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/M12642
October 2005, Nice, F

Source	University of Brescia (Italy) – Signals & Communications Laboratory
Title	**New prediction schemes for scalable wavelet video coding**
Status	Contribution
Sub group	Video Group
Authors	Nicola Adami, Michele Brescianini, Riccardo Leonardi, Alberto Signoroni
	Contacts: **nicola.adami@ing.unibs.it**, **michele.brescianini@ing.unibs.it** **riccardo.leonardi@ing.unibs.it**, **alberto.signoroni@ing.unibs.it**

1 Introduction

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 [1-5]. A classification of SVC architectures has been suggested by the MPEG Ad-Hoc Group on SVC [6]. The so called t+2D schemes (one example is [2]) 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 [7]), 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 [8,9] 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).

In this document we recall the STool scheme principles, already presented in [10]. We present an STool SVC architecture and compare it with respect other SVC schemes. Some main

advancements and new solutions are detailed and the related results presented. Our software implementations are based on the VidWav reference software [11,12].

1.1 The STool 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 [6, 13] the inter-scale prediction (ISP) is obtained by means of data interpolation from level *s* to level *s*+1. The STool idea [10] 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 STool architectures are typically of the 2D+t+2D kind and ISP predictions take place without the need of data interpolation. STool architectures can be designed to be fully space-time-quality scalable [14], and multiple adaptation capabilities [15] can be used without scarifying coding performance. In STool architectures some critical issues that afflicts t+2D and 2D+t schemes are not present.

1.2 STool architectures

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 1 (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 our STool 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} we want 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 2). 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 STool 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.

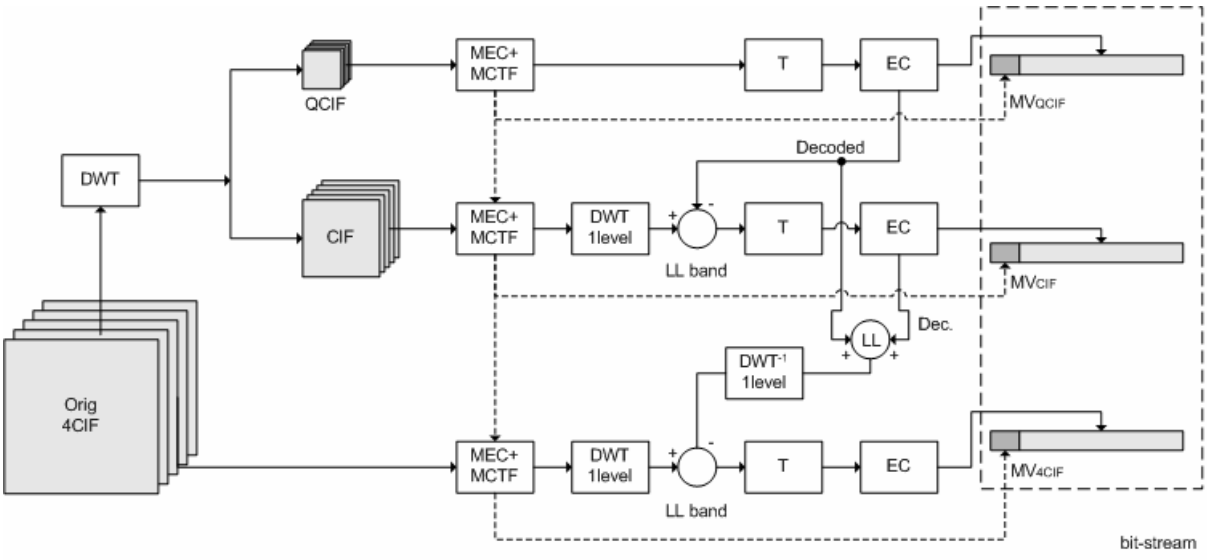

Figure 1. STool coding architecture.

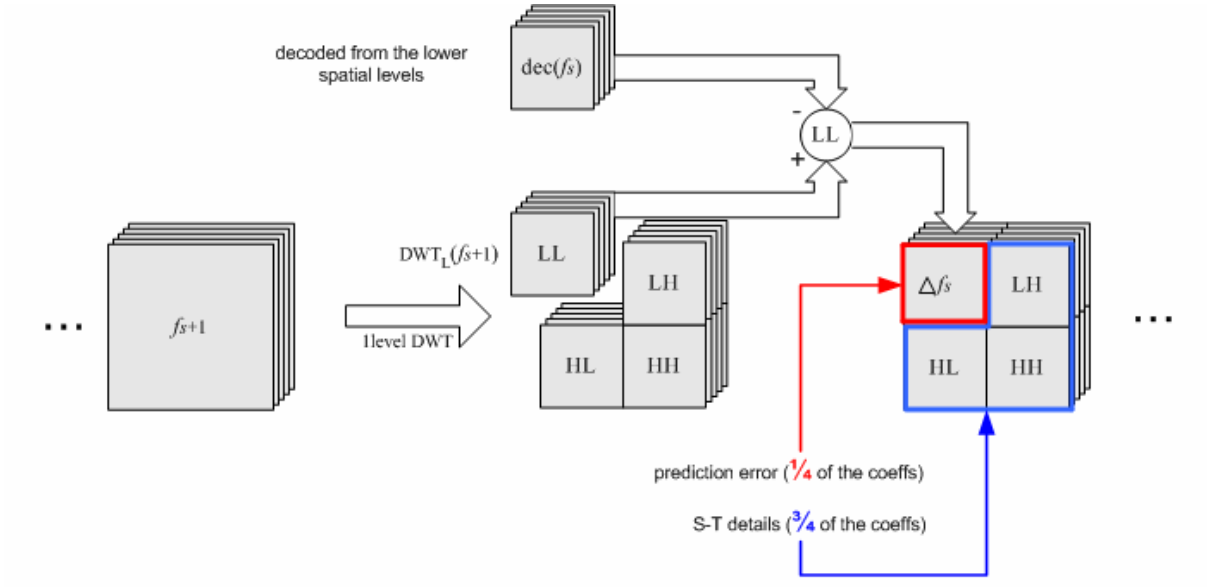

Figure 2. STool prediction detail

1.3 STool and other SVC architectures

We now aim at giving some insight about the differences between the proposed method and other existing techniques for hierarchical representation of video sequences. As explained in 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. We note that this fact is of great importance in wavelet-based video coding scheme, because the differences between the $dec(f_s)$ and $dwt_L(f_{s+1})$ are responsible for a loss in performance in the t+2D schemes as explained hereafter.

1.3.1 T+2D

A deeper analysis of the differences between our 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 [16]) 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.

1.3.2 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 [7]. Motion field coherence among subbands and increased computational complexity are among the residual problems of this approach.

1.3.3 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 predction is made just after the temporal transform but only on intra (not temporally transformed) blocks [13]. These architectures have got their basis in the first hierarchical representation technique introduced for images, namely the Laplacian pyramid [17]. 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 that we have already shown to be problematic due to the relative inefficiency of motion estimation and compensation across the spatial subbands.

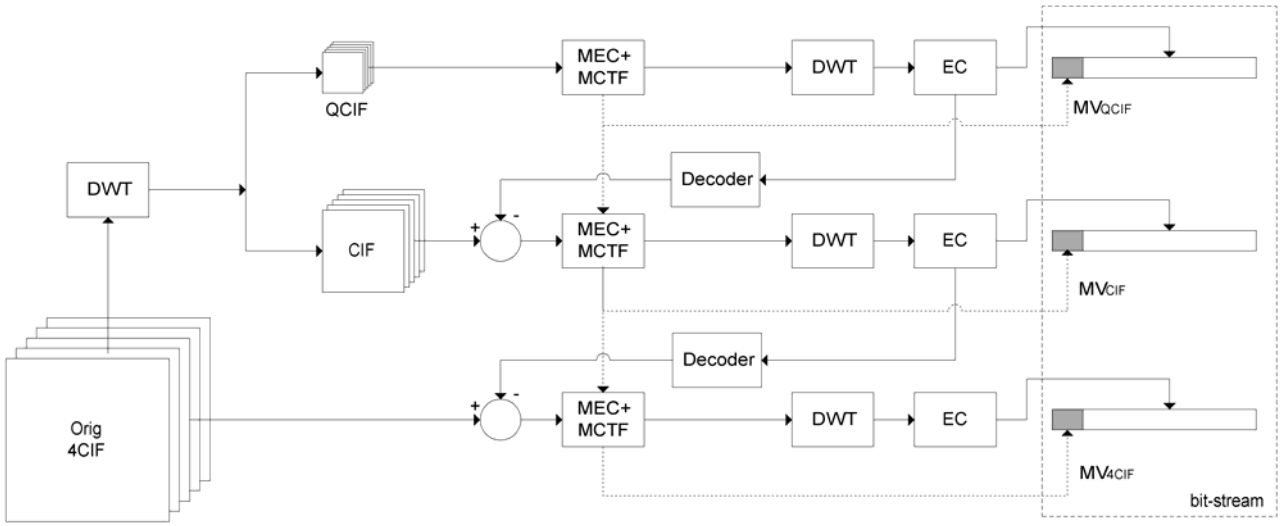

Figure 3. 2D+t+2D pyramidal scheme: prediction with interpolation.

1.3.4 STool 2D+t+2D

Looking at the above issues the 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.

2 STool implementation on the VidWav Reference Software

The advancements and the experimental results presented in the following have been implemented and obtained with modifications of the VidWav reference software [11] as described in the document [12]. T and EC of Fig. 1 are then implemented by DWPT (Discrete Wavelet Packet Transform) and 3D-ESCOT respectively.

3 STool advancements

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 tell us that the STool idea is somehow “robust” and can be used in a non strictly wavelet based coding environment.

3.2 STool 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.

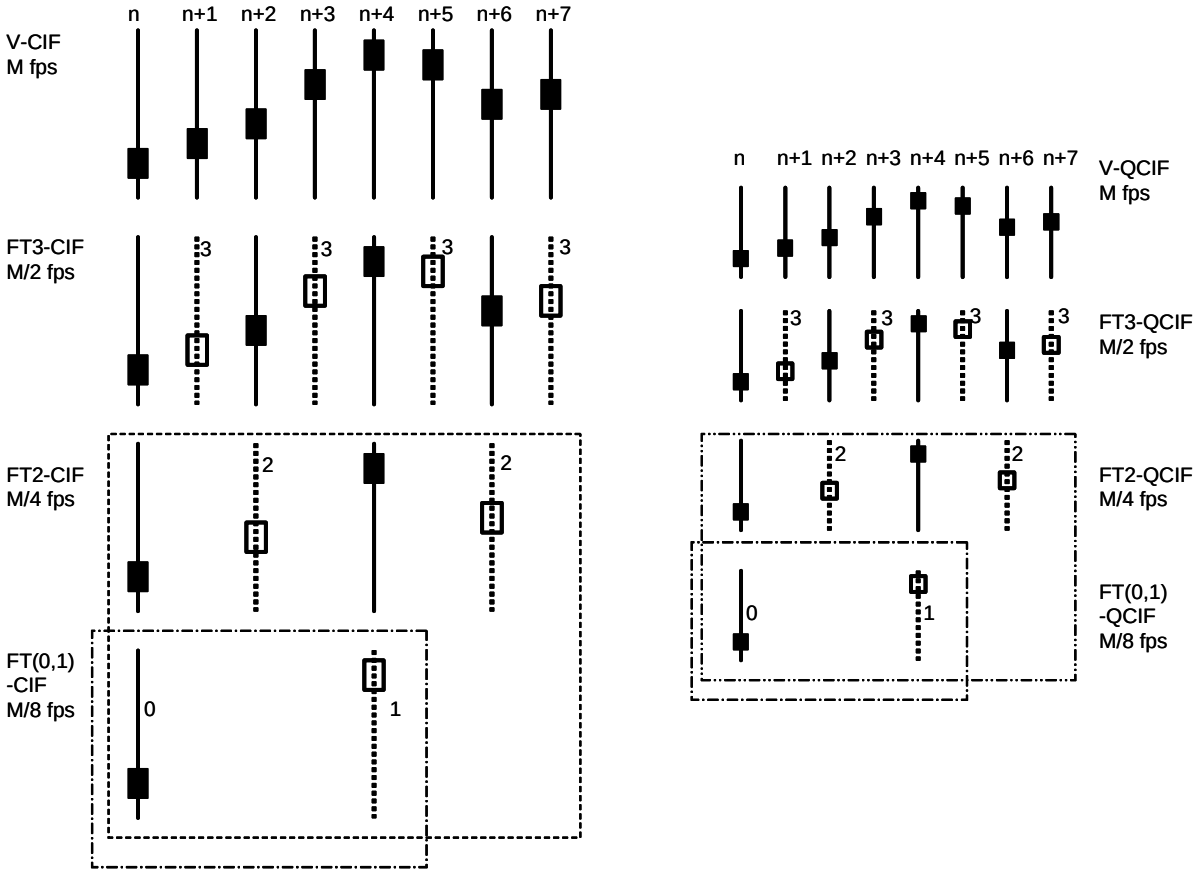

Figure 4. Possible variations of the STool prediction on the MCTF subband hierarchy

3.3 STool 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.

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.

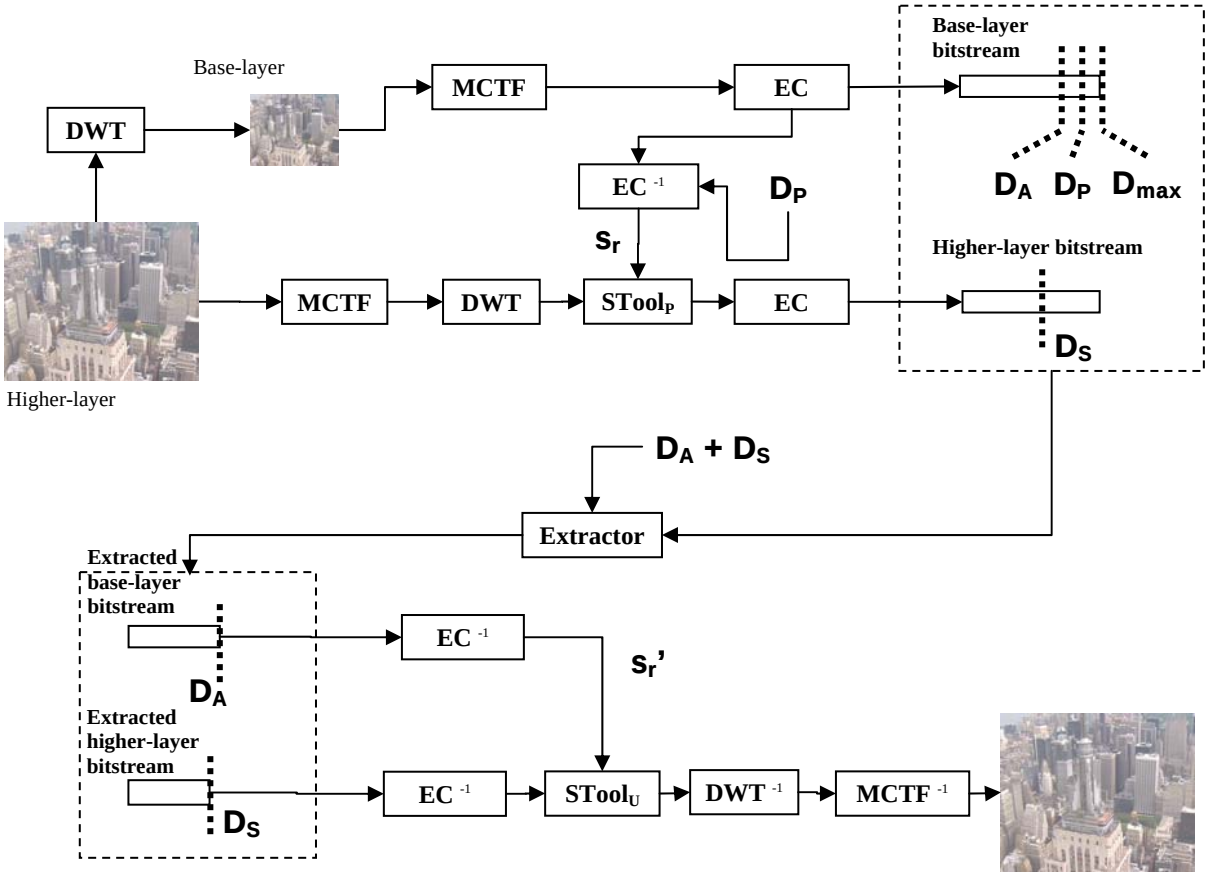

Figure 5. Asymmetric closed loop STool prediction

4 Final considerations and results

Performance evaluation of the current STool implementation on the VidWav Wavelet Video Coding reference software are presented on the document [18].

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

Table 2 reports the average luminance PSNR for the interpolation based pyramidal 2D+t+2D scheme of Figure 3 in comparison with the proposed STool scheme (of Figure 1). *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 STool 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 2. 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 1, (c) coded at 384kbps with the interpolation pyramidal scheme of Figure 3.

4.2 Improvements with respect to the MPEG meeting of Palma (Oct. 2004)

We calculated the one year improvement of the STool and of the JSVM schemes on the lower resolution. We compare today results (current document and [13] respectively) with the results presented at the MPEG Palma Meeting in Oct.2004 ([10] 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.

Table 2: PSNR improvements on the QCIF resolution

Sequence	PSNR Palma Stool	PSNR Palma JSVM	PSNR Nice Stool	PSNR Nice JSVM	Difference Stool	Difference JSVM
Bus	31,49	33,96	32,34	34,02	0,85	0,06
Foreman	33,46	36,52	35,17	36,64	1,71	0,12
Football	32,23	35,91	33,94	36,04	1,71	0,13
Mobile	27,45	30,83	29,77	30,89	2,32	0,06
Harbour	34,69	36,06	34,73	36,06	0,04	0
City	37,07	38,92	37,23	39,73	0,16	0,81
Soccer	35,66	36,71	35,89	37,02	0,23	0,31
Crew	34,09	35,86	34,24	35,84	0,15	-0,02

5 References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

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

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

[10] N. Adami, M. Brescianini, R. Leonardi, 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.

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

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.