmetry. In general, the more suitable value to assign to D_A with respect to constraint D or with respect to an entire extraction path can be decided by the extractor or by other components over the coding-decoding chain. Such a decision can be accomplished by means of heuristic or tabular rules (without requiring complex calculations) or with computational methods (R-D optimization). Moreover, the asymmetric closed-loop approach can be easily extended to the case of more than two spatial levels. At the time, all our tests evidenced that, depending on the operating point, suitable D_A choices can be made which determine actual coding gains.

4. EXPERIMENTAL RESULTS

In Section 4.1, we present experimental results regarding an objective and visual comparison between the pyramidal and STP-Tool 2D+t+2D schemes. Section 4.2 illustrates a PSNR comparison among the VWC reference software in t+2D configuration [5], the same software in STP-Tool configuration [29], and the JVT-SVC reference software (JSVM3.0 [18]). JSVM operated with configuration files provided on the MPEG SVC repository; slightly better results can be obtained by more complex parameters and tools optimizations and by enabling closed-loop hierarchical B-frames prediction [30]. For a more uniform comparison the *AVC base-layer* mode has been enabled in both VWC reference software configurations. For a meaningful objective evaluation, we have tried to handle the manifest problem of different reference signal sequences associated with different schemes, and we propose a solution aimed at fair comparisons.

Finally, in Section 4.3, we show some visual performance results which evidence, in many cases, the superiority of STP-Tool with respect to the t+2D configuration and visual performances not too far from the highly optimized AVC/H.264-based JSVM. One major issue concerning the present implementation of STP-Tool is that MV at various scales is estimated and coded independently instead of being estimated and coded in a scalable way as well.

Despite this fact, obtained experimental results have been favourable. MV estimation and coding improvements are under consideration for future STP-Tool implementations.

Multiple extraction paths as defined and required for the Palma MPEG meeting [19] were always guaranteed in all experiments unless explicitly indicated.

4.1 Improvements with respect to the pyramidal 2D+t+2D scheme

Table 1 reports the average luminance PSNR for the interpolation-based pyramidal 2D+t+2D scheme of Fig. 4 in comparison with the proposed STP-Tool scheme (of Fig. 2). *Mobile Calendar* CIF sequences at 30fps are coded at 256 and 384 Kbps and predicted from a QCIF video coded at 128 Kbps (all headers and coded motion vectors included).

We also compare different configurations of STP-Tool in order to highlight its versatility: 1) STP-Tool prediction made only from the lowest temporal subband of the QCIF video (in this case, which results in the best case, only the 79 Kbps of the lowest temporal subband, without motion vectors, are extracted from the 128 Kbps coded QCIF; then 256-79=177 Kbps or 384-79=305 Kbps can be used for CIF resolution data); 2) like 1) but including all the QCIF sequence to enable multiple adaptations, i.e., the possibility to extract a maximum quality QCIF 30fps from each coded CIF video.

Fig. 7 shows an example of visual results at 384 Kbps. STP-Tool with multiple-adaptation disabled is compared against the interpolation-based ISP (also without multiple-adaptation). The latter scheme generates an overall more blurred image, and a visual quality gap with respect to our system is clearly visible.

SEQUENCE	FORMAT	BIT-RATE (KBPS)	PSNR_Y PYRAMIDAL	PSNR_Y STP-TOOL (MULT. ADAPT. DISABLED)	PSNR_Y STP-TOOL (MULT. ADAPT. ENABLED)
Mobile	CIF 30fps	256	23.85	27.62	26.51
		384	25.14	29.37	28.81

TABLE 1: PSNR COMPARISON AMONG DIFFERENT KINDS OF INTER-SALE PREDICTIONS.

4.2 PSNR comparisons with averaged video signal references

In Fig. 8, we present a complete PSNR comparison at QCIF resolution. PSNR values have been calculated with respect to the reference video sequence at QCIF resolution of each considered system. Unfortunately, each system differs in the way such references are calculated (MPEG downsampling filters for JSVM3.0, 9/7 wavelet filter bank for "t+2D" WVC configuration, and 3-LS wavelet filters [20] for STP-Tool configuration), and then only PSNR trends referred to a single system are meaningful but not absolute PSNR comparison among systems. In particular, due to the poor half-band selectivity of

(a) Original CIF30 (Mobile Calendar)

(b) 384 Kbps coded with STP-tool

(c) 384 Kbps coded with interpolation

FIGURE 7: VISUAL COMPARISON AT 384 KBPS ON MOBILE CALENDAR CIF 30FPS: (A) ORIGINAL FRAME CIF30 (MOBILE CALENDAR), (B) CODED AT 384 KBPS WITH THE STP-TOOL SCHEME OF FIGURE 2, (C) CODED AT 384 KBPS WITH THE INTERPOLATION PYRAMIDAL SCHEME OF FIGURE 4.

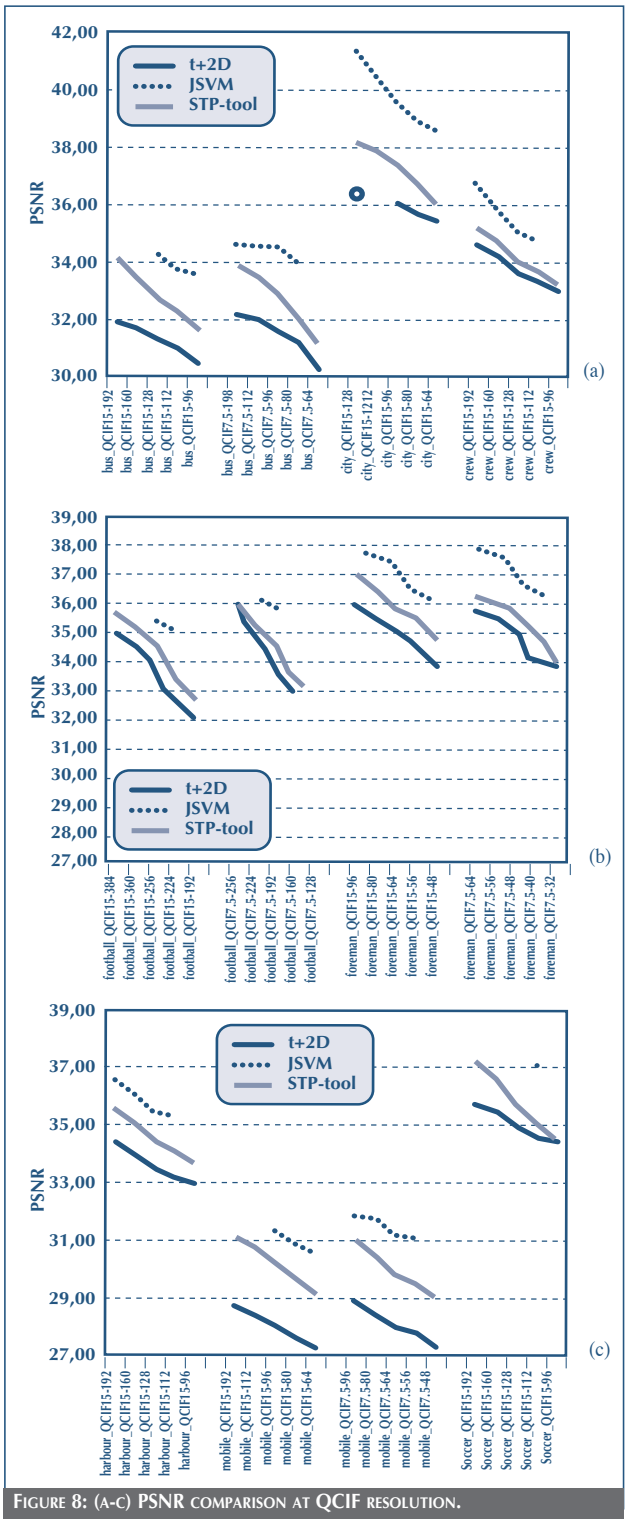

FIGURE 8: (A-C) PSNR COMPARISON AT QCIF RESOLUTION.

wavelet low pass filters, WVC system references are, in general, more detailed and contain more or less visible spatial aliasing. This determines lower PSNR values with respect to those measured with respect to a smoother reference. Therefore, PSNR differences among the three coding schemes lose significance when lower or intermediate resolutions are considered. In the following, we propose a possible solution to this problem which will allow us to "re-interpret" the results of Fig. 8.

A method to create a fair reference between two systems with their own reference video V_1 and V_2 is to create a common weighted reference $V=\alpha_1 V_1+\alpha_2 V_2$. In particular, by selecting $\alpha_1=\alpha_2=1/2$, it can be easily verified that $PSNR(V,V_1)=PSNR(V,V_2)$. This means that V_1 and V_2 are each equally disadvantaged by the creation of V. Moreover, signal V can be reasonably used as a common reference for PSNR comparisons of decoded sequences generated by different systems. In fact, even if a rigorously fair comparison of two very different coding systems with each one using its own reference signal could be seen as an extremely complex and multi-parametric problem to solve, a simple analysis uniquely based on signal spectral content gives us the possibility to confirm that the signal V can be used as a common reference, as stated above. Fig. 9 is a "projection" of the decoding process, which evidences the video frame's frequency content. V_1 is the reference video sequence of WVC system, while V_2 is the one of JSVM (MPEG downsampling generates smoother sequences than those generated by wavelet kernels;

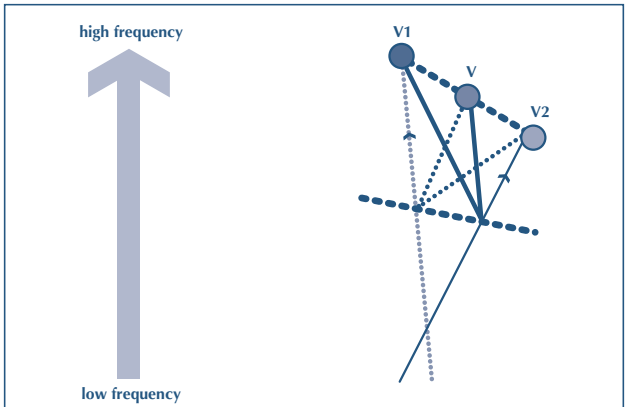

FIGURE 9: COMMON REFERENCE REPRESENTATION.

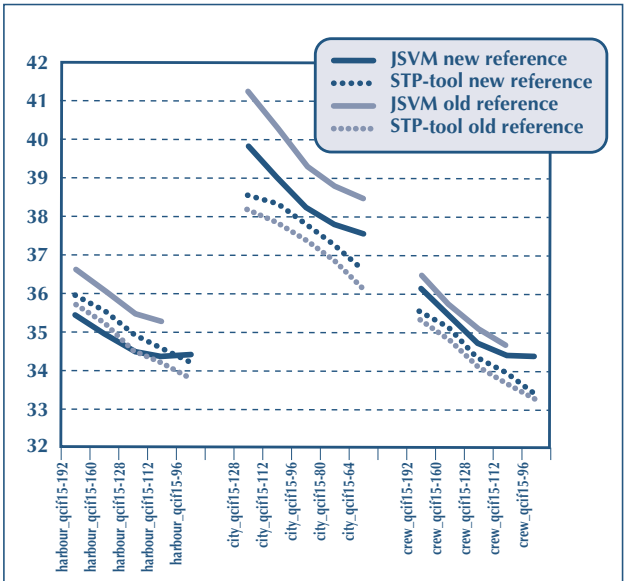

FIGURE 10: PSNR AT QCIF RESOLUTION: COMMON REFERENCE USAGE.

signal smoothing is also a strategy in AVC-H.264-based systems in order to reduce the visual impact of the artifacts related to the block-based DCT and motion model). From a spectral point of view, V can be considered halfway, its being a simple average.

As every transform-based coding system reconstructs the lower spectral part of pictures first, it is plausible, as shown in Fig. 9, that, at a certain bit-rate (represented by the dashed bold line), the WVC reconstructed signal is nearer to the JSVM reference than to its own reference. The converse is not possible, and V actually compensates for this disparity, making a fair comparison possible on a common reference.

In Fig. 10, we compare the PSNR results obtained on two sequences using both system-related references and a common reference for JSVM3 and STP-Tool. By using a common reference, STP-Tool and JSVM PSNR results are significantly closer. Similar results have been obtained at intermediate resolutions, i.e., for CIF sequences extracted from 4CIF (576x704) coded bit-streams. In any case, a certain difficulty in performing quantitative comparisons persists, and rigorous visual comparisons, with additional associated costs, should be made.

4.3 Visual comparisons

The following results are not meant to demonstrate superiority of one system over another but to show how, despite the fact

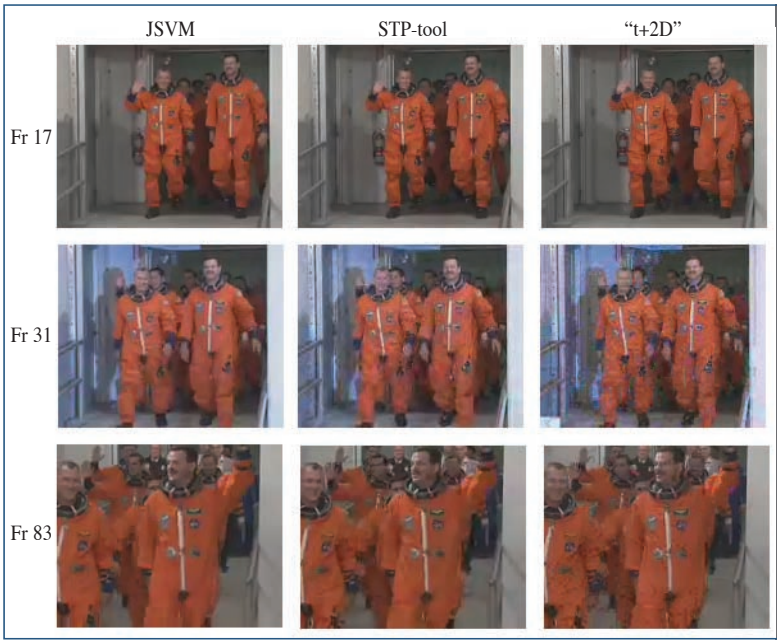

FIGURE 11: VISUAL COMPARISON ON CREW QCIF 15FPS 12 KBPS.

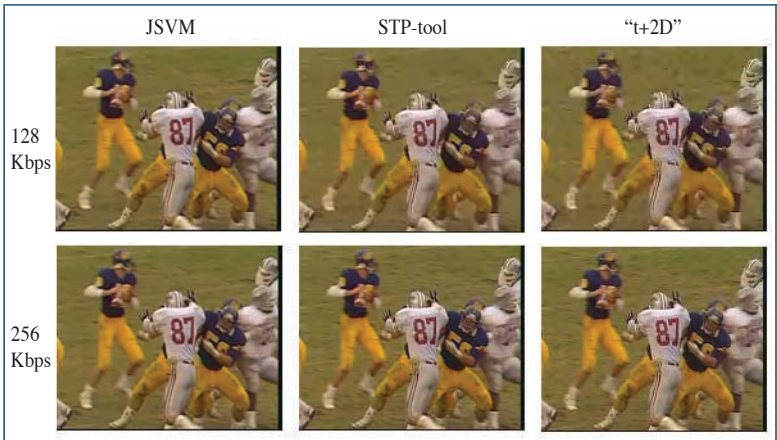

FIGURE 12: VISUAL COMPARISON ON FOOTBALL QCIF 7.5 (FRAME 17).

that JVT-SVC is based on mature technologies, the three systems are already close (in some cases very close) in terms of visual performance, with a superiority of STP-Tool over t+2D, especially for intermediate and lower resolutions. This closeness constitutes a motivation for finding new solutions to improve our STP-Tool system, especially in its peculiar or still unhandled aspects. The following results should then be interpreted as snapshots of systems which are currently in constant development.

In Fig. 11, some 15fps 128kpbs QCIF decoded frames of the *Crew* (originally 4CIF) sequence are displayed, and, in Fig. 12, a representative frame of the 7.5fps decoded *Football* (CIF) sequence is shown for two different bit-rates.

Fig. 13 shows a visual comparison on the *City* (4CIF) sequence at CIF intermediate resolution.

In Fig. 14, visual results on the sequence *Mobile* are presented at full CIF resolution. Decoding visual qualities are still very close. At the highest spatial resolution, the WVC reference software t+2D configuration outperforms most of the time JSVM3.0 in terms of PSNR, but PSNR as well as visual performance degrades at lower spatial resolution due to the analyzed problem of MCTF inversion at lower resolution.

At 4CIF resolution, the current STP-Tool implementation suffers to a greater extent from its still unoptimized blocks, especially for those concerning MV redundancies. In this case, visual performance remains inferior but still comparable with respect to the other schemes.

FIGURE 13: CITY_CIF15-192: (TOP) STP-TOOL (192 KBPS) MEAN PSNR 34.05DB, (MID) "T+2D" REF SW (195 KBPS) MEAN PSNR 33.43DB, (BOTTOM) JSVM3 (192 KBPS) MEAN PSNR 36.76DB


(a)


(b)


(c)

FIGURE 14: MOBILE_CIF30-384: (A) JSVM (384 KBPS) MEAN PSNR 31.04DB (B) STP-TOOL (384 KBPS) 29.63DB (C) "T+2D" (429 KBPS) 31.26DB

5. CONCLUSION

In this paper, we presented our wavelet-based 2D+t+2D SVC architecture called STP-Tool. We summarized the pros and cons of state of the art scalable wavelet video coding architectures and evidenced the potentialities of the STP-Tool solution in terms of flexibility and achievable performance. Then, we detailed some new advantageous architectural solutions which have been implemented into the wavelet video coding MPEG reference software. We showed that there is a noteworthy improvement of the 2D+t+2D STP-Tool approach with respect to the wavelet-based 2D+t+2D pyramidal architecture; then, we compared STP-Tool and t+2D configurations of the WVC reference software and the solution selected for SVC MPEG standardization, both in terms of objective and visual performance. Results can be summarized and interpreted as follows.

Because PSNR comparison with different video references at intermediate and lower spatial resolutions is unfair, we have worked on this problem by proposing a simple method to produce fairer PSNR comparison results. This method determined a sensible approach of JSVM and WVC STP-Tool PSNR values.

Visual performance of the considered schemes is quite close in many cases. This closeness constitutes a motivation for finding new solutions to improve our STP-Tool system, especially in its peculiar or still unhandled aspects. At lower spatial resolutions, WVC STP-Tool configuration shows competitive visual performance with respect to JSVM3.0. At the highest spatial resolution, however, STP-Tool does not perform comparably (especially for low bit-rate operating points) since in the current implementation, motion remains independently encoded for all different spatial resolution layers without exploiting the existing correlation between layer motions. A differential predictive scheme for motion vector coding across layers according to the STP-Tool architecture is under investigation. We are also working on another software system in which a tailored entropy coding solution is considered [31], thus allowing us to overcome some other WVC reference software issues that resulted in some inefficiencies in accommodating the STP-Tool solution.

REFERENCES

[1] ISO/IEC JTC1/SC29/WG11, "REQUIREMENTS AND APPLICATIONS FOR SCALABLE VIDEO CODING V.5," N6505, 69th MPEG Meeting, Redmond, WA, USA, July 2004.

[2] ISO/IEC JTC1/SC29/WG11, "REPORT OF THE SUBJECTIVE QUALITY EVALUATION FOR SVC CE1," N6736, 70th MPEG Meeting, Palma de Mallorca, Spain, Oct. 2004.

[3] N. Adami, M. Brescianini, R. Leonardi and A. Signoroni, "SVC CE1: STP-TOOL - A NATIVE SPATIALLY SCALABLE APPROACH TO SVC," ISO/IEC JTC1/SC29/WG11, M11368, 70th MPEG Meeting, Palma de Mallorca, Spain, Oct. 2004.

[4] N. Adami, M. Brescianini, R. Leonardi and A. Signoroni, "FULLY EMBEDDED ENTROPY CODING WITH ARBITRARY MULTIPLE ADAPTATION," ISO/IEC JTC1/SC29/WG11, M11378, 70th MPEG Meeting, Palma de Mallorca, Spain, Oct. 2004.

[5] ISO/IEC JTC1/SC29/WG11, "JOINT SCALABLE VIDEO MODEL (JSVM) 4.0 REFERENCE ENCODING ALGORITHM

DESCRIPTION," N7556, 74th MPEG Meeting, Nice, France, Oct. 2005.

[6] ISO/IEC JTC1/SC29/WG11, "WAVELET CODEC REFERENCE DOCUMENT AND SOFTWARE MANUAL," N7334, 73rd MPEG Meeting, Poznan, Poland, July 2005.

[7] N. Adami, M. Brescianini, R. Leonardi and A. Signoroni, "PERFORMANCE EVALUATION OF THE CURRENT WAVELET VIDEO CODING REFERENCE SOFTWARE," ISO/IEC JTC1/SC29/WG11, M12643, 74th MPEG Meeting, Nice, France, Oct. 2005.

[8] R. Leonardi, A. Signoroni and S. Brangoulo, "STATUS REPORT - VERSION 1 ON WAVELET VIDEO CODING EXPLORATION," ISO/IEC JTC1/SC29/WG11, N7822, 75th MPEG Meeting, Bangkok, Thailand, Jan. 2006.

[9] J.R. Ohm, "THREE-DIMENSIONAL SUBBAND CODING WITH MOTION COMPENSATION," IEEE Trans. Image Process., vol. 3, no. 5, pp. 559–571, Sept. 1994.

[10] S.-J. Choi and J.W. Woods, "MOTION-COMPENSATED 3-D SUBBAND CODING OF VIDEO," IEEE Trans. Image Process., vol. 8, no. 2, pp. 155–167, Feb. 1999.

[11] S.-T. Hsiang and J.W. Woods, "EMBEDDED VIDEO CODING USING INVERTIBLE MOTION COMPENSATED 3-D SUBBAND/WAVELET FILTER BANK," Signal Processing: Image Communication, vol. 16, pp. 705-724, May 2001.

[12] A. Secker and D. Taubman, "LIFTING-BASED INVERTIBLE MOTION ADAPTIVE TRANSFORM (LIMAT) FRAMEWORK FOR HIGHLY SCALABLE VIDEO COMPRESSION," IEEE Trans. Image Process., vol. 12, no. 12, pp. 1530-1542, Dec. 2003.

[13] V. Bottreau, M. Benetiere, B. Felts and B. Pesquet-Popescu, "A FULLY SCALABLE 3D SUBBAND VIDEO CODEC," in Proc. IEEE Int. Conf. on Image Processing (ICIP 2001), vol. 2, pp. 1017-1020, Oct. 2001.

[14] J. Xu, R. Xiong, B. Feng, G. Sullivan, M.-C. Lee, F. Wu and S. Li, "3-D SUBBAND VIDEO CODING USING BARBELL LIFTING", ISO/IEC JTC1/SC29/WG11, M10569/S05, 68th MPEG Meeting, Münich, Germany, Mar. 2004.

[15] Scalable Video Model 2.0, ISO/IEC JTC1/SC29/WG11, N6520, 69th MPEG Meeting, Redmond, WA, USA, Jul. 2004.

[16] Y. Andreopoulos, M. van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens and J. Cornelis, "COMPLETE-TO-OVERCOMPLETE DISCRETE WAVELET TRANSFORM FOR FULLY SCALABLE VIDEO CODING WITH MCTF," in Proc. Visual Comm. and Image Proc. 2003, SPIE vol. 5150, pp. 719-731, Lugano, Switzerland, July 2003.

[17] ISO/IEC JTC1/SC29/WG11, "SUBJECTIVE TEST RESULTS FOR THE CFP ON SCALABLE VIDEO CODING TECHNOLOGY," M10737, 68th MPEG Meeting, Münich, Germany, Mar. 2004.

[18] ISO/IEC JTC1/SC29/WG11, "JOINT SCALABLE VIDEO MODEL (JSVM) 3.0 REFERENCE ENCODING ALGORITHM DESCRIPTION," N7311, 73rd MPEG Meeting, Poznan, Poland, July 2005.

[19] ISO/IEC JTC1/SC29/WG11, "DESCRIPTION OF CORE EXPERIMENTS IN MPEG-21 SCALABLE VIDEO CODING," N6521, Redmond, WA, USA, July 2004.

[20] V. Bottreau, C. Guillemot, R. Ansari and E. Francois, "SVC CE5: SPATIAL TRANSFORM USING THREE LIFTING STEPS FILTERS," ISO/IEC JTC1/SC29/WG11, M11328, 70th MPEG Meeting, Palma de Mallorca, Spain, Oct. 2004.

[21] Y. Wu and J.W. Woods, "ALIASING REDUCTION FOR SCALABLE SUBBAND/WAVELET VIDEO CODING," M12376, 73rd MPEG Meeting, Poznan, Poland, July 2005.

[22] M. van der Schaar and D. Turaga, "UNCONSTRAINED MOTION COMPENSATED TEMPORAL FILTERING (UMCTF) FRAMEWORK FOR WAVELET VIDEO CODING," in Proc IEEE Int. Conf. Acoust. Speech and Signal Proc., pp. 81–84, , Hong-Kong, China, Apr. 2003.

[23] N. Mehrseresht and D. Taubman, "ADAPTIVELY WEIGHTED UPDATE STEPS IN MOTION COMPENSATED LIFTING BASED ON SCALABLE VIDEO COMPRESSION," in Proc Int. Conf. on Image Processing, Barcelona, Spain, Sept. 2003.

[24] D. Taubman, D. Maestroni, R. Mathew and S. Tubaro, "SVC CORE EXPERIMENT 1, DESCRIPTION OF UNSW CONTRIBUTION", ISO/IEC JTC1/SC29/WG11, M11441, 70th MPEG Meeting, Palma de Mallorca, Spain, Oct. 2004.

[25] D.S. Turaga, M. van der Schaar and B. Pesquet-Popescu, "COMPLEXITY SCALABLE MOTION COMPENSATED WAVELET VIDEO ENCODING", IEEE Trans. on Circuits and Syst. for Video. Technol., vol. 15, no. 8, pp. 982-993, Aug. 2005.

[26] ISO/IEC JTC1/SC29/WG11, "JOINT SCALABLE VIDEO MODEL (JSVM) 5.0," N7796, 76th MPEG Meeting, Montreux, Switzerland, April 2006.

[27] P.J. Burt and E.H. Adelson, "THE LAPLACIAN PYRAMID AS A COMPACT IMAGE CODE", IEEE Trans. on Communications, vol. 31, pp.532-540, Apr. 1983.

[28] J.-R. Ohm, "MULTIMEDIA COMMUNICATION TECHNOLOGY", Chapter 13, Springer-Verlag, 2004.

[29] N. Adami, M. Brescianini and R. Leonardi, "EDITED VERSION OF THE DOCUMENT SC 29 N7334," ISO/IEC JTC1/SC29/WG11, M12639, 74th MPEG Meeting, Nice, Oct. 2005.

[30] M. Wien, "JSVM4 BITSTREAMS FOR VIDWAV VISUAL EVALUATION," ISO/IEC JTC1/SC29/WG11, M13246, 76th MPEG Meeting, Montreux, Switzerland, April 2006.

[31] 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 Proc Visual Comm. and Image Proc. 2005, SPIE vol. 5960 (nr.58), Beijing, China, July 2005.

■ CONTACT: ST.JOURNAL@ST.COM ■