ISO/IEC JTC 1/SC 29/WG1N3954
July 7, 2006

ISO/IEC JTC 1/SC 29/WG 1
(ITU-T SG16)

Coding of Still Pictures

JBIG	**JPEG**
Joint Bi-level Image Experts Group	Joint Photographic Experts Group

TITLE: **Efficient wavelet-based video compression**

SOURCE: Nicola Adami*, Ebroul Izquierdo+, Riccardo Leonardi*, Marta Mrak+, Alberto Signoroni*, & Toni Zgaljic+

+AceMedia consortium[1](Motorola Ltd, Philips Electronics Nederland, Queen Mary University of London, Fraunhofer FIT, Universidad Autonoma de Madrid, Fratelli Alinari, Telefonica I+D, Dublin City University, Centre for Research and Technology - Hellas, INRIA, France Telecom, Belgavox, University of Koblenz - Landau)

*University of Brescia[2], Italy

PROJECT: JPEG 2000, Motion JPEG 2000

STATUS: Draft

REQUESTED ACTION: For discussion and feed-back

DISTRIBUTION: WG 1

INTERNATIONAL ORGANISATION FOR STANDARDISATION

[1] aceMedia is supported by the European Commission under contract FP6-001765.
Contact: Ebroul Izquierdo, ebroul.izquierdo@elec.qmul.ac.uk

[2] Telecommunications Group, DEA, via Branze, 38, 25123 Brescia (Italy).
Contact: Riccardo Leonardi, riccardo.leonardi@ing.unibs.it

ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG1
CODING OF STILL PICTURES

ISO/IEC JTC 1/SC 29/WG 1 N3954

Date: 2006-07-07

Title: Efficient wavelet-based video compression

SOURCE: Nicola Adami*, Ebroul Izquierdo+, Riccardo Leonardi*, Marta Mrak+, Alberto Signoroni*, & Toni Zgaljic+

+AceMedia consortium[3](Motorola Ltd, Philips Electronics Nederland, Queen Mary University of London, Fraunhofer FIT, Universidad Autonoma de Madrid, Fratelli Alinari, Telefonica I+D, Dublin City University, Centre for Research and Technology - Hellas, INRIA, France Telecom, Belgavox, University of Koblenz - Landau)

*University of Brescia[4], Italy

1. Introduction

It is a well known fact that exploiting temporal redundancy in video coding improves compression efficiency. Recent research results have shown that adopting a spatio-temporal multiresolution representation for video coding can represent a flexible base for Scalable Video Coding (SVC). In particular, wavelet-based video coding frameworks provide many attractive features.

Scalability is related to the possibility (in any time and system configuration) of having a direct access to the right amount of coded information (i.e. avoiding over-transmission or data format conversion or transcoding) in order to optimally access, communicate and use the desired video content with respect to the allowable transmission throughput and receiving device features. 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* (e.g. PSNR), *spatial and temporal resolution*s is usually required to optimally and dynamically adapt to the size of displaying terminals, to the related frame-rate reproduction capabilities and/or power saving (temporary or structural) needs as well as to the available throughput on communication networks, channels and distribution nodes.

[3] aceMedia is supported by the European Commission under contract FP6-001765.
Contact: Ebroul Izquierdo, ebroul.izquierdo@elec.qmul.ac.uk

[4] Telecommunications Group, DEA, via Branze, 38, 25123 Brescia (Italy).
Contact: Riccardo Leonardi, riccardo.leonardi@ing.unibs.it

This may turn out a natural evolution of the current JPEG 2000 standard which has already been or may be adopted for handling digital image sequences in a variety of contexts (D-Cinema, E-Cinema, HDTV, secure and efficient content distribution on heterogeneous networks and devices,…).

By adding temporal prediction, improved coding efficiency of video with scalability functionalities is likely to be established, thus leading to great broadening of the standard features. JPEG2000 compatibility can be granted on both intra-frame and residual information in a very natural way. The compatibility could be extended to the motion field coding part especially in consideration of possible innovative motion prediction and compensation scenarios based e.g. on dense (non block-based) motion vector fields (ideally more appropriate to be used in conjunction with non block based transform).

A scalable video codec typically consists of three modules: encoder, extractor and decoder, so that fractions of the bit-stream can be discarded without the need to decode even partially the compressed bit-stream.

Figure 1 shows a typical SVC system, referring to the coding of a video signal at an original resolution CIF (288 height, 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 referred 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”. In Figure 1 it is shown to be arranged between coder and decoder. According to the application field, it can be realised 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 Figure 1, a lower spatial resolution QCIF (144, 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 very low complexity of the extraction block that does not require coding/decoding operations and typically consists in simple “cut and paste” operations on the coded bit-stream.

Given the peculiarity of video signals in time, it is appropriate to use motion-compensated temporal filtering (MCTF) with an adaptive selection of wavelet filters. In the spatial domain an adaptive 2D wavelet transform can be applied. A brief survey of typical approaches is reported in the next section (for more details refer to [1]).

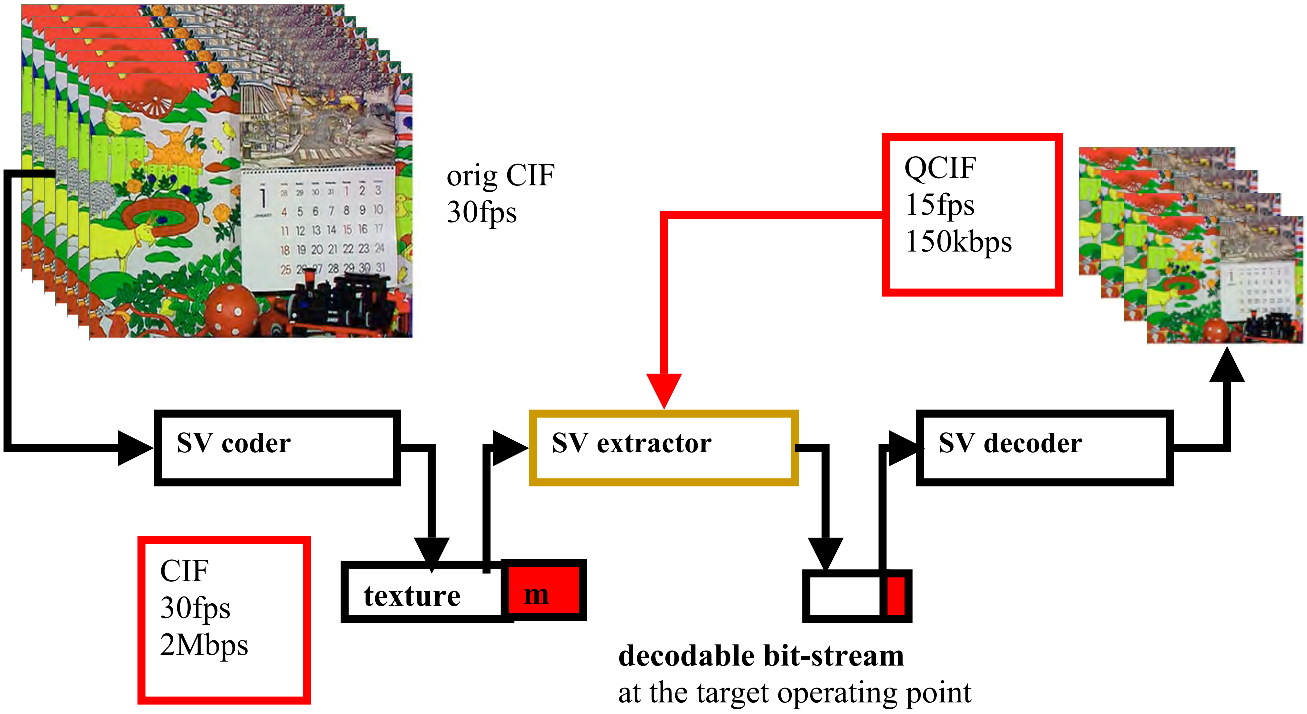


Figure 1. SVC rationale.

2. 3D wavelet based scalable video coding

Wavelet based Scalable Wavelet Video Coding (W-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 [2]. In the last years, stimulated by the advances in the field of wavelet video coding and by the work of a series of MPEG Ad-Hoc Groups on the subject, competitive (in terms of coding performance) fully scalable video coders have begun to appear. These are based on core technologies both inspired to highly optimized single-rate hybrid coding schemes (mainly MPEG-4 AVC/H.264) and to 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 [3] 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 (among which [4,5]) demonstrated similar performances manifesting their limits/potentials and their need of further maturation. Hence, in MPEG, the decision to start a new SVC standard [6] in conjunction with ITU-T (JVT-SVC) was made, considering wavelet based solutions for longer term objectives and applications [7-9]. After this meaningful, but in some aspects unnatural, comparison with respect to today predominant video coding technologies, wavelets based SVC research may 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) [10].

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, then coded with a progressive or quality scalable coding technique [11-15]. Current 3-D wavelet video coding schemes with Motion Compensated Temporal Filtering (MCTF) can be divided into three main categories. The first performs MCTF on the input video sequence directly in the full resolution spatial domain before spatial transform and is often referred to as spatial domain MCTF or "t+2D" (one example is [12]). The second performs MCTF in wavelet subband domain generated by spatial transform, being often referred to as in-band MCTF, or "2D+t" (one example is [16]). The third approach, called pyramidal or "2D+t+2D" (one example is [4]), performs first a spatial transform in order to extract lower resolution reference video signals and codes these signals according to a pyramid of "t+2D" decompositions and by using some Inter-Scale Prediction (ISP) mechanisms. Figure 2 is a general framework which can support the above schemes (a greater detail will be given on pyramidal solutions). Firstly, a pre-spatial decomposition can be applied to the input video sequence. Then a multi-level MCTF decomposes the video frames into several temporal subbands, such as temporal high-pass subbands and temporal low-pass subbands. After temporal decomposition, a post-spatial decomposition is applied to each temporal subband to further decompose the frames spatially.

Each scheme has evidenced its pros and cons [17, 3] 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 to compensate for the shift-variant nature of the wavelet transform (especially for 2D+t schemes)
- in the performance of ISP mechanisms (especially for 2D+t+2D schemes).

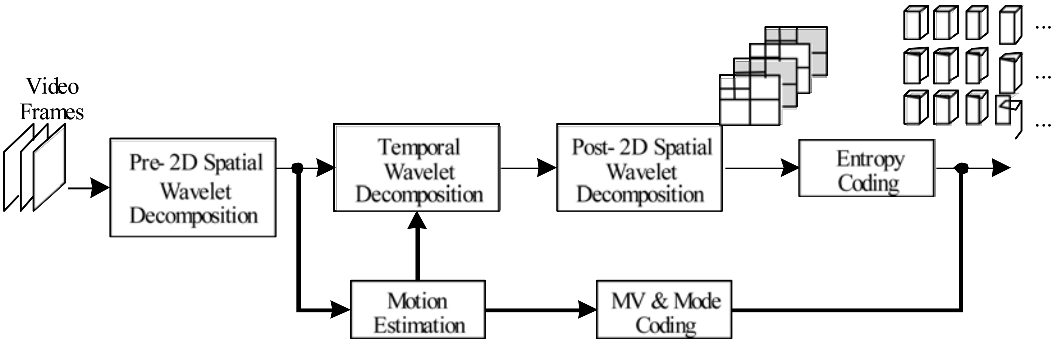

Video Frames
Pre- 2D Spatial Wavelet Decomposition
Temporal Wavelet Decomposition
Post- 2D Spatial Wavelet Decomposition
Entropy Coding
Motion Estimation
MV & Mode Coding
...
...
...

Figure 2: General coding Framework for 3-D wavelet video coding.

To not burden the discussion here, a deeper analysis of the differences between schemes is reported in Annex 1 of this document. Moreover, in Annex 2 and Annex 3 the STP-tool (Spatio-Temporal Prediction tool) W-SVC configuration, proposed by the University of Brescia, and aceSVC solution, implemented by the aceMedia consortium, are described respectively. Those systems are considered in Sec.4 for comparison with JPEG 2000.

3. High resolution applications

The adoption of JPEG 2000 for D-Cinema applications [18] offers a wide degrees of freedom in terms of supported feature such as: colour depth and models, spatial temporal and quality scalability, etc... Despite its potentiality, JPEG 2000 has been originally targeted to still image coding and therefore available tools need to be adapted and improved for video sequence encoding. In addition to the natural exigency to exploit inter frame motion compensated correlation, critical aspects that need to be investigated are related to the separate encoding of each frame, that may originate flickering artefacts [19] and do not make use of rate-distortion approaches to temporally control the rate and/or quality of the encoded signal [20,21].

Other D-Cinema related high-quality video demanding scenarios where full scalability is particularly interesting are E-cinema, HDTV, secure and efficient content distribution on heterogeneous networks and devices. In fact, full scalability can speed up the process of video post-production and/or facilitate and really modulate the mastering and distribution costs when different output formats and resolutions are involved in such processes, as it is the case for advanced and extended applications of D-Cinema.

The development of advanced W-SVC systems is of high relevance in the field of D-Cinema, for at least the following motivations (partially anticipated in [22]):

- Very high resolution makes them potentially easier to compress, in that it is expected that the data better match the smooth basis functions used in the coder. As a consequence, correlation can be expected to exist over a larger number of pixels, hence coders with larger footprint than 8x8 or 4x4 block transforms should be useful.
- High noise levels, e.g. grain noise for film origination and high sensor noise for electronic origination make the frames harder to code, and call into question the effectiveness of predictive coding and classical motion compensation.
- Much greater bit depth, 12 versus 8 bits, is needed by the fact that the motion picture is shown on a huge screen in a darkened auditorium, permitting the human visual system to adapt to low light levels. Film has a relatively huge dynamic range and motion pictures make good use of it. Bit-plane progressive wavelet quantization
- There is the need for long term constant quality with a control on the average bit rate only, or equivalently total file size for the compressed motion picture. This is different from the usual CBR or VBR coding where only a short buffer is being optimized for PSNR performance. In digital cinema, a very long buffer can be used that may hold the entire compressed movie.
- SVC systems based on block based motion models and block DCT transform does not guarantee, even at high bit-rates the absence of blocking artefacts. In addition such schemes tend to an excessive data smoothing at lower resolutions due to the choice of the subsampling filters. Such filters are oriented to reduce the entropic data content for better prediction and lower visual artefact impact for very low bit-rates applications.

In conclusion, starting from the experience in the MPEG exploratory activities on wavelet video coding, [1] we are persuaded that wavelet based coding techniques are particularly suited for fully scalable coding systems targeted to high definition and high decoding quality video content. Temporal correlation cannot be completely ignored or discarded. The need to reproduce, with high fidelity, fine details and even noise could weaken the hypotheses at the base of motion compensated temporal prediction and lead to reconsider the adequateness of block based compensation; nevertheless temporal correlation remains and alternative motion representation and coding models should thus be explored in accordance with suitable temporal prediction mechanisms. Motion compensated temporal filtering (MCTF) using wavelet kernels (prediction-update) lifting implementations, unconstrained MCTF (only the prediction stage implemented), hierarchical B-frame prediction are all possible technical solutions for temporal correlation exploitation within scalable systems.

The choice of the DCI consortium to select JPEG 2000 as the reference system for D-Cinema coding is justified by important technical, economical and implementation factors. JPEG 2000 is a mature standard solution, it fulfils DC format requirements, it does not introduce blocking artefacts and it is royalty free. This however does not prejudice or exclude any further improvement especially on aspects involving coding efficiency with preservation and fulfilment of scalability and complexity requirements. In fact, low delay and random frame access requirements, which are useful for postproduction editing phases and needs JPEG 2000 intra-frame coding, are not essential for the distribution stages. New commercial scenarios and advanced distribution of D-Cinema and related applications will benefit of fully scalable and highly efficient video coding solutions: heterogeneous storage and transmission systems as well as display and projection devices are interested by future and extended D-Cinema scenarios. The 4k content projected on cinema theatres should be easily scaled (without transcoding, i.e. simply cut from the original bit-stream) to 2k resolution for smaller theatres but it should be also further scaled (in terms of spatial, temporal, quality and colour depth) for HD or SD home theatre equipment or even for mobile display terminals. To satisfy at every level both high quality reproduction and sustainable complexity system demands advanced scalable video coding solutions, oriented to high quality and artefact free reproduction, must be explored. Wavelet based video technologies, despite their maturity could not be compared with advanced hybrid coders (e.g. AVC and JSVM) one, already offers very promising solution (e.g. those proposed by the MPEG VidWav exploration group). Market requirements for advanced D-Cinema distribution and state of art video coding technology fully justify the research on alternative solutions to intra-frame JPEG 2000 coding, such solutions being conceivable to be backward compatible with the JPEG 2000 standard itself.

4. Performance gain with respect to JPEG2000

To demonstrate functionalities and coding efficiency of scalable video coding the results of several tests are presented. The experiments have been designed to show quality scalability performance as well as temporal and spatial resolution scalabilities. A typical video coding scenarios that require high compression and adaptability of the video bit-stream is considered. From a bit-stream of once encoded sequence using SVC, different rate and resolution points have been decoded. Two test sequences have been used - "City" and "Crew", both of 4CIF resolution (704 × 576), 60 frames per second (Hz), sampling format 4:2:0 and 600 frames.

For JPEG 2000 coding and decoding Kakadu executables have been used [23] and scripts that perform encoding and decoding of all frames in a sequence to a specific rate point are derived from those available from [24]. The same bit-rate has been allocated for each frame at the specific rate point. The length of JPEG 2000 header has been discarded. Frames have been encoded targeting bit-rate points at original resolution. For decoding at lower resolution, JPEG 2000 transcoding has been used. Input compressed frames for transcoding correspond to the highest rate of one higher resolution. During the transcoding for lower resolution the required bit-rates have been targeted.

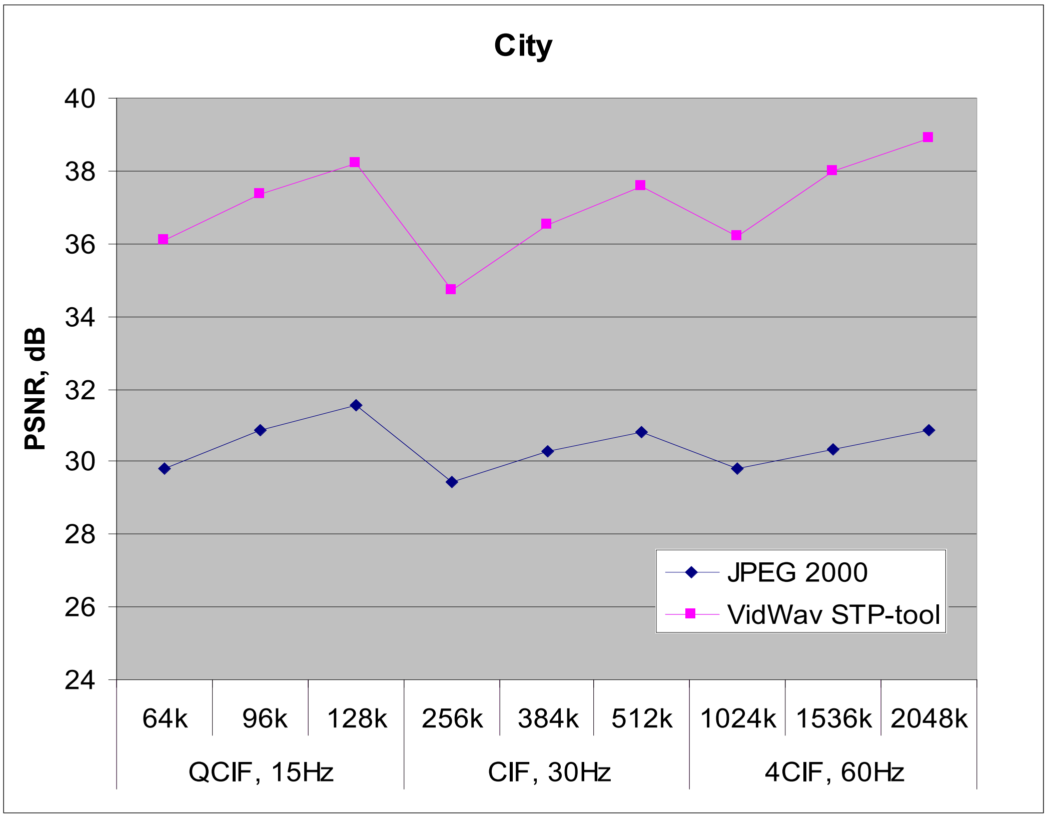
In Fig. 3, PSNR performances of the MPEG reference wavelet video coding exploration software (VidWav) in SPT-tool configuration (see Annex 2) and of JPEG 2000 are shown. In Fig. 4, PSNR performances of the aceSVC system (see Annex 3) and of JPEG 2000 are shown. Two different test sequences (City, Crew) have been considered in YUV format, and PSNR values are averaged on the three components according to the expression $PSNR=(4PSNR_Y+PSNR_U+PSNR_V)/6$. The reported results reflect the performance of W-SVC systems on which authors of this document were directly involved. Similar performance can however been reached by other W-SVC systems or configurations.

4.1 STP-tool

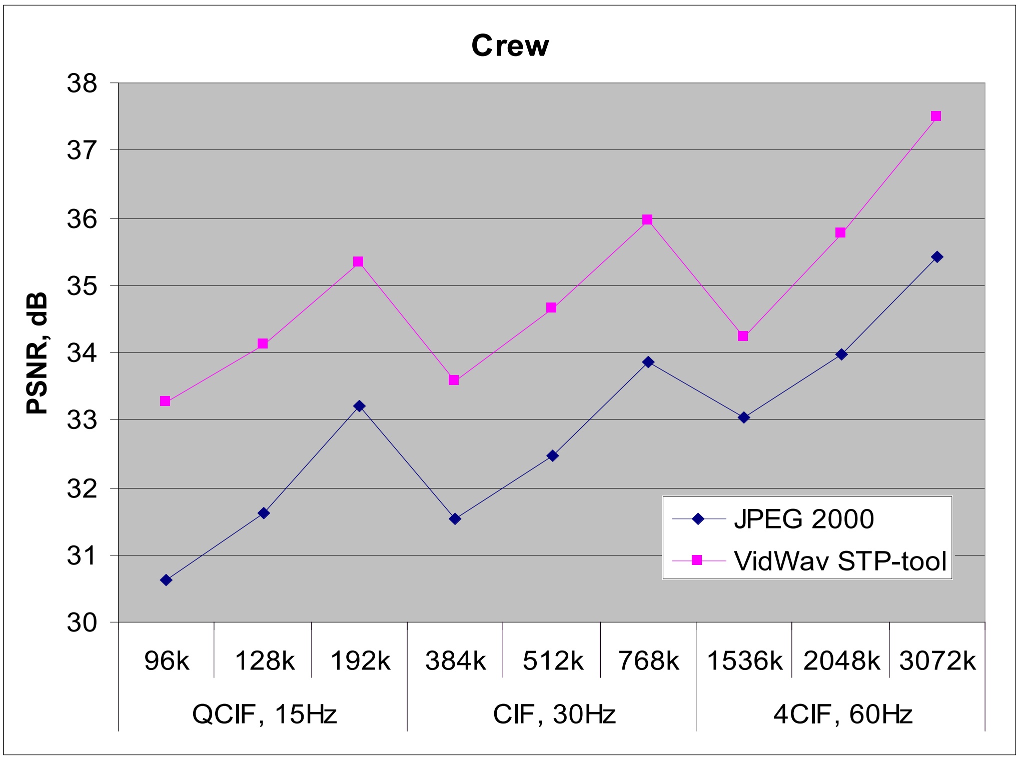
MPEG reference wavelet video coding exploration software (VidWav) in SPT-tool configuration (see Annex 2) results are taken from [25]

Modifications to the Wavelet Video Coding Reference Software [26] to accommodate the 2D+t+2D STP-tool interband prediction mechanism have been described in [27].

Coding gain for the two sequences are very different: 6.5dB in average for City and 2dB for Crew. This is because of the related nature of motion. In particular in the Crew sequences many discontinuities due to flash lamps are present and this have a negative impact on the system. However, a great part of the problems is due to a non-optimized handling of motion field discontinuities in the present version of the software, so it can be assumed that the PSNR gain could be sensibly improved even for this kind of irregular sequences (this can be also seen in the next subsection). Another thing that must be pointed out is that, at lower resolutions, reference video sequences for STP-tool system has been obtained by applying more selective wavelet filters than the 9/7 ones, in order to limit visual effects of spatial aliasing. Hence the CIF and QCIF resolutions reference sequences are not exactly the same between the two systems. This however could determine PSNR shifts of only some fraction of dB and does not impair our comparison (this issue is not present in the following system comparison).



(a)



(b)

Figure 3. Performance comparison: VidWav STP-tool, JPEG 2000

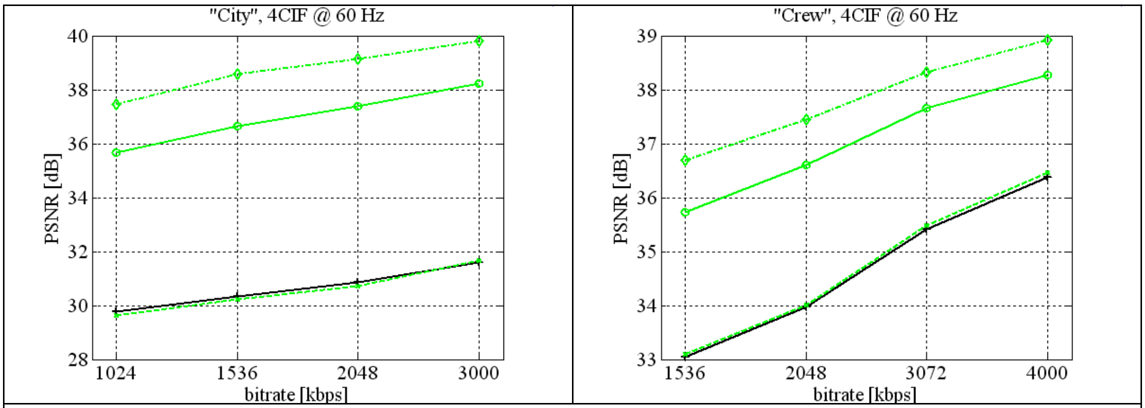
4.2 aceSVC

As W-SVC codec, aceSVC has been employed here. SVC results are compared to JPEG 2000 for the same sequences considered before.

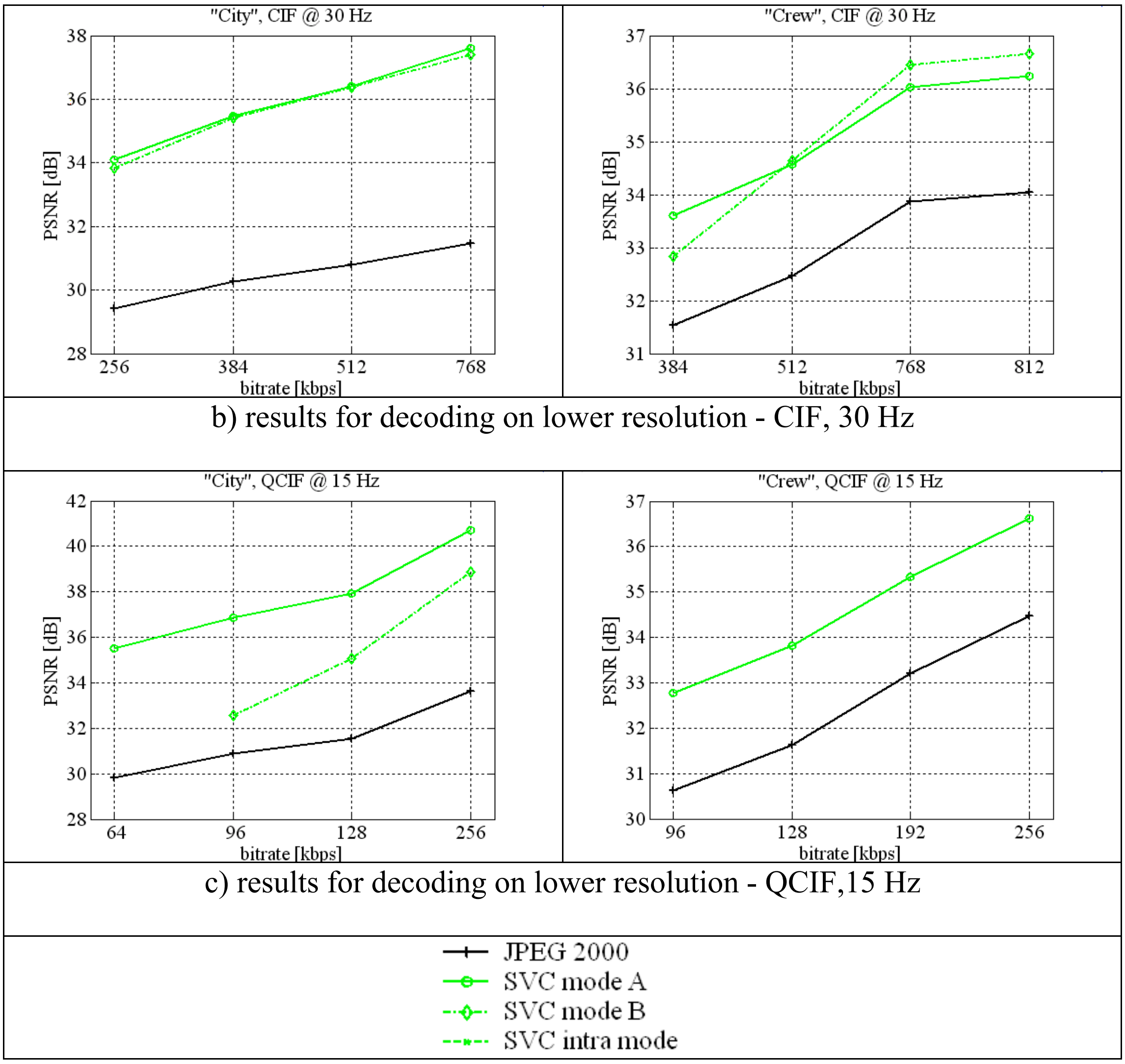
Although the main goal is to present compression gain that can be obtained using motion compensated prediction, two different wavelet based coders have firstly been compared for intra coding only. In Figure 4 results are presented for coding JPEG 2000 and SVC of single frames, i.e. for SVC in intra mode. Similarly as in JPEG 2000 settings, the same bit-rate has been allocated to each frame for SVC coding. It can be seen that both coders give similar results so it can be assumed that the coding gain shown in the following analysis is due to motion compensated prediction.

As scalable video coding allows for many different decomposition structures, in this experiment two different configurations have been used - t + 2D (named "SVC mode B") and a version of GSTS performing t + 2D + t + 2D + ... (named "SVC mode A"). See Annex 3 for further details. The main difference in those two configurations is that the first one provides better performance on the original resolution while the latest provides better performance on lower resolutions. In both schemes spatial transform employed uses 9/7 wavelet and temporal transform adaptively selects 5/3 or Haar wavelets without the update step.

All results are presented in terms of PSNR averaged over all three components. On original spatial resolution (Figure 4.a) it can be seen that application of motion compensation in wavelet based video coding system introduces in average 5 dB gain over for selected decoding points. For decoding on lower resolutions, SVC encoded sequences have only been adapted using simple bit-stream parsing, instead of transcoding. The results are summarised in Figure 4.b and 4.c. Here it can be observed that the performance relative to the encoding without motion compensation depends on chosen decomposition scheme. Some low rate points are not available for t + 2D decomposition scheme due to non-scalable motion coding used in this test. However, high quality can be reached using more sophisticated schemes based on GSTS (here "SVC mode A") which provide in average 3.8 dB gain for CIF resolution and 4.2 dB gain for QCIF resolution over intra only coding.



a) results for decoding on original spatio -temporal resolution - 4CIF, 60 Hz



b) results for decoding on lower resolution - CIF, 30 Hz

c) results for decoding on lower resolution - QCIF,15 Hz

Figure 4. Performance comparison: aceSVC, JPEG 2000

5. Conclusion

While Motion JPEG 2000 provides effective compression and easy access to individual video frames, as well as scalability functionalities, it prevents high compression of videos since it does not take into account temporal relations between frames. High compression is a requirement on video coding in all applications that need video delivery. Therefore by adopting motion compensation in JPEG standards the range of targeted application would be highly increased.

This contribution has attempted to report the recent advances of wavelet based video compression, that have lead to the video coding exploration effort in MPEG. Though MPEG has decided to discontinue the wavelet coding exploration activity, because of the current JVT effort for SVC, the still competitive performance of wavelet based video coding may lead to substantial benefits in JPEG, for which the targeted application domains are complementary, in particular since high resolution and high quality material is often a target objective. In addition, efficient processing may be possible on the compressed bit-stream to enable effective content description for search and retrieval.

6. References

[1] R. Leonardi, T. Oelbaum, J.-R- Ohm, "Status Report on Wavelet Video Coding Exploration " ISO/IEC JTC1/SC29/WG11, N8043, 76th MPEG Meeting, Montreux, Switzerland, Apr. 2006.

[2] ISO/IEC JTC1/SC29/WG11, "Requirements and Applications for Scalable Video Coding v.5," N6505, 69th MPEG Meeting, Redmond, WA, USA, July 2004.

[3] ISO/IEC JTC1/SC29/WG11, "Report of the Subjective Quality Evaluation for SVC CE1," N6736, 70th MPEG Meeting, Palma de Mallorca, Spain, Oct. 2004.

[4] N. Adami, M. Brescianini, R. Leonardi and A. Signoroni, "SVC CE1: STool - a native spatially scalable approach to SVC," ISO/IEC JTC1/SC29/WG11, M11368, 70th MPEG Meeting, Palma de Mallorca, Spain, Oct. 2004.

[5] 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.

[6] ISO/IEC JTC1/SC29/WG11, "Joint Scalable Video Model (JSVM) 4.0 Reference Encoding Algorithm Description," N7556, 74th MPEG Meeting, Nice, France, Oct. 2005.

[7] ISO/IEC JTC1/SC29/WG11, "Wavelet Codec Reference Document and Software Manual," *N7334*, 73rd MPEG Meeting, Poznan, Poland, July 2005.

[8] 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.

[9] 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.

[10] J.R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Trans. Image Process.*, vol. 3, no. 5, pp. 559–571, Sept. 1994.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[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] M.W. Marcellin and A. Bilgin, “JPEG2000 for Digital Cinema,” in Proc. of VidTrans and SMPTE Advanced Motion Imaging 2005, Atlanta, Georgia, 2005.

[19] A. Becker, W. Chan, D. Poulouin, "Flicker reduction in intraframe codecs", in Proc. of Data Compression Conference (DCC 2004), pp. 252-261, Utah (USA), 2004.

[20] J.C. Dagher, A. Bilgin and M.W. Marcellin, "Resource-constrained rate control for Motion JPEG2000," Image Processing, IEEE Transactions on , vol.12, no.12pp. 1522-1529, Dec. 2003.

[21] A. Ortega and K. Ramchandran, "Rate-distortion techniques in image and video compression," IEEE Signal Processing Magazine, vol. 15, no. 6, November 1998.

[22] J.-R. Ohm, M. van der Schaar, J.W. Woods, "Interframe wavelet coding.motion picture representation for universalscalability," Signal Processing: Image Communication, vol. 19, pp. 877-908, 2004.

[23] URL http://www.kakadusoftware.com/

[24] URL http://www.sprljan.com/nikola/matlab/jpeg2000.html

[25] 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.

[26] ISO/IEC JTC1/SC29/WG11, “Wavelet Codec Reference Document and Software Manual”, N7334, 73rd MPEG Meeting, Poznan, Poland, July 2005.

[27] N. Adami, M. Brescianini and R. Leonardi, “Edited version of the document SC 29 N 7334”, ISO/IEC JTC1/SC29/WG11, M12639, 74th MPEG Meeting, Nice, France, Oct.’05.

[28] 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.

[29] Y. Wu and J.W. Woods, “Aliasing reduction for scalable subband/wavelet video coding,” M12376, 73rd MPEG Meeting, Poznan, Poland, July 2005.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] ISO/IEC JTC1/SC29/WG11, “Joint Scalable Video Model (JSVM) 5.0,” N7796, 76th MPEG Meeting, Montreux, Switzerland, April 2006.

[35] T. Kimoto and Y. Miyamoto, “Multi-resolution motion compensated temporal filtering for 3d wavelet coding,” ISO/IEC JTC1/SC29/WD11 M10569/S09, March 2004, munich, Germany.

[36] G. Baud, M. Duvanel, J. Reichel, and F. Ziliani, “VisioWave scalable video CODEC proposal,” ISO/IEC JTC1/SC29/WG11M10569/S20, March 2004, munich, Germany.

[37] Y. Andreopoulos, A. Munteanu, J. Barbarien, M. van der Schaar, J. Cornelis, and P. Schelkens, “Inband motion compensated temporal filtering,” Signal Processing: Image Communication 19, pp. 653–673, August 2004.

[38] Y. Wang, S. Cui, , and J. E. Fowler, “3D video coding using redundant-wavelet multihypothesis and motioncompensated temporal filtering,” in Proceedings of the International Conference on Image Processing, 2, pp. 755–758, (Barcelona, Spain), September 2003.

[39] Davide Maestroni, Marco Tagliasacchi and Stefano Tubaro, In-band adaptive update step based on local content activity in Proc Visual Comm. and Image Proc. 2005, SPIE vol. 5960 (nr.19), Beijing, China, July 2005.

[40] Y. Wang, S. Cui, , and J. E. Fowler, “3D video coding using redundant-wavelet multihypothesis and motioncompensated temporal filtering,” in Proceedings of the International Conference on Image Processing, 2, pp. 755–758, (Barcelona, Spain), September 2003.

[41] H.-W. Park and H.-S. Kim, “Motion estimation using low-band-shift method for wavelet-based movingpicture coding,” IEEE Transactions on Image Processing 9, pp. 577–587, April 2000.

[42] J. C. Ye and M. van der Schaar, “Fully scalable 3-D overcomplete wavelet video coding using adaptive motion compensated temporal filtering,” in Visual Communications and Image Processing, T. Ebrahimi and T. Sikora, eds., pp. 1169–1180, Proc. SPIE 5150, (Lugano, Switzerland), July 2003.

[43] X. Li, “Scalable Video Compression via Overcomplete Motion Compensated Wavelet Coding”, Signal Processing: Image Communication, vol. 19, no. 7, pp. 637-651, 2004

[44] 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 Proc Visual Comm. and Image Proc. 2005, SPIE vol. 5960 (nr.38), Beijing, China, July 2005.

[45] 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.

[46] ISO/IEC JTC1/SC29/WG11, “Joint Scalable Video Model (JSVM) 4.0 Reference Encoding Algorithm Description,” N7556, 74th MPEG Meeting, Nice, France, Oct. 2005.

[47] 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.

[48] ISO/IEC JTC1/SC29/WG11, “Description of Core Experiments in MPEG-21 Scalable Video Coding,” N6521, Redmond, WA, USA, July 2004.

ANNEXES

Annex 1: W-SVC approaches

An analysis of the differences between W-SVC schemes is reported in the following.

A1.1. "t+2D"

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 fame-rate (partial versus complete MCTF inversion) and SNR quality; instead, if a lower spatial resolution version is needed, the inversion process discloses an incoherence with respect to the forward decomposition. The problem consists in 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 [28] or locally adaptive spectral shaping acting on the quantization parameters inside each spatial subband [29]. 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 artefacts that appears on the low pass temporal subbands when MC is not applied or when it fails due to unreliable motion vectors or to inadequate motion model. Such ghosting artefacts come 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) [30] which basically consists in 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 reducing temporal aliasing. Omitting it can cause visual coding performance worsening on reduced frame rate decoding, while as stated above, keeping it causes ghosting artefacts where the MC model fails. A solution that tries to adaptively weight the update step according to a motion field reliability model parameter has been proposed in [31, 32].

In the common motion compensated temporal filtering cases (e.g. with Haar or 5/3 kernels) an UMCTF approach actually lead to temporal open-loop versions of classical motion compensated (respectively uni- or bi-directional) 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. [33]). A closed loop version of UMCTF has been recalled "hierarchical B-frames prediction" and adopted in the latest version of JSVM [34].

A1.2. “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 MVs. Visual artefacts 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. Different coding systems have also been proposed [35–44] which are based on a 2D+t wavelet spatio-temporal decomposition.

A1.3. 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 (inter-scale prediction). What has been proposed to this end in the video coding literature is to use prediction between the lower resolution and the higher one 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. Figure 5 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 [45]. So, 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 is spread on a high number of coefficients and efficient encoding is hardly achievable. In the case of image coding, this drawback favoured 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 [46] falls in this pyramidal family in that prediction is made just after the temporal transform but only on intra (not temporally transformed) blocks.

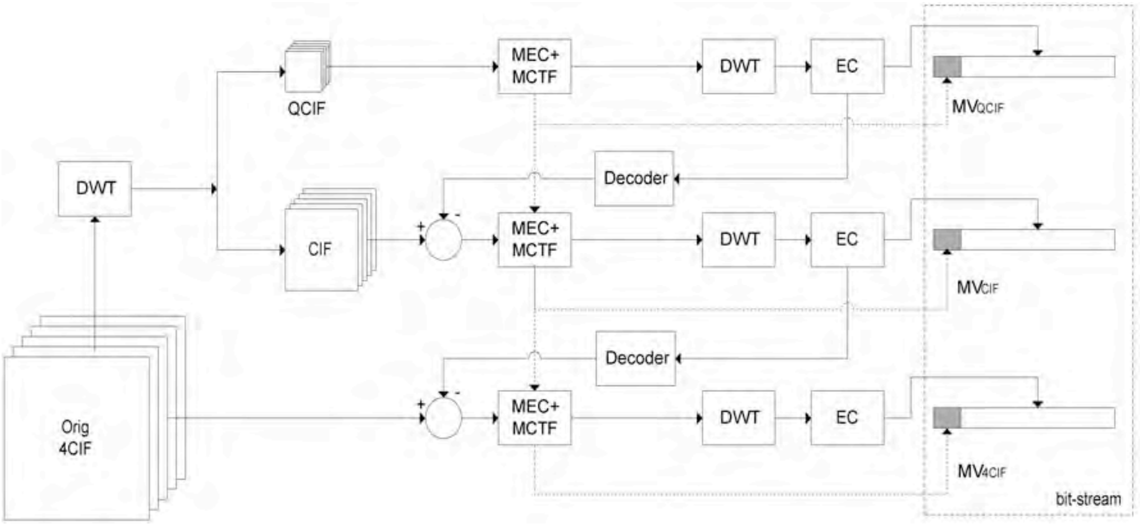


Figure 5. 2D+t+2D pyramidal scheme: ISP with interpolation.

Annex 2: STP-tool approach

As shown spatial scalability can be obtained 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 2D+t+2D pyramidal approach the ISP is obtained by means of data interpolation from level *s* to level *s*+1. The STP-tool idea [4,47] consists in 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, and multiple adaptation capabilities [48] 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.

A2.1 Closed-loop ISP STP-tool architecture

The closed-loop ISP STP-tool architecture is presented in Figure 6 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 Figure 6 (MEC stands for motion estimation and coding, T stands for spatial transform and EC stands for entropy coding, with coefficients 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 fs+1, starting from the decoded MCTF subbands at spatial level s, dec(fs);
2. rather than interpolating the decoded subbands, a single level spatial wavelet decomposition is applied to the portion of temporal subband frames fs+1 we want to predict. The prediction is then applied between dec(fs) (closed-loop STP-tool) and the low-pass (LL) component of the spatial wavelet decomposition, namely DWTL(fs+1). This has the advantage of feeding the quantization errors of dec(fs) 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 $\mathrm{DWT}_{\mathrm{L}}(f_{s+1})$ 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, a spatial decomposition followed by a temporal one in the second case (2D+t)). For the s+1 resolution, the prediction error $\Delta f_s = \mathrm{DWT}_{\mathrm{L}}(f_{s+1}) - \mathrm{dec}(f_s)$ is further coded instead of $\mathrm{DWT}_{\mathrm{L}}(f_{s+1})$.

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.

The STP-tool idea leads to valid alternative approaches with respect other W-SVC architectures. It efficiently introduces the idea of prediction between different spatial resolution levels within the framework of spatio-temporal wavelet transforms. Compared with the schemes described in Annex 1 it has several advantages. First of all, different spatial resolution levels both undergo a MCTF, and this prevents from the MC inversion incoherency of t+2D schemes. 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), and this bypasses the 2D+t schemes 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 the prediction signal at the lower resolution. So, there is a clear distinction between the coefficients that are interested in the prediction and the coefficients that are associated to 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.
The 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.

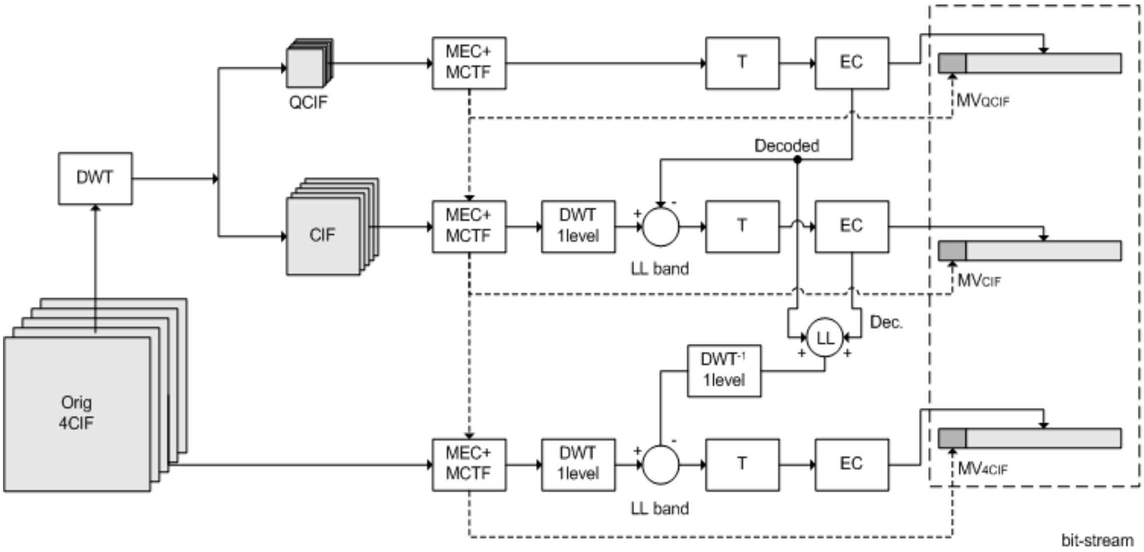


Figure 6. STP-tool coding architecture.

Annex 3: aceSVC approach

The architecture of aceSVC is based on the generalised spatio-temporal scalability (GSTS). This is the base for the consequential unified framework for spatio-temporal decomposition underpinning the targeted scalability functionalities.

The approach taken in the aceSVC implementation was to build the core of the codec with the GSTS as a target, allowing any decomposition path. The advantage of such design is twofold. Firstly, the support for two main architectures, namely t + 2D and 2D + t, is easily achieved as they represent special cases of decomposition paths. Moreover, with the implemented modular subband structure, other architectures not entirely based on the subband representation, i.e. 2D + t + 2D architecture, or a solution supporting a specific base layer, are feasible and the effort required for the necessary modifications becomes minimal. Secondly, with freedom in the selection of decomposition steps, the compression performance can be optimised with respect to the specified set of the decoding points. In other words, the idea is that subbands are decomposed only if it improves compression efficiency or if it is necessary for providing the requested decoding points.

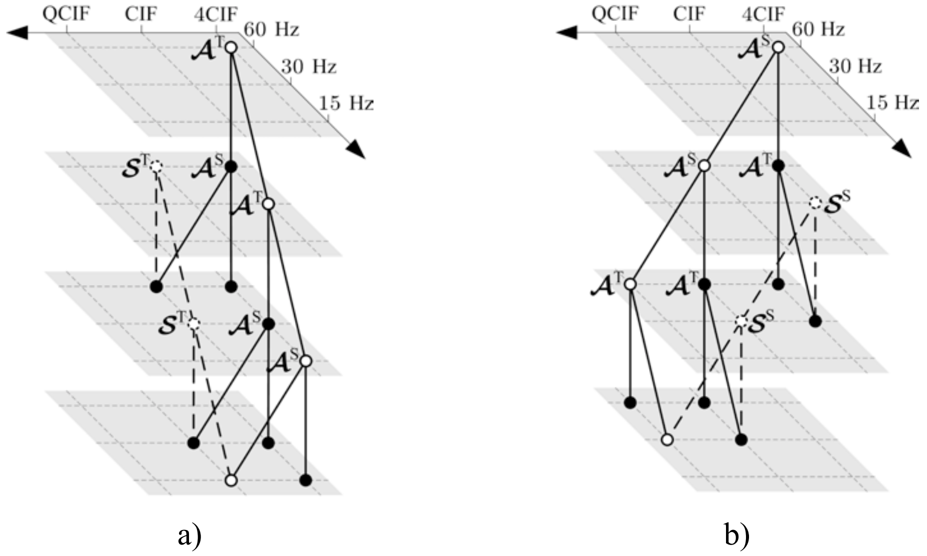

QCIF
CIF
4CIF
60 Hz
30 Hz
15 Hz
$\mathcal{A}^{\mathrm{T}}$
$\mathcal{S}^{\mathrm{T}}$
$\mathcal{A}^{\mathrm{S}}$
$\mathcal{A}^{\mathrm{T}}$
$\mathcal{S}^{\mathrm{T}}$
$\mathcal{A}^{\mathrm{S}}$
$\mathcal{A}^{\mathrm{S}}$
a)
QCIF
CIF
4CIF
60 Hz
30 Hz
15 Hz
$\mathcal{A}^{\mathrm{S}}$
$\mathcal{A}^{\mathrm{S}}$
$\mathcal{A}^{\mathrm{T}}$
$\mathcal{S}^{\mathrm{S}}$
$\mathcal{A}^{\mathrm{T}}$
$\mathcal{A}^{\mathrm{T}}$
$\mathcal{S}^{\mathrm{S}}$
b)

Figure 7. Two main types of open-loop SVC architectures for the example of source video of 4CIF resolution and 60 Hz frame rate: a) t + 2D b) 2D + t. White circles - nodes containing the low-pass spatio-temporal subband on the particular decomposition level. Dashed circles - nodes not present at the encoder. Black circles - nodes containing high-pass spatio-temporal subbands.

The main concepts introduced in the codec, that serves to support the GSTS, are the *spatio-temporal node* and the *decomposition tree*. The decomposition tree is composed of spatio-temporal nodes, each one representing a collection of subbands corresponding to the particular spatio-temporal resolution. An example illustrating this concept for t+2D and 2D+t architectures is shown in Figure 7, where A and Σ are the wavelet analysis and synthesis operators, while the temporal transform is denoted with T and the spatial with S. Each step of the decomposition involves critical subsampling, so that the number of coefficients in the children nodes is equal to the number of coefficients in the parent node. As it is shown in Fig. 7(a), in the t + 2D architecture the temporal decomposition is performed before the spatial one. In this particular example, where the original sequence is of 4CIF resolution and 60 Hz frame rate, a two level temporal decomposition is performed, producing the lowest temporal subband sequence corresponding to the frame rate of 15 Hz. In the subsequent step a spatial decomposition of one level is performed, leading to the low-pass spatial subband corresponding to the CIF resolution. By reconstructing the sequence, *i.e.*, by applying the inverse transform, a video sequence can be represented on any of the spatio-temporal decomposition nodes that were visited in the encoder (white circles). Moreover, by combining the available subbands a different decomposition path can be chosen (dashed lines), producing the decoding points that were not present at the encoder (dashed circles). Following the same principle, Fig. 7(b) shows the decomposition path for 2D + t scheme.

Temporal modules

In the aceSVC implementation of MCTF, the temporal wavelets of different lengths can be used. The chosen motion model is block based, with flexible macroblocks and block sizes, i.e. sizes of macroblock and minimal block sizes are derived from encoder input parameters. Such approach has been taken in order to support higher flexibility according to different requirements. Supported sub-pixel accuracy of motion is 1/32 and interpolation filters are based on *sinc* interpolation with different kernel sizes.

Spatial modules

Any of the specified wavelet filters can be used to perform spatial decomposition. For wavelets that are specified with lifting steps a new adaptive wavelet transform is supported. The adaptive transform uses the concept of connectivity-map to enable Motion Driven Adaptive Transform (MDAT). The connectivity-map describes the underlying irregular structure of a regularly sampled data. In the applied scheme the connectivity values are used in lifting as the weights applied to the neighbouring signal samples. The connectivity-map is generated from motion information, which is available at both the encoder and the decoder sides. Thus, no additional transmission overhead is introduced and separate coding of intra blocks is also realised.

Bitstream and adaptation

The aceSVC bit-stream is very simple, thus providing a fast adaptation. The bit-stream is divided into GOPs, where each GOP is composed of a GOP header and the atoms. An atom is a basic unit of adaptation, designed to be processed by some content-agnostic adaptation tool. An extractor module is used to adapt the bitstream to the given spatio-temporal resolution with fine-granular scalability. Each atom of the bitstream corresponds to a particular spatial, temporal and quality coordinate, and carries the data necessary for decoding at that point. This data can be bit-planes of the corresponding spatio-temporal subbands, and/or the motion vectors (in full quality or coded in a scalable way). Depending on the performed decomposition, a different number of atoms are created in a GOP. The example provided in Figure 8**Error! Reference source not found.** shows a resulting bit-stream when a complete spatio-temporal decomposition of T levels in temporal direction and S levels in spatial direction is performed.

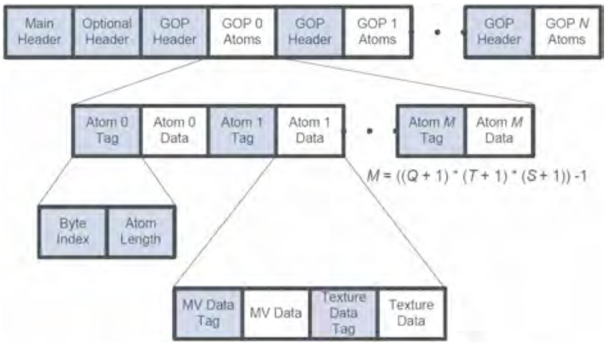


Figure 8. High level description of the aceSVC bitstream