INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11
MPEG2005/N7571
October 2005, Nice, F

Source	Video
Title	**Draft Status Report on Wavelet Video Coding Exploration**
Status	Approved
Sub group	Video Group
Authors	S. Brangoulo, R. Leonardi, M. Mrak, B. Pesquet Popescu, Jizheng Xu
	Contacts: Riccardo.Leonardi@ing.unibs.it, Beatrice.Pesquet@enst,fr

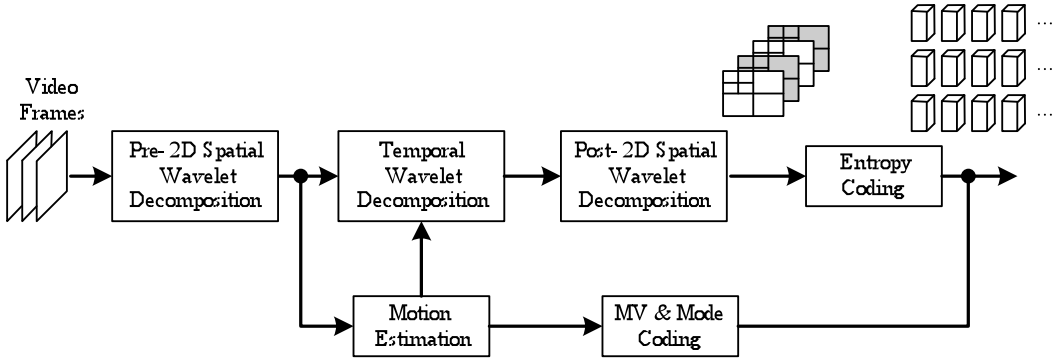
Update versions of this document will be discussed on the AhG group reflector (mpeg-vidwav@lists.rwth-aachen.de)

1 Video Coding with Wavelets

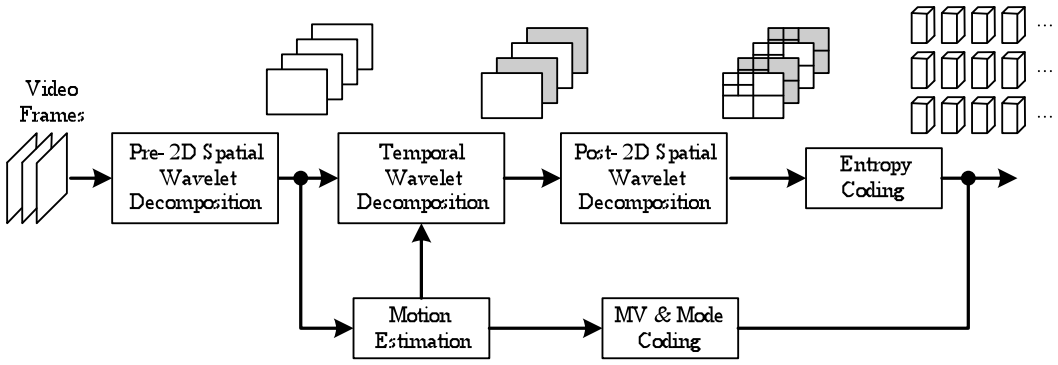
Current 3D wavelet video coding schemes with Motion Compensated Temporal Filtering (MCTF) can be divided into two main categories. The first performs MCTF on the input video sequence directly in the full resolution spatial domain before spatial transform and is often referred to as spatial domain MCTF. The second performs MCTF in wavelet subband domain generated by spatial transform, being often referred to as in-band MCTF. Figure 1(a) is a general framework which can support both of the above two schemes. 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 highpass subbands and temporal lowpass subbands. After temporal decomposition, a post-spatial decomposition is applied to each temporal subband to further decompose the frames spatially.

In the framework, the whole spatial decomposition operations for each temporal subband are separated into two parts: pre-spatial decomposition operations and post-spatial decomposition operations. The pre-spatial decomposition can be void for some schemes while non-empty for other schemes. Figure 1(b) shows the case of the T+2D scheme where pre-spatial decomposition is empty. Figure 1(c) shows the case of the 2D+T+2D scheme where pre-spatial decomposition is usually a multi-level dyadic wavelet transform. Depending on the

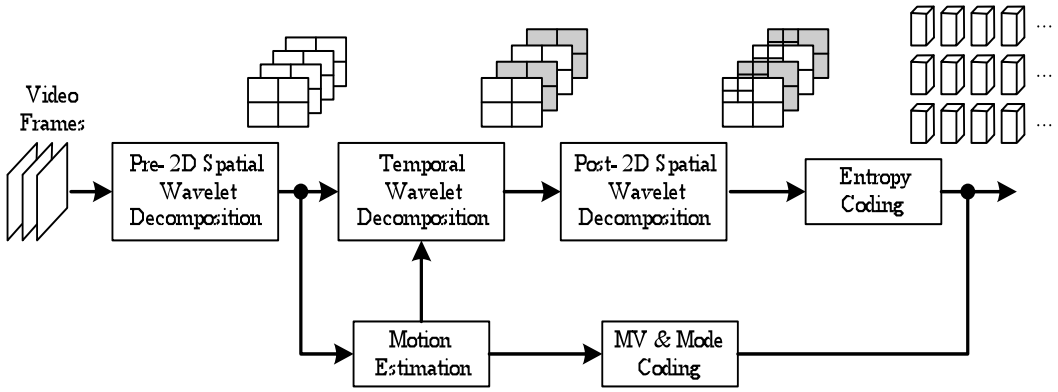
results of pre-spatial decomposition, the temporal decomposition should perform different MCTF operations, either in spatial domain or in subband domain.



(a) The general coding framework



(b) Case for the T+2D scheme (Pre-spatial decomposition is void)



(c) Case for the 2D+T+2D scheme (Pre-spatial decomposition exists)

Figure 1: Framework for 3D wavelet video coding.

A deep analysis on the difference between schemes is here reported.

A simple T+2D scheme acts on the video sequences by applying a temporal decomposition followed by a spatial transform. The main problem arising with this scheme is that the inverse temporal transform is performed on the lower spatial resolution temporal subbands by using the same (scaled) motion field obtained from the higher resolution sequence analysis. Because

of the non ideal decimation performed by the low-pass wavelet decomposition, a simply scaled motion field is, in general, not optimal for the low resolution level. This causes a loss in performance and, even if some means are being designed to obtain a better motion field, this is highly dependent on the working rate for the decoding process, and is thus difficult to estimate it in advance at the encoding stage. Furthermore, as the allowed bit-rate for the lower resolution format is generally very restrictive, it is not possible to add corrective measures at this level so as to compensate the problems due to inverse temporal transform.

In order to solve the problem of motion fields at different spatial levels a natural approach has been to consider a 2D+T scheme, where the spatial transform is applied before the temporal one. Unfortunately, this approach suffers from the shift-variant nature of wavelet decomposition, which leads to inefficiency in motion compensated temporal transforms on the spatial subbands. This problem has found a partial solution in schemes where the motion estimation and compensation take place in an overcomplete (shift-invariant) wavelet domain.

From the above discussion it comes 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, and thus both lower and higher resolution sequences must be MCTF filtered.

In this perspective, a possibility for obtaining good performance in terms of bit-rate and scalability is to use an Inter-Scale Prediction scheme. What has been proposed in the literature is to use prediction between the low resolution and the higher one before applying 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. This architecture has a clear basis on what have been the first hierarchical representation technique, introduced for images, namely the Laplacian pyramid. So, even if from an intuitive point of view the scheme seems to be well motivated, it has the typical disadvantage of overcomplete transforms, namely that of leading to a full size residual image. This way the information to be encoded as refinement is spread on a high number of coefficients and coding efficiency is hardly achievable.

A 2D+T+2D scheme that combines a layered representation with interband prediction in the MCTF domain appears now as a valid alternative approach. It efficiently combines the idea of prediction between different resolution levels within the framework of spatial and temporal wavelet transforms. Compared with the previous schemes it has several advantages. First of all, the different spatial resolution levels have all undergone an MCTF, which prevents the problems of T+2D schemes. Furthermore, the MCTF are applied before spatial DWT, which solves the problem of 2D+T schemes.

Moreover, the prediction is confined to the same number of transformed coefficients that exist in the lower resolution format. So, there is a clear distinction between the coefficients that are associated to differences in the low-pass bands of high resolution format with respect to the low resolution ones and the coefficients that are associated to higher resolution details. This constitutes an advantage between the prediction schemes based on interpolation in the original sequence domain. Another important advantage is that it is possible to decide which and how many temporal subbands to use in the prediction. So, one can for example discard the temporal high-pass subbands if when a good prediction cannot be achieved for such “quick” details. Alternatively this allows for example a QCIF sequence at 15 fps to be efficiently used as a base for prediction of a 30 fps CIF sequence.

A Scalable Video Coder (SVC) can be conceived according to different kinds of spatio-temporal decomposition structures which can be designed to produce a multiresolution spatio-temporal subband hierarchy which is then coded with a progressive or quality scalable coding technique [x-y]. A classification of SVC architectures has been suggested by the MPEG Ad-Hoc Group on SVC [x]. The so called t+2D schemes (one example is [x]) performs first 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 [x]), 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.

Each scheme has evidenced its pros and cons [x,y] in terms of coding performance. From a theoretical point of view, the critical aspects of the above SVC scheme 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 inter-scale prediction (ISP) mechanisms (especially for 2D+t+2D schemes).

2 A Comparison of Wavelet Based Video Coding Architectures in MPEG

The Exploration Video Wavelet Coding group has been considering three working modalities for wavelet video coding:

- A modified 2D+t+2D scheme presented in [x]
- A t+2D architecture as described in [x]
- A 2D+t architecture as described in [x]

2.1 The modified 2D+t+2D idea

Spatial scalability can be obtained by coding schemes where the lower spatial resolution information (at spatial level s) is used as a base-layer from which the finer resolution spatial level s+1 can be predicted. According to a common pyramidal approach [x,y] the inter-scale prediction (ISP) is obtained by means of data interpolation from level s to level s+1. The adopted idea [x] consists in performing an ISP where, by means of proper (e.g. reversible) spatial transforms, information is compared at the same spatial resolution (possibly after having be subjected to the same kind of spatio-temporal transformations). The deriving architectures are typically of the 2D+t+2D kind and ISP predictions take place without the need of data interpolation. These architectures can be designed to be fully space-time-quality scalable [x], and multiple adaptation capabilities [x] can be used without scarifying coding performance. In these architectures some critical issues that afflicts t+2D and 2D+t schemes are not present.

A main characteristic of the proposed (SNR-spatial-temporal) scalable video coding schemes is their native dyadic spatial scalability. Accordingly, this implies a spatial resolution driven complexity scalability. Spatial scalability is implemented within a scale-layered scheme (2D+t+2D). For example, in a 4CIF-CIF-QCIF spatial resolutions implementation three different coding-decoding chains are performed, as shown in Figure 2a (MEC stands for motion estimation and coding, T stands for spatial transform and EC stands for entropy

coding, with coefficients quantization included). Each chain operates at a different spatial level and presents temporal and SNR scalability. Being the information from different scale layers not independent of each other, it is possible to re-use the decoded (in a closed loop implementation) information (at a suitable quality) from a coarser spatial resolution (e.g. spatial level s) in order to predict a finer spatial resolution level s+1. This can be achieved in different ways. In the adopted approach, the prediction is performed 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). However, rather than interpolating the decoded subbands, a single level spatial wavelet decomposition is applied to the portion of temporal subband frames f_{s+1} one wants to predict. The prediction is then applied only between dec(f_s) and the low-pass (LL) component of the spatial wavelet decomposition, namely $\mathrm{dwt_L}(f_{s+1})$. 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 $\mathrm{dwt_L}(f_{s+1})$ 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, a spatial decomposition followed by a temporal one in the second case (2D+t)). For the s+1 resolution, the prediction error $\Delta f_s = \mathrm{dec}(f_s) - \mathrm{dwt_L}(f_{s+1})$ is further coded instead of $\mathrm{dwt_L}(f_{s+1})$ (see the related detail in Figure 2b). The question of whether the above predicted and predicting 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 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.

Starting from the proposed idea different kind of architectures can be envisaged. A main distinction can be made between open loop and closed loop solutions. In a purely closed loop scheme (the prediction signal is obtained from the decoded information) 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 (this is detailed in Fig. 1 for the prediction at the 4CIF level). In a purely open loop scheme the MCTF transformed signal at spatial resolution s is directly taken as the prediction signal, then prediction at spatial level s+1 only depends from the spatial level s. However, open loop schemes, especially at low bit-rates, undergo to the drift problems at the decoder side and then are nor further considered here.

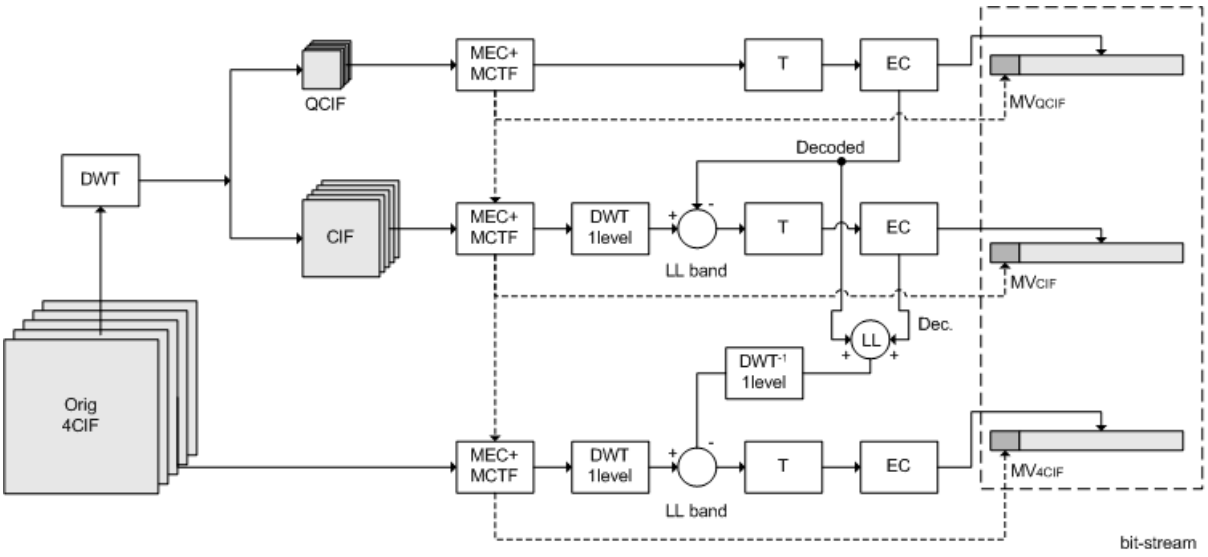


Figure 2a. 2D+t+2D coding architecture.

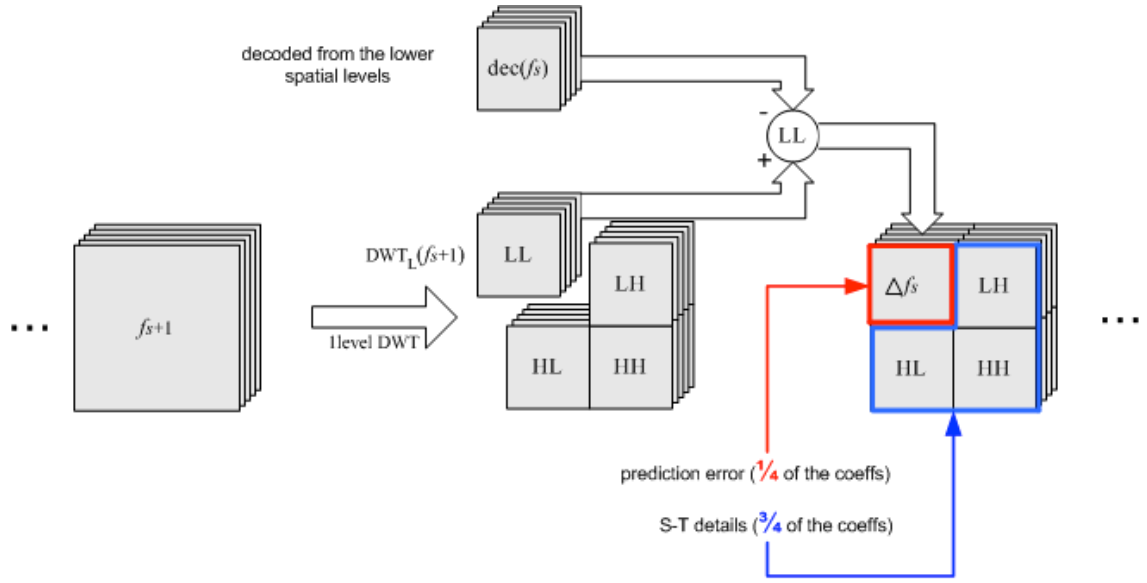


Figure 2b. 2D+t+2D prediction detail

2.2 2D+t+2D adopted architectures versus other SVC architectures

Next some insights are described about the differences between the proposed method and other existing techniques for hierarchical representation of video sequences. As explained in detail 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. The first 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 in different points with respect to the temporal filtering process. This implies that even prior to quantization, due to the shift variant nature of the motion compensation, these coefficients are in general

different. Thus, 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 before and after the temporal decomposition. This fact is of great importance in wavelet-based video coding scheme, because the differences between the $\mathrm{dec}(f_s)$ and $\mathrm{dwt}_L(f_{s+1})$ are responsible for a loss in performance in the t+2D schemes as explained hereafter.

2.2.1 T+2D

A deeper analysis of the differences between the 2D+t+2D scheme and the t+2D one reveals several advantages of the former one. A t+2D scheme acts on the video sequence by applying a temporal decomposition followed by a spatial transform. If the full spatial resolution is required, the process is reversed at the decoder to obtain the reconstructed sequence; if instead a lower resolution version is needed the inversion process differs in the fact that before the temporal inverse transform, the spatial inverse DWT is performed on a smaller number of resolution levels (higher resolution details are not used). The main problem arising with this scheme is that the inverse temporal transform is performed on the lower spatial resolution temporal subbands by using the same (scaled) motion field obtained in the higher resolution sequence analysis. Because of the non ideal decimation performed by the low-pass wavelet decomposition, a simply scaled motion field is, in general, not optimal for the low resolution level. This causes a loss in performance and even if some solutions can be conceived to obtain better motion fields (see for example [x]) these usually show dependencies from the operating point of the decoding process and then they are hardly optimally applicable during the encoding. Furthermore, as the allowed bit-rate for the lower resolution format is generally very restrictive, it is difficult to add corrections at this level so as to compensate the problems due to inverse temporal transform.

2.2.2 2D+t

In order to solve the problem of the motion fields scaling at different spatial levels an alternative 2D+t approach has been considered. In 2D+t schemes the spatial transform is applied before the temporal ones. Unfortunately, this approach suffers from the shift-variant nature of the wavelet decomposition, which leads to the inefficiency of motion compensated temporal transforms on 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 [x]. Motion field coherence among subbands and increased computational complexity are among the residual problems of this approach.

2.2.3 Pyramidal 2D+t+2D

From the above discussion it comes 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, and thus both higher and lower resolution sequences must be MCTF filtered. In this perspective, a possibility to 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 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 3 shows such an interpolation based inter-scale prediction scheme. The current reference model JSVM3 falls in this pyramidal family in that prediction is made just after the temporal transform but only on intra (not temporally transformed) blocks [x]. These architectures have got their basis in the first hierarchical representation technique introduced for images, namely the Laplacian pyramid [x]. 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 comes 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 which has already been shown to be problematic due to the relative inefficiency of motion estimation and compensation across the spatial subbands.

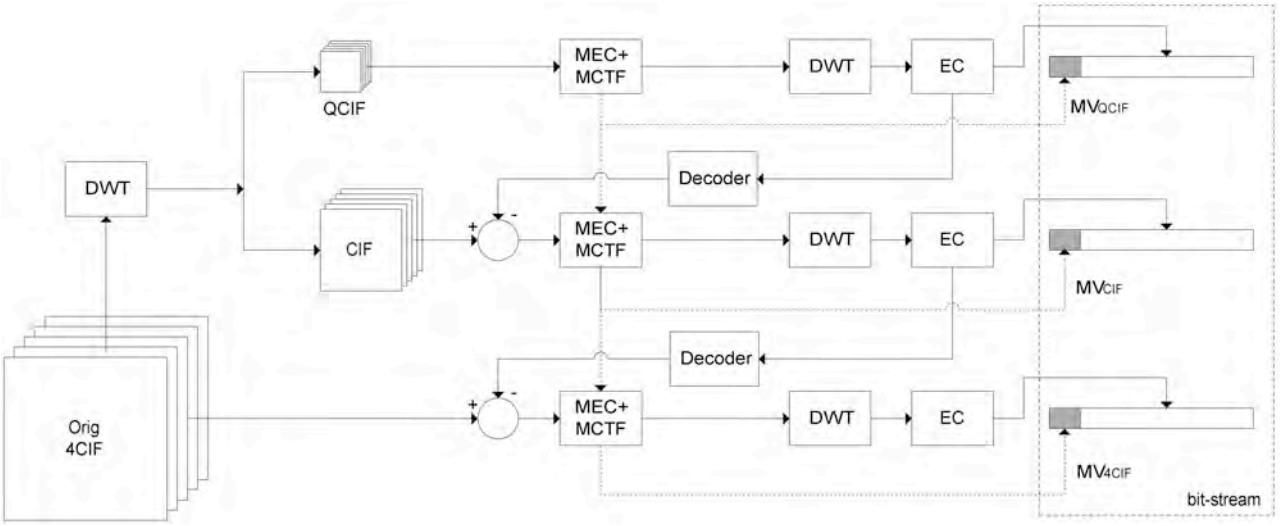


Figure 3. 2D+t+2D pyramidal scheme: prediction with interpolation.

2.2.4 Adopted 2D+t+2D architecture : « STool »

Looking at the above issues the so called STool idea leads to valid alternative approaches. It efficiently introduce 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 both undergone a MCTF, and this prevents from the problems of t+2D schemes. Furthermore, the MCTFs are applied before spatial DWT, and this bypasses the problems of 2D+t schemes. Moreover, 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. An STool architecture is highly flexible in that it permits several adaptations and additional features which in turns allow to improve the scalability and coding performance. These aspects will be considered in the following.

For a detailed presentation of the current STool based architectures, please see [1, 2, 3].

2.3 Latest improvements in 2D+t+2D architecture

2.3.1 AVC base-layer

STool is compatible with the use of an external base-layer bit-stream. We used the AVC base-layer functionality of the VidWav reference SW also 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 that the STool idea is somehow "robust" and can be used in a non strictly wavelet based coding environment.

2.3.2 Prediction on a subset of temporal frames

The STool prediction can be limited on a subset of the MCTF subbands, while the remaining subbands are directly coded. Figure 4 shows an example of MCTF decomposition on CIF and QCIF sequences and indicates, by the line-dot rectangle, that only the (0,1) subbands are involved in STool prediction instead of the whole group (3,2,1,0). The selection criterion can be empirical or computational (data content based or R-D based). It is also important to note that the above degree of freedom is not allowed for data domain prediction schemes (as the scheme o Fig.3 based on interpolation). At the time, we explored several empirical solutions and remarked that a coding gain can be obtained by using this degree of freedom.

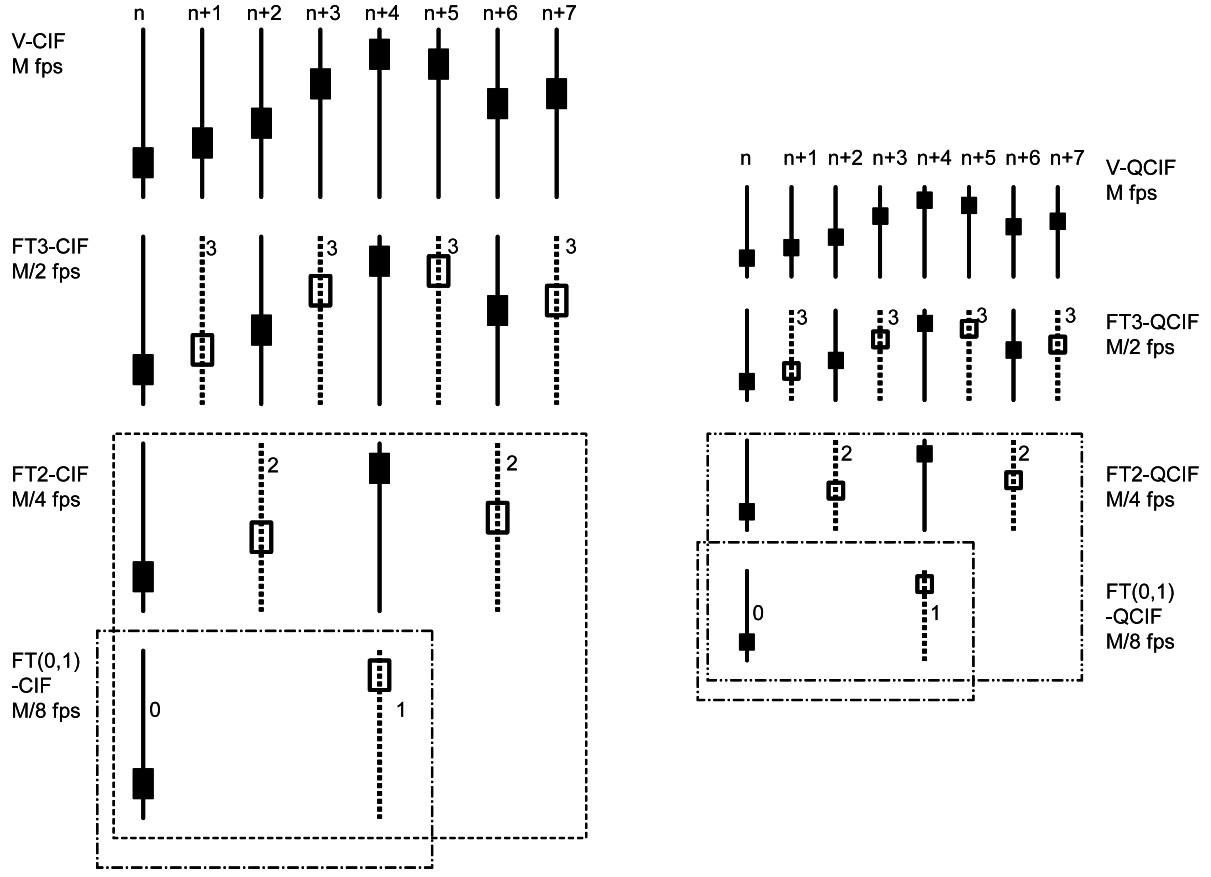


Figure 4. Possible variations of the STool prediction on the MCTF subband hierarchy

2.3.3 Prediction on an adapted temporal decomposition depth

Another degree of freedom in using the STool prediction mechanism consists in adapting the temporal decomposition level at which the prediction take place according to the target temporal resolution. It may happens that the temporal decomposition depth or the target frame

rate is not the same for all spatial resolutions in a prediction pyramid. In likely applications higher resolutions are associated to higher frame rate reproductions. In the example of Figure 4, two temporal decomposition 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 apply the STool prediction in two opposite ways (and other halfway ones):

1. execute a 3 level temporal decomposition for the CIF video in order to be able to perform a STool prediction adapted to the needed temporal decomposition depth at QCIF level (in Fig. 4 the dashed rectangle contains the additional subbands),
2. temporal transform the reference videos according to their needed levels (e.g. 1 for CIF and 3 for QCIF) and in order to perform the STool prediction partially inverse the overmuch levels (in Fig. 4 the double-dot line rectangle contains the inverted subbands in our example).

We tested both the solution and remarked that the second one usually performs better in that prediction on low-pass temporal subbands is more appropriate.

2.3.4 Asymmetric closed loop STool prediction

Another degree of freedom that we have in implementing a STool SVC architecture is the possibility to use 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. The idea is depicted in Fig.5 where for clearness only two spatial levels are considered. The coded base layer bit-stream can be entirely used (until the maximum of its assigned dimension, D_{max}) for base-layer video reconstruction. The ordinary closed loop STool mechanism consist in using, at the encoder site, a bitstream portion D_P, 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 site in order to update the prediction error decoded data. Instead, an asymmetry in this mechanism actually permit the use of a sub-portion D_A of the portion D_P for updating the prediction (causing s_r' to differ form s_r). Keeping the (D_P-D_A) spread limited within certain limits, and considering the fact that a target extraction of a higher resolution operating point undergo a $D=D_A+D_S$ target dimension, a coding gain can be achieved by exploiting this asymmetry. In general, the decision about a suitable value to assign to D_A with respect to a constraint D or with respect to an entire extraction path can be inserted into the extractor or otherwise distributed over the coding-decoding chain and can be realized by means of heuristic or tabular rules (without requiring complex calculations) or with computational methods (R-D optimization). Moreover the asymmetric closed loo approach can be easily extended to the case of more than two spatial levels. At the time all our tests concern heuristic D_A choices and are intended to illustrate the coding gain opportunity.

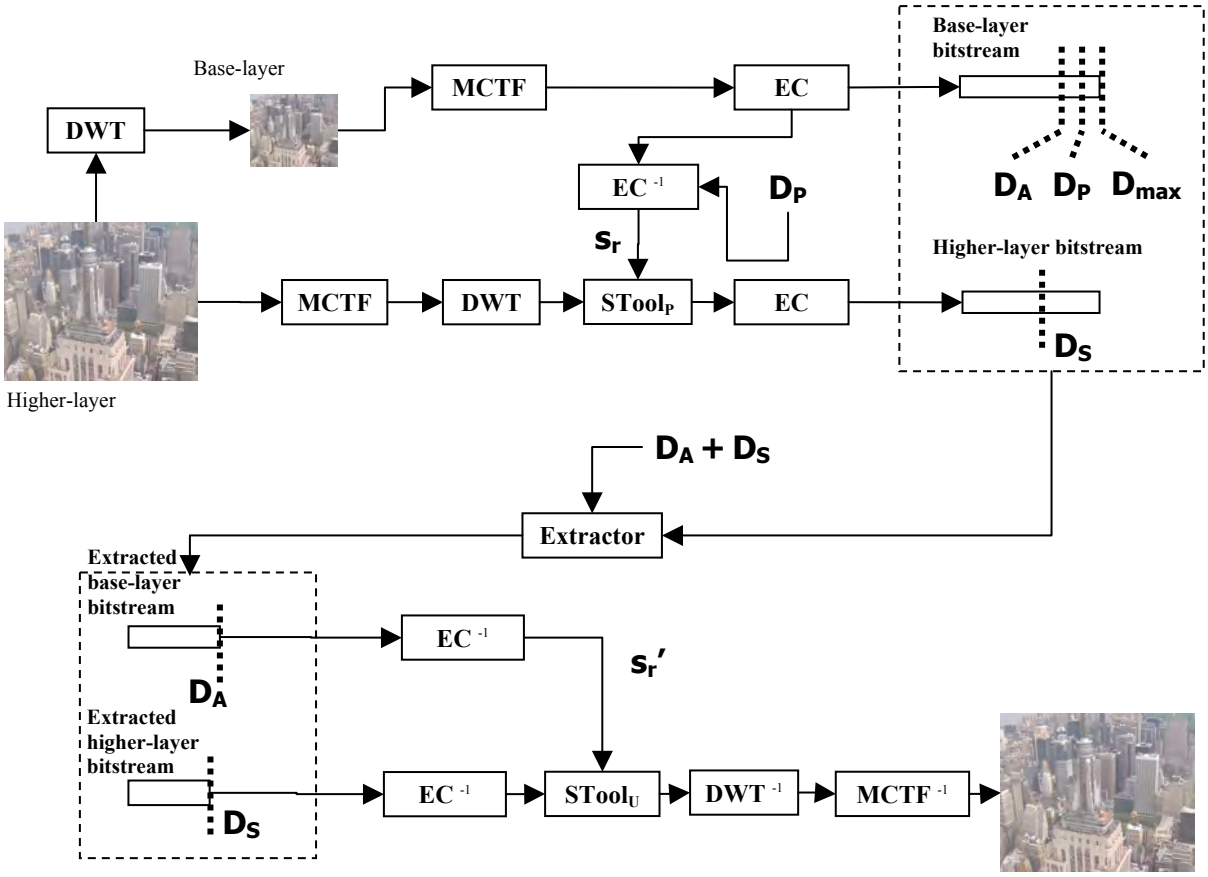


Figure 5. Asymmetric closed loop STool prediction

2.4 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 Figure 3 in comparison with the scheme presented in Figure 2. *Mobile Calendar* CIF sequences at 30fps are coded at 256 and 384kbps and predicted from a QCIF video coded at 128kbps (all headers and coded motion vectors included). We also compare different configurations of the STool architecture in order to highlight its versatility: 1) STool prediction made only from the lowest temporal subband of the QCIF video (in this case, which results to be the best case, only the 79kbps of the lowest temporal subband, without motion vectors, are extracted from the 128kbps coded QCIF, then 256-79=177kbps or 384-79=305kbps can be used for CIF resolution data); 2) like 1) but including all the QCIF sequence to enable multiple adaptations, i.e. extraction of a maximum quality QCIF 30fps from each coded CIF video.

Table 1. PSNR comparison among different kind of inter-scale predictions

Sequence	Format	Bitrate (kbps)	PSNR_Y pyramidal	PSNR_Y STool (mult. adapt. disabled)	PSNR_Y STool (mult. adapt. enabled)
Mobile	CIF 30fps	256	23.85	27.62	26.51
		384	25.14	29.37	28.81

Figure 6 shows an example of visual results at 384 Kbps. The STool with multiple adaptation disabled case is compared against the interpolation based ISP (also without multiple

adaptation). The latter scheme generates an overall more blurred image, and the visual quality gap with respect to our system is clearly visible.

(a) Original CIF30 (Mobile Calendar)

(b) 384kbps coded with STool prediction

(c) 384kbps coded with interpolation

Figure 6. Visual comparison at 384kbps on Mobile Calendar CIF 30fps: (a) original frame CIF30 (Mobile Calendar), (b) coded at 384kbps with the STool scheme of Figure 2, (c) coded at 384kbps with the interpolation pyramidal scheme of Figure 3.

2.5 Improvements with respect to the 70th meeting, Palma (10/2004)

One year improvement of the STool and of the JSVM schemes on the lower resolution. We compare today results (current document and [x] respectively) with the results presented at the MPEG Palma Meeting in Oct.2004 ([x] System 1 based on the MSRA SVC software and HHI SVC proposal and software respectively). In Tab. 2 we calculated, for each test sequence, a PSNR measure which is the average PSNR on the whole set of QCIF multiple extracted Palma points allowable for each sequence. PSNR are calculated with respect to each system reference i.e. 3-LS filtered and MPEG downsampling filtered sequences respectively. The PSNR improvements (difference) are free from the bias related to the different reference sequence.

Sequence	PSNR Palma Stool	PSNR Palma JSVM	PSNR Nice Stool	PSNR Nice JSVM	Difference Stool
Bus	31,49	33,96	32,34	34,02	0,85
Foreman	33,46	36,52	35,17	36,64	1,71
Football	32,23	35,91	33,94	36,04	1,71
Mobile	27,45	30,83	29,77	30,89	2,32
Harbour	34,69	36,06	34,73	36,06	0,04
City	37,07	38,92	37,23	39,73	0,16
Soccer	35,66	36,71	35,89	37,02	0,23
Crew	34,09	35,86	34,24	35,84	0,15

Table 2: PSNR improvements on the QCIF resolution

3 Tailored Wavelet Video Coding applications and functionalities

Wavelet video coding appears promising for applications:

1. targeting storage of high definition content (no delay constraint), with non predefined scalability range (e.g., Digital Cinema)
2. targeting a very high number of spatio-temporal decomposition levels (e.g. surveillance)
3. targeting non dyadic decompositions (e.g., video editing, conversion format between SD and HDTV, ...)
4. targeting fast moving region of interest tracking over time (e.g., surveillance)

Wavelet video coding may also provide advantages for

1. Multiple Description Coding which would lead to better error-resilience (e.g., wireless broadcasting)
2. easily providing means to optimally prioritize temporal versus spatial information for fast decoding purposes (surveillance, video browsing)
3. extremely fine grain SNR scalability (naturally implemented given the multiresolution framework enabled by wavelet representation).
4. enabling efficient similarity search in large video databases (fast indexing of multimedia documents for e.g. browsing, information retrieval)
5. better R(D) performance for very high resolution material (HD, DC, medical imaging), since it naturally will naturally deal with redundancy at various scales.
6. “better” compatibility with J2K and MJ2K.

4 Performance evaluation

4.1 Quality assessment in a scalable video coding framework

4.1.1 Problem statement

4.1.2 Objective measures

A method to create a fair reference between two system which use their own reference video V_1 and V_2 is to create a weighted reference $V=\alpha_1 V_1+\alpha_2 V_2$, and in particular with $\alpha_1=\alpha_2=1/2$ it can be easily verified that $PSNR(V,V_1)=PSNR(V,V_2)$. This means that V_1 and V_2 are equally disadvantaged by the creation of the common reference V, and then V is a fair common reference for both.

4.1.3 Visual tests

4.2 Latest performance results

A comparison among the decoded sequences by JSVM3.0, VidWav reference software in "t+2D" working condition and AVC base-layer (with optimal configuration files, provided by MSRA) and Vidwav reference software in "2D+t+2D" working condition (with configuration as described in the m12642 document [x]) is reported. All the points have been extracted following the Palma extraction path and the bitstream size have been verified: VidWav reference SW in "t+2D" configuration and JSVM3.0 do not respect the bit-rate constraint in all the sequences.

All PSNR results are reported in the excel file attached to document m12643 [x].

NOTE: it was not possible to correctly extract some JSVM3 working points with the available configuration files.

4.2.1 Lowest spatial resolution results

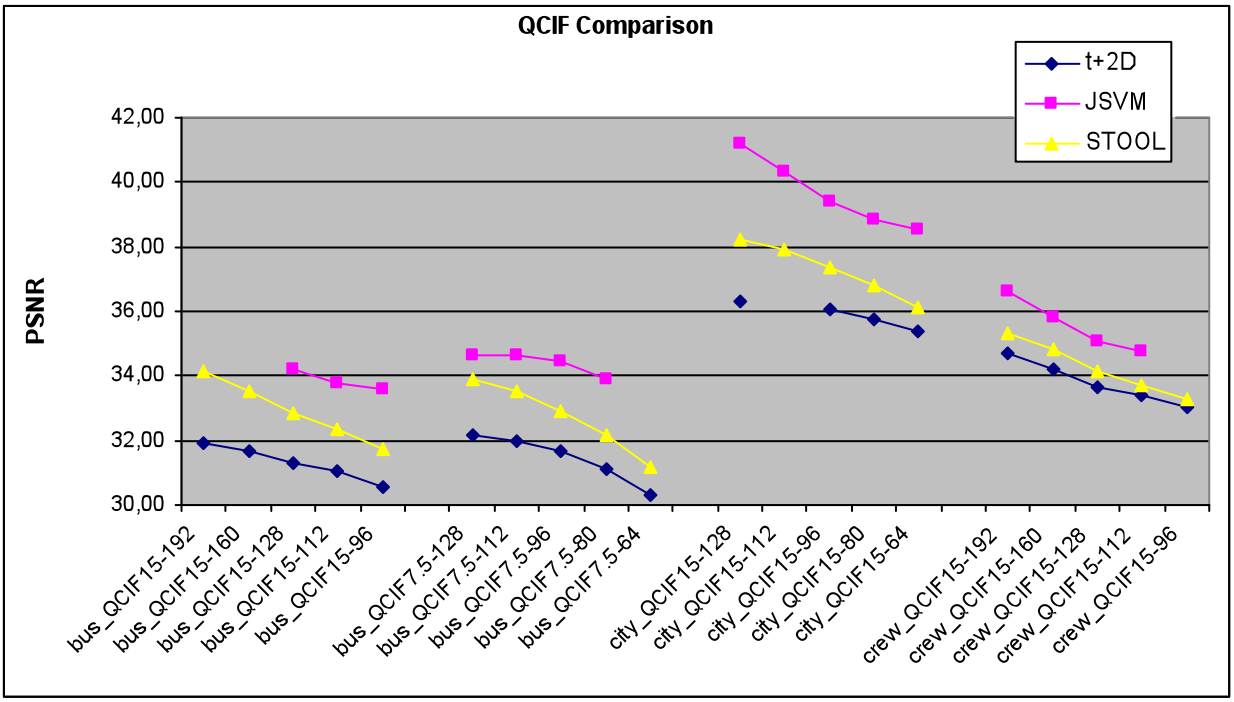
4.2.1.1 PSNR comparison with original references

Figure7 presents a complete PSNR comparison at QCIF resolution. As known only trends for each system are meaningful since the different coding schemes use different reference sequences (MPEG downsampling filters for JSVM3.0, 9/7 wavelet filterbank for "t+2D" Vidwav Reference Software configuration, 3-LS filters for "2D+t+2D" configuration) relative difference in PSNR between the three coding schemes lose significance.

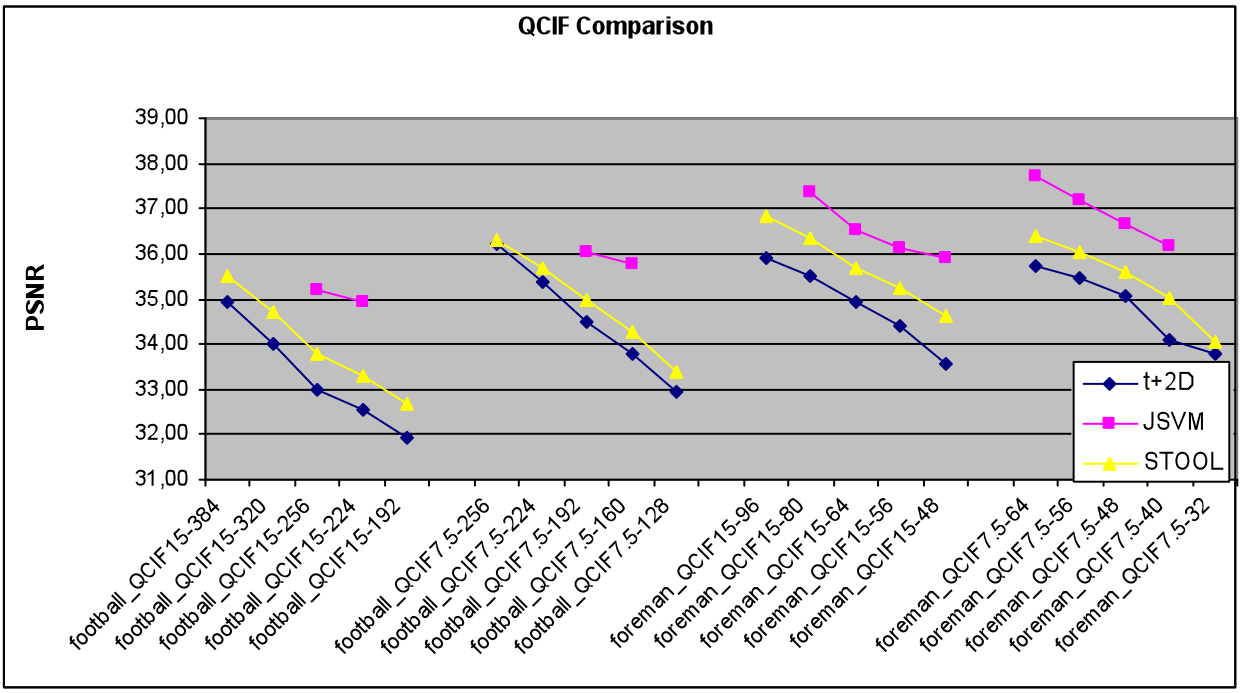
4.2.1.2 PSNR comparison with averaged references

In Figure 8 we compare the PSNR results obtained on two sequences using both system related references and a common reference for JSVM3 and STool.

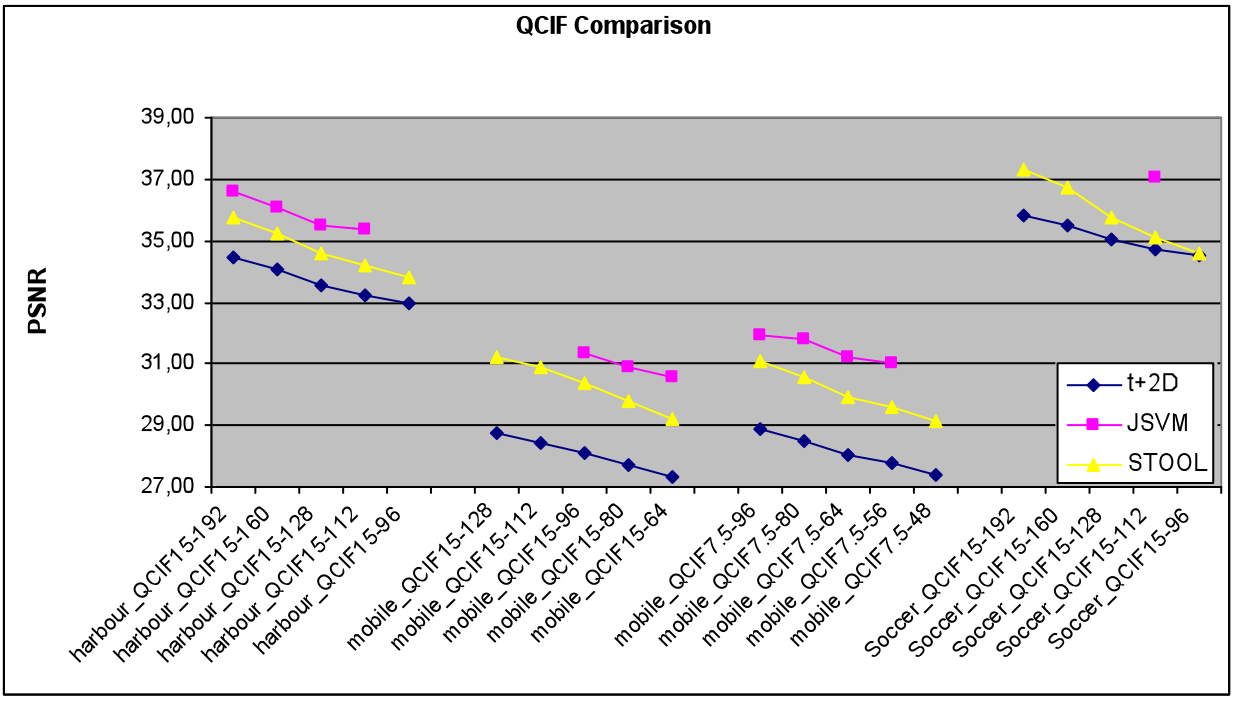
Results in Fig. 8 indicate that using a common reference, "2D+t+2D" configuration PSNR results are very close (and sometimes outperforms) those of JSVM3.


QCIF Comparison
PSNR
42,00
40,00
38,00
36,00
34,00
32,00
30,00
t+2D
JSVM
STOOL
bus_QCIF15-192
bus_QCIF15-160
bus_QCIF15-128
bus_QCIF15-112
bus_QCIF15-96
bus_QCIF7.5-128
bus_QCIF7.5-112
bus_QCIF7.5-96
bus_QCIF7.5-80
bus_QCIF7.5-64
city_QCIF15-128
city_QCIF15-112
city_QCIF15-96
city_QCIF15-80
city_QCIF15-64
crew_QCIF15-192
crew_QCIF15-160
crew_QCIF15-128
crew_QCIF15-112
crew_QCIF15-96

(a)


QCIF Comparison
PSNR
39,00
38,00
37,00
36,00
35,00
34,00
33,00
32,00
31,00
t+2D
JSVM
STOOL
football_QCIF15-384
football_QCIF15-320
football_QCIF15-256
football_QCIF15-224
football_QCIF15-192
football_QCIF7.5-256
football_QCIF7.5-224
football_QCIF7.5-192
football_QCIF7.5-160
football_QCIF7.5-128
foreman_QCIF15-96
foreman_QCIF15-80
foreman_QCIF15-64
foreman_QCIF15-56
foreman_QCIF15-48
foreman_QCIF7.5-64
foreman_QCIF7.5-56
foreman_QCIF7.5-48
foreman_QCIF7.5-40
foreman_QCIF7.5-32

(b)



(c)

Figure 7. (a-c) PSNR comparison at QCIF resolution

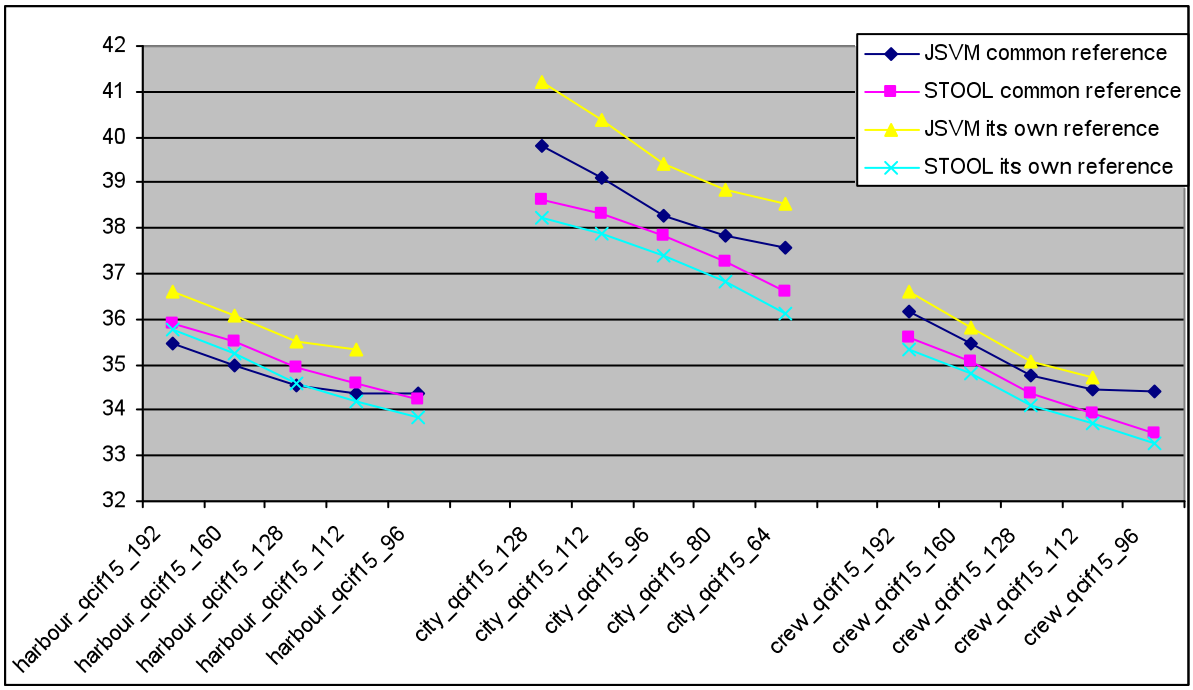


Figure 8. PSNR at QCIF resolution: common reference usage

4.2.1.3 Visual comparison at QCIF resolution

We show a visual comparison among some sample frames. In Fig.9 some 15fps 128kpbs decoded frames of the CREW sequence are displayed, and in Fig.10 a representative frame of the 7.5fps decoded FOOTBALL sequence is shown for 2 different bit-rates.

	JSVM	STool	"t+2D"
Fr 17			
Fr 31			
Fr 83			

Figure 9. visual comparison on CREW QCIF 15fps 128kbps

	JSVM	STool	"t+2D"
128 kbps			
256 kbps			

Figure 10. visual comparison on FOOTBALL QCIF 7.5 (frame 17)

4.2.2 Intermediate spatial resolution

This is the case of CIF sequences extracted from 4CIF coded bit-streams. In this situation the STool interscale prediction is applied once while the VidWav "t+2D" applies the inverse MCTF using one level downscaling of the motion field. Figure 11 shows a visual comparison on the CITY sequence. All three sequences are visually close. In Fig. 12 we show some PSNR results with or without using a common reference. Similar remarks on the common reference usage , previously made for QCIF resolution, apply also in this case.

4.2.3 Highest spatial resolutions

4.2.3.1.1 CIF originals

We propose a visual comparison for the sequences FOOTBALL (Fig.13) and MOBILE (Fig.14). In these cases the CIF resolution is the highest one. We remarked that from a visual point of view the decoded sequences are very close.

Figure 11. City_CIF15-192: (top) STool (192kbps) mean PSNR 34.05dB, (mid) "t+2D" ref sw (195kbps) mean PSNR 33.43dB, (bottom) JSVM3 (192kbps) mean PSNR 36.76dB

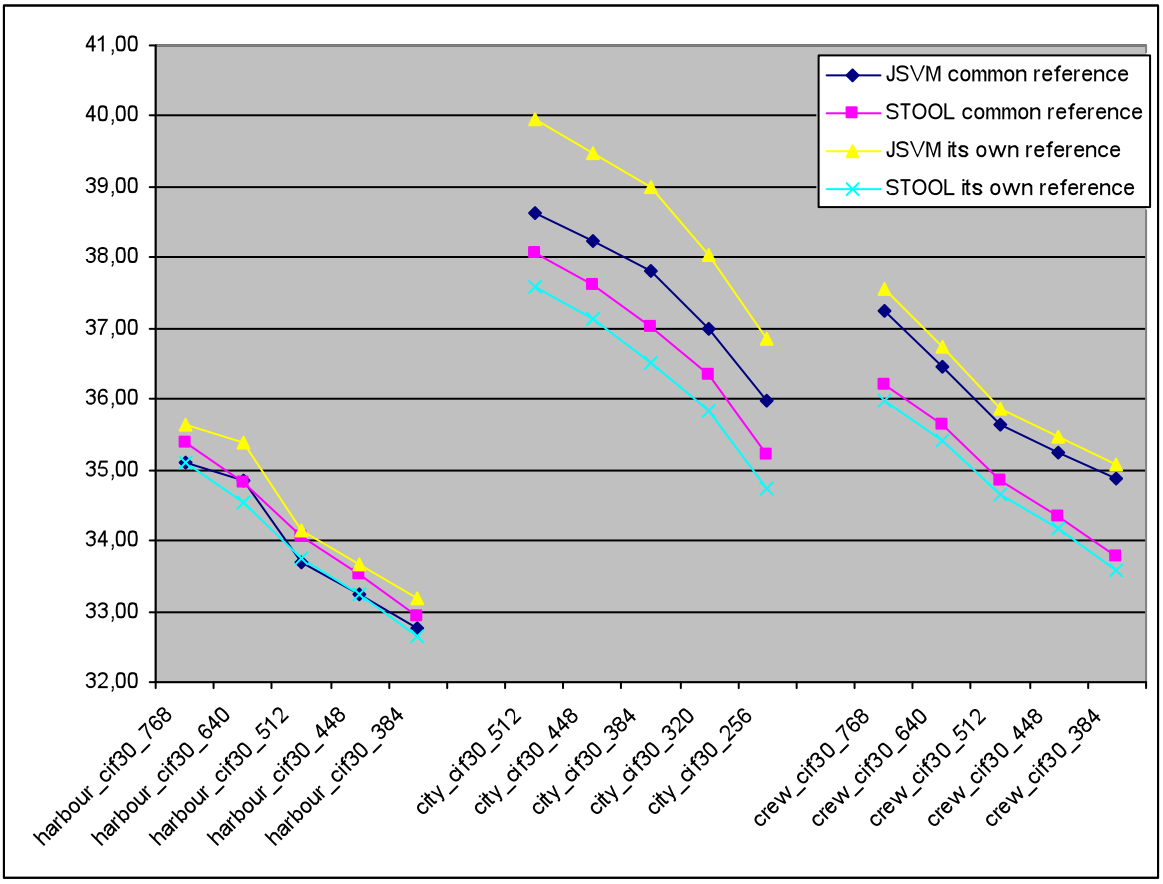


Figure 12. some PSNR results at CIF resolution (with and without a common reference)


(a)


(b)


(c)

Figure 13: Football_CIF30-1024: (a) JSVM mean PSNR 35.95dB (b) STool 34.62dB (c) "t+2D" (1.128Mbps) 36.0db


(a)


(b)


(c)

Figure 14: Mobile_CIF30-384: (a) JSVM (384kbps) mean PSNR 31.04dB (b) STool (384kbps) 29.63dB (c) "t+2D" (**429**kbps) 31.26dB

4.2.3.2 4CIF originals

In this case (Fig. 15), even if the current "2D+t+2D" STool VidWav implementation suffer from the redundancy of the motion vector representation (this find correspondence in terms of PSNR performance) visual performance remains inferior but comparable with respect to the other schemes.

(a)

(b)

(c)

Figure 15: HARBOUR 4CIF 30fps 1024kbps: (a) STool PSNR 33.02dB, (b) "t+2D" PSNR 34.45dB, (c) JSVM3 PSNR 32.58dB

4.3 Performance Evaluation Summary

A comparison among the decoded sequences by JSVM3.0, t+2D with AVC base-layer (with the currently best configuration files) and the current 2D+t+2D scheme (Stool).

At the highest point t+2D is the best coder, but it shows problems at lower resolution (for example in QCIF and CIF soccer sequences there are evident artefacts). STool solves these problems and at lower resolution has performance competitive with JSVM3.0 (visual comparison is suggested). At highest resolution, STool scheme does not appear yet to be competitive with JSVM 3.0 since on one hand, motion information is independently coded in the different spatial resolution layer, without considering the relation between the layer motion, on the other hand inconsistent mode decision take place at times across different spatial resolution layers.

5 Perspectives towards future improvements

Motion estimation resolution problem in current t+2D implementation

In order to support spatial resolution scalability, temporal levels that will be decoded on targeted lower spatial resolutions must use large macroblocks (64x64, 32x32), since from

an implementation point of view, decoding is not working at low resolution for smaller size blocks.

2D+t+2D (Stool) inter-layer issues (currently not supported)

- consistent mode decision (e.g. intrablock) across spatial resolution layers
- consistent motion estimation across spatial resolution layers, for ensuring:
  - good prediction of LL on higher spatial resolution
  - optimal coding of motion field

New tools

Replace Intra-coding mode with Motion Adaptive Transform (better tuned to small areas of uncovered background)

Entropy coding (both for t+2D and 2D+t+2D architectures)

- Same scale temporal and spatial subbands appear to be coded separately (which means at the level of individual subband level), but given the 3D EBCOT used, good context requires the use of motion information which is not available within any given subband, and should be predicted or estimated to take into account the advantage of context information.

6 VidWav history

This section provides an overview of the history of VidWav AhG from its establishment during the 70th MPEG meeting (Palma, ES). In the first subsection all the documents produced within the VidWav are summarised, while in the second subsection the participants are listed.

6.1 Meetings and input documents

6.1.1 Meeting 71 Hong-Kong, China:

10 input documents:

11680	Ruiqin Xiong Jizheng Xu Feng Wu Dongdong Zhang	Studies on Spatial Scalable Frameworks for Motion Aligned 3D Wavelet Video Coding
11681	Dongdong Zhang Jizheng Xu Hongkai Xiong Feng Wu	Improvement for In-band Video Coding with Spatial Scalability

11713	Markus Beermann Mathias Wien	Application of the Bilateral Filter for Quality-Adaptive Reconstruction
11732	Christophe Tillier Beatrice Pesquet-Popescu	CBR 3-band MCTF
11738	Gregoire Pau Beatrice Pesquet-Popescu	Optimized Prediction of Uncovered Areas in Wavelet Video Coding
11739	Gregoire Pau Beatrice Pesquet-Popescu	Four-Band Linear-Phase Orthogonal Spatial Filter Bank in Wavelet Video Coding
11741	Gregoire Pau Jerome Vieron Beatrice Pesquet-Popescu	Wavelet Video Coding with Flexible 5/3 MCTF Structures for Low End-to-end Delay
11748	G.C.K. Abhayaratne Ebroul Izquierdo	Wavelets based residual frame coding in t+2D wavelet video coding
11750	Marta Mrak Nikola Sprljan G.C.K. Abhayaratne Ebroul Izquierdo	Scalable motion vectors vs unlimited precision based motion compensation at the decoder in t+2D wavelet video coding
11757	Woo-Jin Han Kyohyuk Lee	Comments on wavelet-based scalable video coding technology

During the 71st meeting, wavelet based software from Microsoft Research Asia (MSRA) has been chosen as the common software for the investigation and evaluation within the VidWav.

6.1.2 Meeting 72 Busan, Korea :

7 input documents:

11844	Z. K. Lu W. S. Lin Z. G. Li K. P. Lim X. Lin S. Rahardja E. P. Ong S. S. Yao	Perceptual Region-of-interest (ROI) based Scalable Video Coding
11952	ChinPhek Ong ShengMei Shen MenHuang Lee Yoshimasa Honda	Wavelet Video Coding - Generalized Spatial Temporal Scalability (GSTS).
11975	Ruiqin Xiong Jizheng Xu	Coding Perfromance Comparison Between MSRA Wavelet Video Coding and JSVM

	Feng Wu	
11976	Yihua Chen Jizheng Xu Feng Wu Hongkai Xiong	Improvement of the update step in JSVM
12008	Markus Beermann Mathias Wien	De-ringing filter proposal for the VIDWAV Evaluation software
12056	Christophe Tillier Grégoire Pau Béatrice Pesquet-Popescu	Coding performance comparison of entropy coders in wavelet video coding
12058	Grégoire Pau Béatrice Pesquet-Popescu	Comparison of Spatial *M*-band Filter Banks for *t+2D* Video Coding

6.1.3 Meeting 73 Poznan, Poland:

7 input documents:

12176	Vincent Bottreau Grégoire Pau Jizheng Xu	Vidwav evaluation software manual
12286	Ruiqin Xiong Jizheng Xu Feng Wu	Responses to Vidwav EE1
12303	Grégoire Pau Maria Trocan Béatrice Pesquet-Popescu	Bidirectional Joint Motion Estimation for Vidwav Software
12339	Ruiqin Xiong Xiangyang Ji Dongdong Zhang Jizheng Xu Grégoire Pau Maria Trocan Vincent Bottreau	Vidwav Wavelet Video Coding Specifications
12374	Markus Beermann	Joint reduction of ringing and blocking for VidWav
12376	Yongjun Wu John Woods	Aliasing reduction for subband/wavelet scalable video coding
12410	Soroush Ghanbari Leszek Cieplinski	Results of Vidwav Exploration Experiment 3

1 output document:

7334	Wavelet Codec Reference Document and Software Manual

6.1.4 Meeting 74 Nice, France:

7 input documents:

12616	Gregoire Pau Beatrice Pesquet-Popescu	Proposal of Vidwav OBMC bug fix
12633	Nikola Sprljan Marta Mrak Naeem Ramzan Ebroul Izquierdo	Motion Driven Adaptation of Spatial Wavelet Transform
12639	Nicola Adami Michele Brescianini Riccardo Leonardi	Edited version of the document SC 29 N 7334
12640	Markus Beermann Mathias Wien	Wavelet Video Coding EE4: Joint Reduction of Ringing and Blocking
12642	Nicola Adami Michele Brescianini Riccardo Leonardi Alberto Signoroni	New prediction schemes for scalable wavelet video coding
12643	Nicola Adami Michele Brescianini Riccardo Leonardi Alberto Signoroni	Performance evaluation of the current Wavelet Video Coding Reference Software
12699	Ruiqin Zhong	Verification of Vidwav EE4 results of RWTH

3 output documents:

	Status Report on Wavelet Video Coding Exploration
	Description of Exploration Experiments in Wavelet Video Coding
	Wavelet Codec Reference Document and Software Manual WCS 1.1

6.2 VidWav participation

6.2.1 Academic Institutions

- ENST Paris
- University of Brescia, Italy
- RWTH Aachen University
- Queen Mary, University of London, United Kingdom.
- Rensselaer Polytechnic Institute
- Institute of Computing Technology, Chinese Academy of Sciences
- Image Communication Institute, Shanghai Jiao Tong University

6.2.2 Research Institutions and Industry

- Institute for Infocomm Research, Singapore
- IRISA/INRIA Rennes
- Microsoft Research Asia
- Mitsubishi Electric ITE-VIL
- Samsung Electronics

- Thomson R&D

7 References

[1] 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.

[2] 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. of VCIP 2005, SPIE vol. 5960 (nr.58), Beijing, China, July 2005.

[3] N. Adami, M. Brescianini, R. Leonardi and A. Signoroni, "New prediction schemes for scalable wavelet video coding", ISO/IEC JTC1/SC29/WG11, M12642, 74th MPEG Meeting, Nice, France, Oct. 2005.

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

[5] 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. 2005.

[6] ISO/IEC JTC1/SC29/WG11, "Description of Core Experiments in MPEG-21 Scalable Video Coding," N6521, Redmond, July 2004.

[7] 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.

[8] 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.

[9] A. Secker and D. Taubman, "Lifting-Based Invertible Motion Adaptive Transform (LIMAT) Framework for Highly Scalable Video Compression," IEEE Trans. Image Processing, vol. 12, no. 12, pp. 1530-1542, Dec. 2003.

[10] 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.

[11] Jizheng Xu, Ruiqin Xiong, Bo Feng, Gary Sullivan, Ming-Chieh Lee, Feng Wu, Shipeng Li: "3-D Subband Video Coding Using Barbell Lifting", ISO/IEC JTC1/SC29/WG11, M10569/S05, 68th MPEG Meeting, Münich, Germany, Mar. 2004.

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

[13] 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. of VCIP 2003, SPIE vol. 5150, pp. 719-731, Lugano (CH), July 2003.

[14] Subjective test results for the CfP on Scalable Video Coding Technology, ISO/IEC JTC1/SC29/WG11, M10737, 68th MPEG Meeting, Munich, Germany, Mar. 2004.

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

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

[17] 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. 2005.

[18] ISO/IEC-JTC1 and ITU-T, "Joint Scalable Video Model (JSVM) 3.0 Reference Encoding Algorithm Description", ISO/IEC JTC1/SC29/WG11, N7311, 73th MPEG Meeting, Poznan, Poland, July 2005.

[19] ISO/IEC-JTC1 and ITU-T, "Joint Scalable Video Model (JSVM) 3.0 Reference Encoding Algorithm Description", ISO/IEC JTC1/SC29/WG11, N7311, 73th MPEG Meeting, Poznan, Poland, July 2005.

[20] ISO/IEC JTC1/SC29/WG11, "Requirements and Applications for Scalable Video Coding v.5," N6505, Redmond, July 2004.

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

[22] 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.

[23] 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.

[24] N. Adami, M. Brescianini and R. Leonardi, "Performance evaluation of the current Wavelet Video Coding Reference Software", ISO/IEC JTC1/SC29/WG11, M12643, 74th MPEG Meeting, Nice, France, Oct. 2005.