A fully scalable video coder with inter-scale wavelet prediction and morphological coding

Nicola Adami, Michele Brescianini, Marco Dalai, Riccardo Leonardi, Alberto Signoroni*
Signals and Communications Lab., Electronics for Automation Dept., University of Brescia (Italy)

ABSTRACT

In this paper a new fully scalable - wavelet based - video coding architecture is proposed, where motion compensated temporal filtered subbands of spatially scaled versions of a video sequence can be used as base layer for inter-scale predictions. These predictions take place between data at the same resolution level without the need of interpolation. The prediction residuals are further transformed by spatial wavelet decompositions. The resulting multi-scale spatio-temporal wavelet subbands are coded thanks to an embedded morphological dilation technique and context based arithmetic coding. Dyadic spatio-temporal scalability and progressive SNR scalability are achieved. Multiple adaptation decoding can be easily implemented without the need of knowing a predefined set of operating points. The proposed coding system allows to compensate some of the typical drawbacks of current wavelet based scalable video coding architectures and shows interesting visual results even when compared with the single operating point video coding standard AVC/H.264.

Keywords: Scalable video coding, prediction, wavelets, morphology

1. INTRODUCTION

Scalability is one of the hottest features for future emerging video coding standards. Researchers as well as industries are more and more convinced that a high degree of scalability could improve existing applications and create new scenarios for the exploitation of digital video technologies. Scalable Video Coding (SVC) is in need of technologies which enable scalability in various dimensions: spatial and temporal resolution, SNR and/or visual quality, complexity and sometimes others.[1] The discrete wavelet transform (DWT) is a congenial tool to be used in this perspective. In fact, a digital video can be decomposed according to a compound of spatial DWT and wavelet based motion compensated temporal filtering (MCTF).[2] Different kinds of spatio-temporal decomposition structures can be designed to produce a multiresolution spatio-temporal subband hierarchy which is then coded with a progressive or quality scalable coding technique. [3-7] A classification of SVC architectures has been suggested by the MPEG Ad-Hoc Group on SVC.[8] The so called t+2D schemes (one example is [4]) 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 [9]), 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.[10,11] Critical aspects reside, for example:

- 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 mechanisms (especially for 2D+t+2D schemes).

This paper introduces a new SVC architecture which demonstrates good scalability performance over a wide range of operating points. The described approach is called STool.[12] It falls under the category of 2D+t+2D approaches. More precisely, 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. As a main innovative component, Inter-Scale Prediction (ISP) can be obtained without the need to interpolate data from lower to higher resolutions (as typically performed when using a layered pyramidal representation of the information).

* Authors' e-mails: *firstname.lastname*@ing.unibs.it;
Address: DEA – University of Brescia, via Branze, 38, I-25123 Brescia (Italy) - Tel. +39 030 3715434, Fax +39 030 380014

In Section 2 the STool architecture is presented and compared with respect to other SVC architectures. The following section describes how coefficient quantization and entropy coding is obtained using a group of frames (GOF) version of the Embedded Morphological Dilation Coding technique[13], called GOF-EMDC.[14] The scalability features and the bit-stream structure and handling are discussed in Section 4, where multiple adaptation capabilities are also highlighted. Finally, Section 5 shows some experimental results, offering in particular a subjective and quantitative comparison of the proposed SVC architecture with respect to AVC/H.264.

2. A MULTILAYER PYRAMID WITH INTER-SCALE WAVELET PREDICTION

2.1. STool scheme description

A main characteristic of the proposed (SNR-spatial-temporal) scalable video coding scheme is its native dyadic spatial scalability. Accordingly, this implies a spatial resolution driven complexity scalability. Spatial scalability is implemented within a scale-layered pyramidal scheme (2D+t+2D). For example, in a 4CIF-CIF-QCIF spatial resolution implementation three different coding-decoding chains are performed, as shown in Figure 1 (MEC stands for motion estimation and coding and EC stands for entropy coding, with coefficients quantization included). Each chain operates at a different spatial level and presents temporal and SNR scalability. Obviously the information from different scale layers are not independent of each other. One may thus re-use the decoded 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 approach, that we called STool, 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 each temporal subband frame f_{s+1}. The prediction is then applied only between dec(f_s) and the low-pass 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 preserved.

2.2. Comparison with 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 $\mathrm{dwt_L}(f_{s+1})$ 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, due to the shift variant nature of the motion compensation that, even prior to quantization, 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 $\mathrm{dwt_L}(f_{s+1})$ are responsible for a loss in performance in the t+2D schemes as explained hereafter.

A deeper analysis of the differences between our scheme and the simple t+2D one reveals several advantages of the former one. A simple 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 means are being designed to obtain better motion field (see for example [15]), this is highly dependent on the working rate for the decoding process, and is thus difficult to estimate 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 corrections on this level so as to compensate the problems due to inverse temporal transform. These facts represent in our view the main reasons for which a t+2D wavelet scheme has not been able to outperform as of today more traditional schemes that ensure spatial scalable video compression.

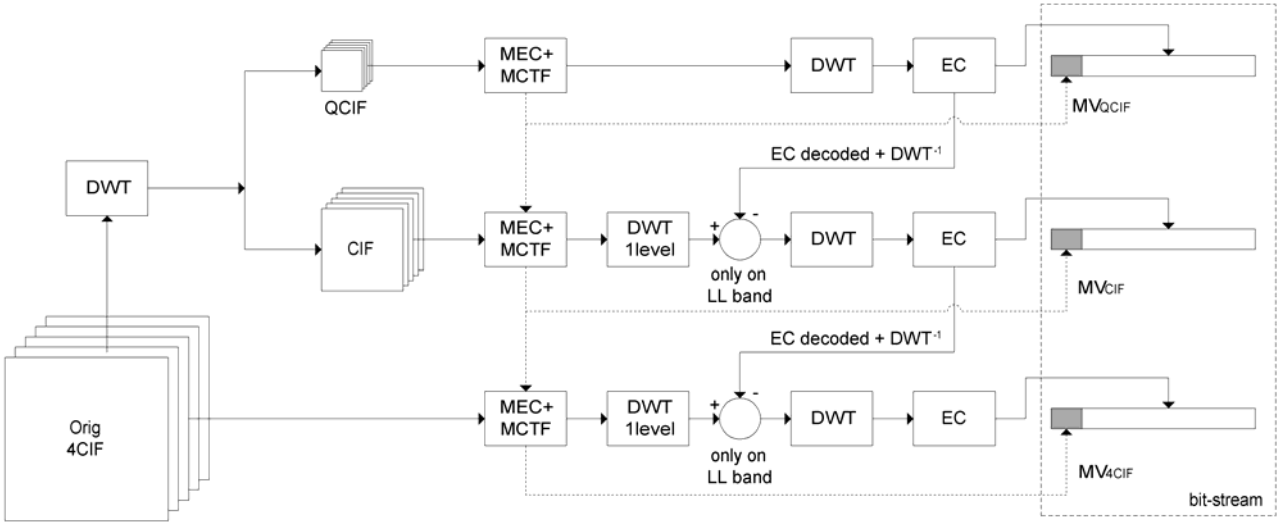

Figure 1. Overall coding architecture.

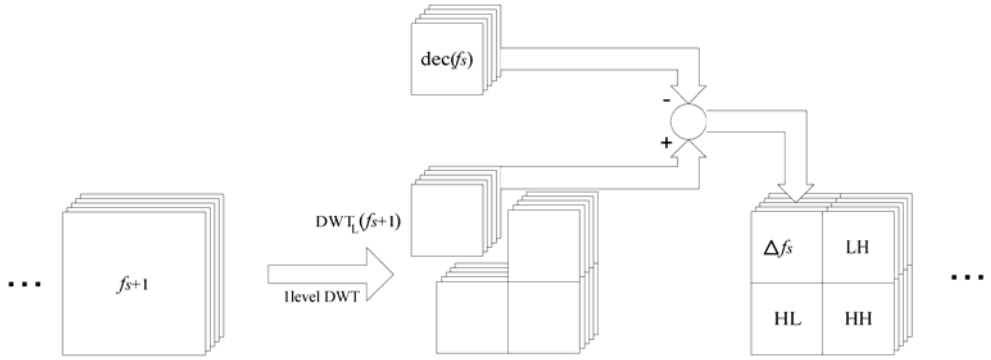

Figure 2. Inter-scale prediction (detail).

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 the 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 motion estimation and compensation take place in an overcomplete (shift-invariant) wavelet domain.[9]

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 for obtaining good performance in terms of bit-rate and scalability is to use ISP. What has been proposed in the literature towards this end is to use prediction between the lower 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. Figure 3 shows such an interpolation based inter-scale prediction scheme. This architecture has a clear basis on what have been the first hierarchical representation technique, introduced for images, namely the Laplacian pyramid.[16] 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 efficient encoding is hardly achievable. In the case of image coding, this reason led to the use of the critically sampled wavelet transforms 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 inefficiency of motion compensation across the spatial subbands.

The method proposed in this paper appears now as a valid alternative approach. This method efficiently mixes up the idea of prediction between different resolution levels within the framework of spatial and temporal wavelet transforms. Compared with the above mentioned schemes it has several advantages. First of all, different spatial resolution levels have both undergone as MCTF, which prevent from the problems of t+2D schemes. Furthermore, the MCTF are applied before spatial DWT, which resolves the problem of 2D+t schemes.

Moreover, the prediction is restricted to a number of coefficients of the same size of 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 for prediction. So, one can for example disregard the temporal high-pass subbands if a good prediction is not achievable for such “quick” details.

Alternatively this allows a QCIF at 15 fps to be efficiently used as a base for prediction of a 30 fps CIF. In order to concretely show the advantages of the proposed methods with respect to the use of interpolation in the original domain we refer to Section 5 where an experimental comparison is presented.

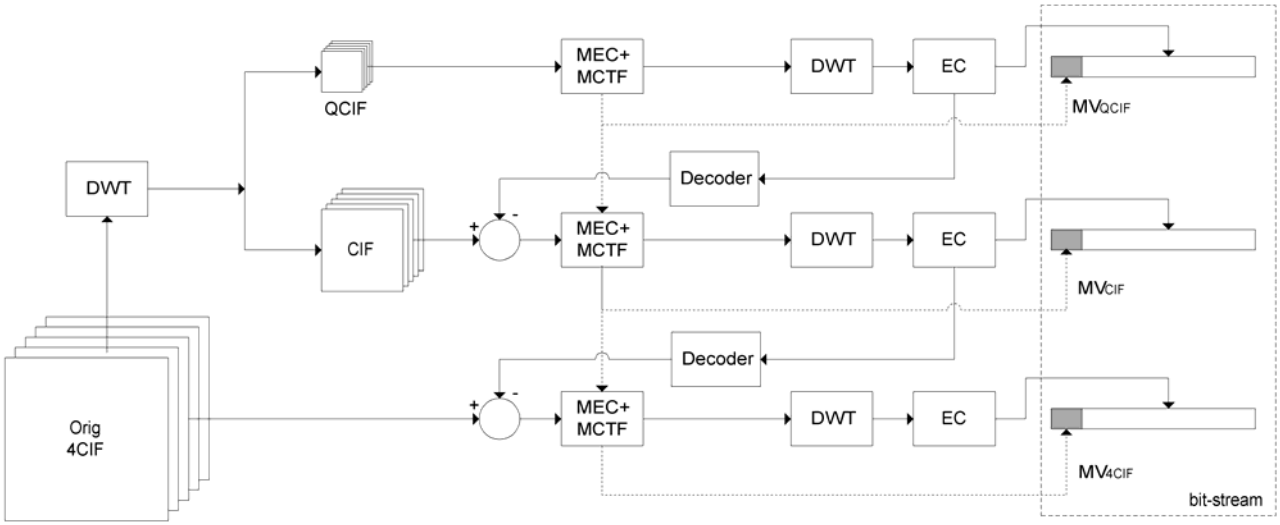

Figure 3. Pyramidal prediction with interpolation.

3. MORPHOLOGICAL SUBBAND CODING WITH GOF-EMDC

Being the last block of the SVC coding chain, Entropy Coding (EC) does not only enable the quality scalability, but it is also responsible of many other bit-stream syntax specifications which in turn are necessary to meet specific SVC requirements. In our implementation of STool, we use the Embedded Morphological Dilation Coding (EMDC) algorithm which has already been tested for 2D images[13] and 3D volumes.[17] EMDC is an embedded progressive significance map coder integrated with a context based arithmetic coder; the coding is organized according to a bit-plane based refinement of the quantization step. At each bit-plane the coefficient scanning order and the coding process follow the analysis work of a *multiresolution dilation morphological operator* which directly explore the significance map.
EMDC is the most performing codec among the family of morphological coders; its performance are comparable with the state-of-the-art wavelet coding while its complexity remains similar to the popular embedded zerotree based schemes.
In the EMDC philosophy (see the block diagram of Figure 4) each newly found significant coefficient is tested to be a seed of a significant coefficient cluster (Intra-subband dilation step). Once a cluster has been detected other hypothesis are tested: a) clusters of significant coefficients are likely to be organized in a parent child relationship so that the presence of significant coefficients are searched in the child subbands from the parent scaled positions (Inter-subband significance tree prediction step); b) when the above hypothesis are used up, before looking for the next significant coefficient, the already found cluster boundaries are explored (Extended connectivity dilation step). This last hypothesis is weaker than the others but it has been observed to contribute to the coding gain. This can be explained by the fact that in the subband domain and given a certain quantization threshold the coefficient clusters are quite irregular structures which often look more like an archipelago than a single island. Thus a weakened connectivity based dilation has been adopted in order to code this feature.

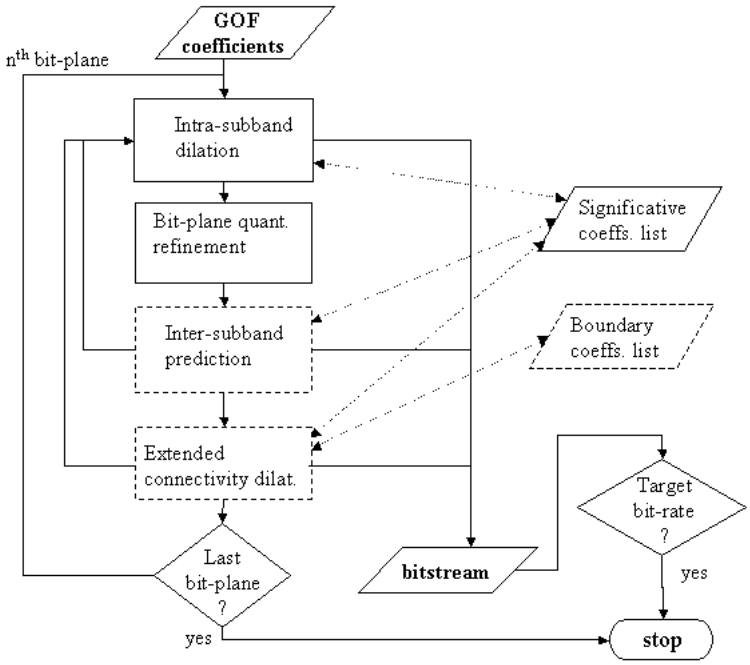

Figure 4. EMDC block diagram

In our STool scheme we implemented a GOF-EMDC (we prefer not to use the 3D term in this context) where, in this framework, a GOF is a group of MCTF generated temporal subbands (some of which has undergone to the STool prediction), followed by a spatial wavelet decomposition. Coefficient scanning inside a GOF can take place in different ways, e.g. on a per frame basis or by prioritising homologous subbands across GOF frames. Consequently, context

based arithmetic coding statistics can evolve along an entire GOF and this improves the coding performance with respect to the use of a single frame approach.
In the present GOF-EMDC implementation, the progressive bit plane quantization and entropy coding are achieved in six different stages for each bit plane. These processing stages can be prioritised differently. The selected priority order has been experimentally determined to provide the best RD tradeoff.[18] These steps are: 1) Intra-subband dilation step, 2)/3) Inter-subband significance tree prediction referred to the current/previous bit planes respectively, 4) Extended connectivity dilation, 5) Explicit position coding, 6) Quantization refinement. These steps are shown in the GOF-EMDC block diagram of Figure 4. Significance, sign, refinement bits and side information are adaptively entropy coded with the Moffat et al. arithmetic coder.[19] About 30 contexts have been properly defined. They are associated to the current coding step, the considered subband, the neighbouring and parent coefficient values.
The fact that information source with very diverse statistics (low resolution temporal subband frames, detail temporal frames at various level of resolution, STool prediction residuals) are input to the same coder must to be considered with a particular attention. Despite the fact that EMDC is quite versatile, a relevant change of the source characteristics should have an impact on the basic hypothesis on which the coding strategy is based. This effect is under our attention and, at the moment, it seems that only small changes in the coding procedure need to be made. In other words, the clustering hypothesis on the spatio-temporal wavelet coefficients appears still valid, at least for the intra-subband level. In fact, motion related residuals should be considered clustered as well. In addition, if one or more EMDC passes (e.g. those which block is dashed in Figure 4) would become inefficient (in a RD sense) or would lose sense (e.g. because DWT is not used), they can be easily dropped from the coding algorithm.
The GOF-EMDC technique also allows a decoder driven extraction and multiple adaptation paths (i.e. the possibility to generate a path of extractions at each operating point, starting from the previously extracted bit-stream). This has the great advantage that there is no advance requirement at the encoder of a precise specification of the operating points. Bit-stream and decoding features are further described in the next section.

4. BIT-STREAM AND DECODING FEATURES

As already stated, GOF-EMDC coding and the STool architecture allow a decoder driven extraction and multiple adaptation paths. This means that the quantity and/or precise specification of extraction points need not to be known in advance by the encoder. Let us consider a coding example based on three spatial scalability layers and four temporal scalable resolutions. In this case a possible structure of the bit-stream generated by our system is shown in Figure 5. As shown, we can adopt the usual video signal partition in independent groups of pictures GOP. There is a GOP header which is shared by all GOPs and contains the fundamental characteristics of the video and the pointers to the beginning of each independent GOP bit-stream. Within each GOP bit-stream we recognize three sub-bit-streams, one for each spatial level. Note that the QCIF level is coded first and its decoded version, according to our STool scheme, serves as base layer for the CIF coding. The same for CIF with respect to 4CIF resolution. As we will understand the only thing the encoder should know, in order to correctly perform the ISP, is the bit-rate with respect to which ISP prediction take place at various resolution. This bit-rate should be suitable for a good prediction and is usually selected to correspond to the maximum quality level related to the considered spatial resolution. However, other choices are possible which depends on the actual utilisation and application requirements. The above decision on which quality (and related bit-rate) must be set for ISP at various resolutions depends on wether it is suitable that in the whole bit-stream there should be parts not used for some extraction points. For example this would be the case of a refinement stream for the QCIF base layer, never used for CIF and 4CIF decoding.
Let us proceed with the bit-stream analysis. Within each spatial resolution level there is a temporal resolution partition. For example the 7.5 fps part contains data that allow to reconstruct the lowest temporal subbands of the considered spatial resolution original reference video. Such coded data contain spatio-temporal subbands and prediction residuals according to the STool ISP adopted scheme. Similarly the 15fps (30fps,60fps) part refers to the MCTF subbands at the first (second, highest) level of detail. During MCTF motion vectors are estimated and coded (MEC), they are naturally subdivided in temporal resolution levels, so they can be inserted in the bit-stream as shown in Figure 5. In our implementation of GOF-EMDC we considered a GOF as formed by all the temporal subbands of a certain spatial scale and temporal level. As GOFs are coded independently, this allows to fully enable spatio-temporal scalability and, at the same time, to fully exploit the statistics similarities of the various GOF signal sources. However, further GOF partitioning (see Figure 5 on the bottom) should be necessary for low latency decoding applications. In this case the GOF length must be properly selected, in the various temporal subband levels, depending on the delay requirements.

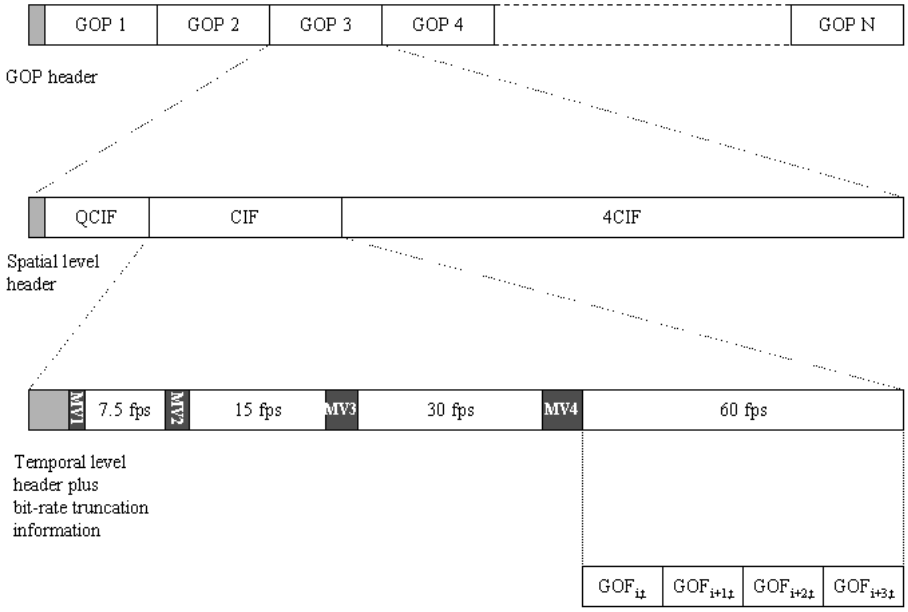

Figure 5. Bit-stream structure

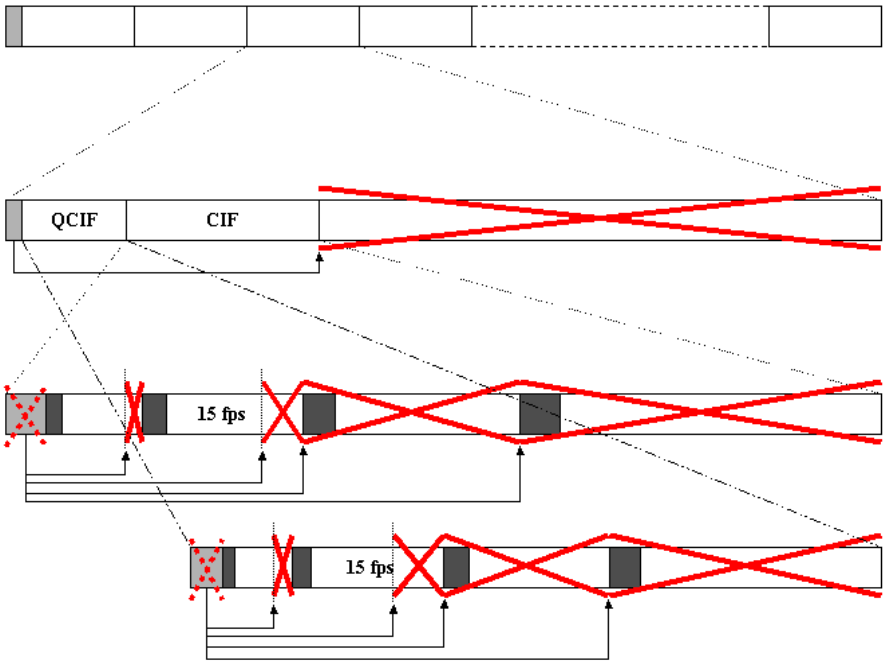

Figure 6. extraction of a sub-stream from the original bitstream

In fact, in order to reconstruct a single video frame, a certain number of temporal subbands are needed. How many of that subbands depends on the MCTF implementation and is not further discussed here.
With the so described bit-stream, the extraction of an operating point (scale, frame-rate, quality), on the basis of a budget bit-rate, consists in detecting the pertinent GOFs and (thanks to the progressivity of GOF-EMDC) truncating them according to a suitable policy. Such a multiple GOF truncation policy should be effective in a R-D sense and then derived from computational techniques or suitable heuristics. In our system, in order to let the operating point extraction to be a simple "cut and paste operation" on the whole bit-stream, some pre-computed information, related to each GOF, are inserted as heading data in order to assist the multiple-truncation. Such a data are strictly related to the GOF-EMDC structure. In fact, a table must be created for each GOF which rows correspond to the bit-plane number. The more essential table we envision is the one containing only the bit-planes end point bit-stream lengths (only one column). A more articulated possibility consists in having more columns each one associated to some of the coding step which are repeated for each bit plane (see Figure 4). In this case we gain the possibility to realize a "fractional" bit-plane truncation. Each table cell contains the bit-stream length at that point of the coding, i.e. at the end of the coding phase detected by the bit-plane/EMDC-step pair. The produced tables are then lossless coded in order to not influence the video coding performance, especially at low bit-rates. The coded tables pertaining to a certain scale are put together within the temporal header (see Figure 5). At this point multiple GOF truncation is straightforward: the whole set of table sheets must be considered and the maximum bit plane containing non empty cells is detected. From that bit-plane downwards we begin to sum-up the table cells until the target bit-budget is reached. Summation is done by scanning the table cells as follows: in decreasing order of priority, from low to thigh temporal resolutions (sheets), from first to last coding-step (columns), from highest to lowest bit-plane (rows). Exact bit-rate budget can be reached by truncating the sub-stream related to the last considered cell, according to the residual available bits.
A bit-plane based truncation generates a shorter header with respect to fractional bit-plane truncation. However, the latter leads to a better R-D behaviour. This is due to the fact that, as stated in the previous section, the GOF-EMDC coding steps have been ordered in terms of their R-D performance.
Every allowed combination in terms of spatial, temporal and quality scalability can be achieved and decided at the decoder side. Figure 6 give a pictorial representation of a target bit-rate extraction of a CIF 15fps coded sequence, from the 4CIF 60fps bit-stream of the original video.
Multiple adaptation (chained extractions along a path) can also be made very freely (of course along every advisable path) and still without any a-priori encoding settings. Multiple adaptation extraction differs from a single extraction only in one aspect: for multiple adaptations the above truncation tables must be used and re-inserted (throwing out what become useless) into the extracted bit-stream. In other words, there is no need to insert (and spend bits for) the truncation tables if no further extraction will be done on the extracted bit-stream. We observed that the additive cost of propagating the truncation table doesn't prejudice coding performance and represents a little price to pay in order to have the above described decoder driven multiple adaptation.

5. EXPERIMENTAL RESULTS

The performance of our scheme are compared here with respect to H.264/AVC (which works on single operating points), and with respect to the 2D+t+2D scheme of Figure 3. We used the Barbell lifting MCTF decomposition along with the motion estimation and coding as implemented in the Microsoft Research Asia MPEG SVC reference software.[7]
It must be said that our current implementation is not yet optimised in some aspects. One major issue concerning our present implementation is that MV at various scale are estimated and coded independently. Instead, motion vectors can be estimated and coded in a scalable way as well.[20] This would generate a twofold beneficial effect in our system. First a bit-saving on coded MV and second a further inter-scale coherence which would be beneficial on STool ISP performance. Despite these facts, obtained experimental results are favourable.

Table 1 reflects the average luminance PSNR for H.264/AVC in comparison with the proposed scheme. Coding results are reported for the *Harbour* and *Soccer* test sequences. It is important to specify that in our scheme decodable bit-streams have been extracted along a multiple adaptation path. At a first sight, the values indicate that for about a half of the operating points we have similar coding performance (in one case we do better) while there are some critical situations. However, this quantitative comparison presents some incoherencies which make an objective comparison difficult to do. First of all, except for the original 4CIF resolution video, there is a reference video problem. Original video are not the same at lower resolutions because video downsampling takes place differently depending on the

adopted scheme: with AVC an MPEG downsampling filter is normally used, while in the proposed scheme the low-pass subband of a spatial wavelet transform is taken. For the calculus of the *italic*-styled values in Table 1, the AVC downsampled videos have been used as the reference for CIF and QCIF resolutions. This leads to an apparent performance loss of the proposed method which is difficult to meaningfully quantify.
Moreover, while AVC results are optimised for each operating point, the operating points in the proposed method have been extracted along a multiple adaptation path. As a consequence, an ideal extraction cannot be granted for every point. This is the case of QCIF 15Hz 192kbps extracted from a CIF 30Hz 384kbps in Table 1. Of course, when multiple adaptation is not essential, this problem is automatically solved, since a lower point can always be extracted from a point which has enough information.
Figure 7 shows an example of visual results on a frame of the City sequence at low resolution and low bit-rate conditions: QCIF resolution, frame-rate 15fps, and bit-rate 64kbps. Visually our results appears more detailed even if some wavelet artefacts (more visible in the zoomed picture) can be detected. AVC, on this low bit-rate operating point, generates an overall more blurred image.

Table1. PSNR values (AVC and proposed method along an extraction path)

Sequence	Format	Bitrate (kbps)	PSNR_Y(dB) proposed method	PSNR_Y(dB) H.264/AVC
Harbour	QCIF 15Hz	96	*27,61*	30,57
	QCIF 15Hz	192	*28,50*	33,85
	CIF 30Hz	384	*28,49*	29,34
	CIF 30Hz	750	*30,51*	32,07
	4CIF 30Hz	1500	30,10	31,38
	4CIF 60Hz	3000	33,49	32,96
Soccer	QCIF 15Hz	96	*31,53*	33,82
	QCIF 15Hz	192	*33,73*	37,23
	CIF 30Hz	384	*31,34*	31,81
	CIF 30Hz	750	*34,10*	34,43
	4CIF 30Hz	1500	32,88	35,28
	4CIF 60Hz	3000	36,15	36,57

Table 2 reports the average luminance PSNR for the ISP-interpolation 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; 3) like 2) but STool prediction made on each temporal subband.

Table 2. PSNR comparison among different kind of inter-scale predictions

Sequence	Format	Bitrate (kbps)	PSNR_Y ISP with interpolation	PSNR_Y STool on LP t-subb (best case)	PSNR_Y STool on LP t-subb (multiple ad.)	PSNR_Y STool on all t-subb
Mobile	CIF 30fps	256	23.85	27.62	26.51	25.64
		384	25.14	29.37	28.81	27.79

The first thing we note is that it is convenient to use the QCIF-CIF STool prediction only for the low-pass temporal subband instead of using it on all the temporal subbands. Obviously intermediate solution are possible as well and this may lead to data dependent optimal choices. This opportunity actually represents an exploring field for future

optimisations. It is also important to note that the above degree of freedom is not allowed for data domain prediction based on interpolation. A second thing we note is that the multiple adaptation feature may impact not only on the enriched heading information to propagate but also on the amount of data to retain from the base layers. Obviously there are strategies to reduce this overhead, for example one may decide to discard some data and improve CIF quality by enabling the extraction only of QCIF at 15fps instead of 30fps.
Figure 8 shows an example of visual results at 384 Kbps. STool best case is compared against the interpolation based ISP. The latter scheme generates an overall more blurred image, and the visual quality gap with respect to our system is clearly visible.

Figure 7. Coding results for a frame of CITY at QCIF resolution, 15Hz, 64kbps.

6. CONCLUSIONS

In this work we presented a new ISP scheme, named STool, for wavelet based SVC. The overall architecture has been compared to other existing SVC solutions. Justification in the use of STool has been provided with reference to a more classical pyramidal ISP where inter-scale interpolation is used. Then the EC part (which include coefficients approximation) has been described. This is a GOF extension of the EMDC algorithm, where in the proposed framework a GOF is a group of temporal subbands at the same temporal decomposition level. The overall bit-stream structure has been also described in detail and the allowed feature of multiple adaptation with respect to decoded driven extraction path has been evidenced. Therefore, the overall system presents fully scalability in space (dyadic), time (dyadic) and quality (bitwise progressive). Visual and objective comparisons have been presented with respect to H.264/AVC. Even if not optimised in some aspects, the proposed system shows interesting performances. Finally, the STool architecture has been demonstrated to be sensibly superior, both objectively and visually, to a pyramidal ISP interpolating scheme.

(a) Original CIF30 (Mobile Calendar)

(b) 384kbps coded with STool prediction

(c) 384kbps coded with interpolation

Figure 8. Visual comparison at 384kbps on Mobile Calendar CIF 30fps:
(a) original frame, (b) coded with the STool scheme of Figure 1 (best case), (c) coded with the interpolation ISP scheme of Figure 3.

REFERENCES

1. ISO/IEC JTC1/SC29/WG11, “Requirements and Applications for Scalable Video Coding v.5,” N6505, Redmond, July 2004.
2. J.R. Ohm, “Three-dimensional subband coding with motion compensation,” IEEE Trans. Image Process., vol. 3, no. 5, pp. 559–571, Sept. 1994.
3. 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.
4. 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.

5. 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.
6. 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.
7. 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.
8. Scalable Video Model 2.0, ISO/IEC JTC1/SC29/WG11, N6520, 69th MPEG Meeting, Redmond, USA, Jul. 2004.
9. 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.
10. Subjective test results for the CfP on Scalable Video Coding Technology, ISO/IEC JTC1/SC29/WG11, M10737, 68th MPEG Meeting, Munich, Germany, Mar. 2004.
11. Report of the Subjective Quality Evaluation for SVC CE1, ISO/IEC JTC1/ SC29/ WG11, N6736, 70th MPEG Meeting, Palma de Mallorca, Spain, Oct. 2004.
12. 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.
13. F. Lazzaroni, R. Leonardi, & A. Signoroni: "High-performance embedded morphological wavelet coding", IEEE Signal Processing Letters, SPL-10(10): 293-295, Oct. 2003.
14. N. Adami, M. Brescianini, R. Leonardi, 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.
15. 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.
16. 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.
17. A. Signoroni, F. Lazzaroni and R. Leonardi, "Selective coding with controlled quality decay for 2D and 3D images in a JPEG2000 framework", in Proc. of VCIP 2003, SPIE vol. 5150, pp.830-841, Lugano, CH, July 2003.
18. F. Lazzaroni, R. Leonardi and A. Signoroni, "High-performance embedded morphological wavelet coding", in Proc. IWDC 2002, pp.319-326, Capri, Italy, Sep.2002.
19. A. Moffat, R. M. Neal and I. H. Witten, "Arithmetic Coding Revisited," ACM Trans. Information Systems, vol.16, pp. 256-294, Jul. 1998.
20. M. Mrak, N. Šprljan, G.C.K. Abhayaratne and E. Izquierdo, "Scalable generation and coding of motion vectors for highly scalable video coding," in Proc. of Picture Coding Symposium (PCS) 2004, San Francisco, CA, Dec. 2004.