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

ISO/IEC JTC1/SC29/WG11
MPEG2004/M11368
October 2004, Palma de Mallorca, ES

Source	University of Brescia (Italy) – Signals & Communications Laboratory
Title	SVC CE1: STool - a native spatially scalable approach to SVC.
Status	Contribution to the 70th MPEG meeting
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

1.1 Overview

This documents describes the UNIBS-SCL proposal in response to the MPEG21 SVC CE1 [1]. Our scalable video coding scheme, called STool, is based on a 2D+t+2D structure and is implemented using a modified version of the Microsoft Research Asia (MSRA) reference software [2] plus some modifications and tools which has been used in substitution. The STool architecture has been implemented in two different systems. In System-1 the modules provided in the MSRA software have been used to build the new STool architecture. In System-2 we test a new entropy coder, called GOF-EMDC, which is an extended version of the EMDC coder [3]. At the time GOF-EMDC codec and other parts of System-2 have not been optimized in many aspects, therefore we can expect better performance from our system in the next future. Despite this fact System-2 provides similar coding performances when compared to System-1. In addition, System-2 is much more flexible in many aspects, it guarantees a major number of functionalities and better fulfill the requirements list. Therefore with System-1 we intend to demonstrate the characteristics of the STool architecture, especially with respect to the reference software used, while with System-2 we customize and add functionalities to STool.

1.2 SVC-CE1 submission

We submitted extraction and decoding software for both Systems-1 and System-2, System-1 coded sequences for both scenarios 1 and 2 and System-2 coded sequences for scenario 2 only. For System-2 scenario 1 we only had deadline problems. No technical problems actually exist to produce such sequences.

2 STool: a native spatially scalable SVC scheme

The main characteristic of our (SNR-spatial-temporal) scalable video coding scheme is the native spatial scalability. This in turns imply a spatial resolution driven complexity scalability. This native spatial scalability is implemented within a 2D+t+2D approach (a mixture of the 2D+t and t+2D approach) where the lower spatial resolution information (at spatial level s) is used as a base-layer on which to predict the finer resolution spatial level s+1. For a 4CIF-CIF-QCIF implementation three different coding-decoding chains are used (Figure 1). Each chain acts at a different spatial scale level and presents temporal and SNR scalability. The idea is to use the decoded information (at a suitable quality) at a lower spatial level in order to predict the higher resolution info. In principle this can be done in different ways. Our idea consists in possibly predict the MCTF temporal subband at spatial level s+1, f_{s+1}, starting from the decoded MCTF subband (before $MCTF^{-1}$) at spatial level s, $dec(f_s)$. In order to work at the same spatial level s we perform our prediction on the lowpass subband extracted after 1 DWT level on f_{s+1}, i.e. $dwt_L(f_{s+1})$. In fact at that point the predicted subbband and the predicting ones have been subjected to the same number and kind of transformations, but in a different order (t+2D and 2D+t respectively). The prediction error $\Delta f_s = dec(f_s) - dwt_L(f_{s+1})$ can now substitute $dwt_L(f_{s+1})$ in the coding scheme (Figure 2). We call the produced frame at spatial level s+1 a *delta*-frame, and we called our approach "Substi-Tool", or more concisely STool. Obviously the question is whether the above predicted and predicting subbands actually resemble. In our experience this strongly depends on 1) the exact kind of spatio-temporal transformations and 2) the way the motion is estimated and coded at the various spatial levels. It can be expected that a certain degree of coherence must be guaranteed in the structure and precision of the motion fields in order to realize the objective to minimize the energy of the prediction Δf_s. However, in our actual implementation, we let the motion field to be estimated and coded independently at each spatial level. In fact, even if MSRA motion fields could have a layered coding, we encountered problems in doing exactly what we needed in imposing inter-level MV coherence when using the MSRA software. Then leading a spatial inter-level MV coherence should be, in our opinion, a good test field, for improving booth textural and MV coding performance.

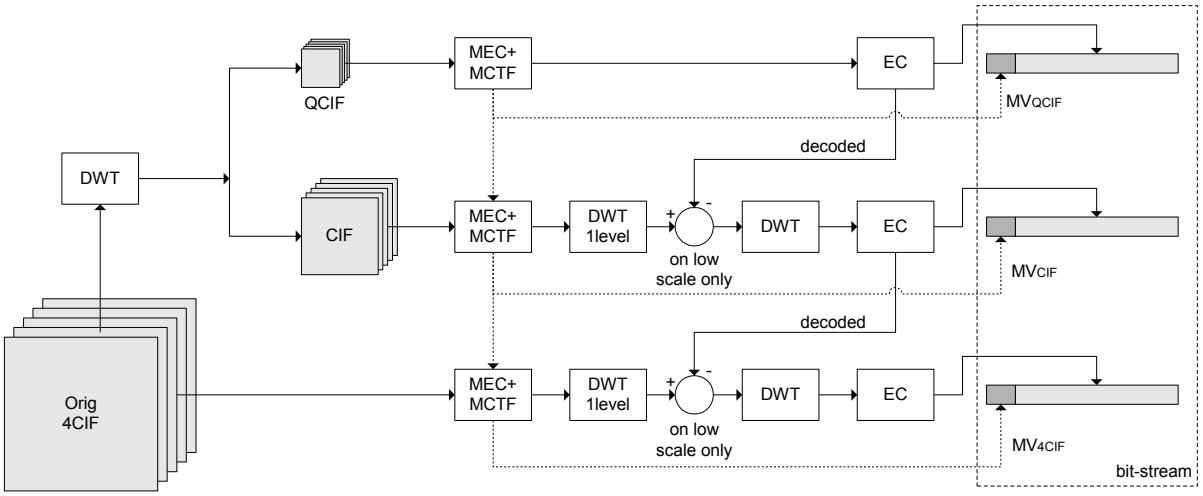

Figure 1 STool: overall coding scheme

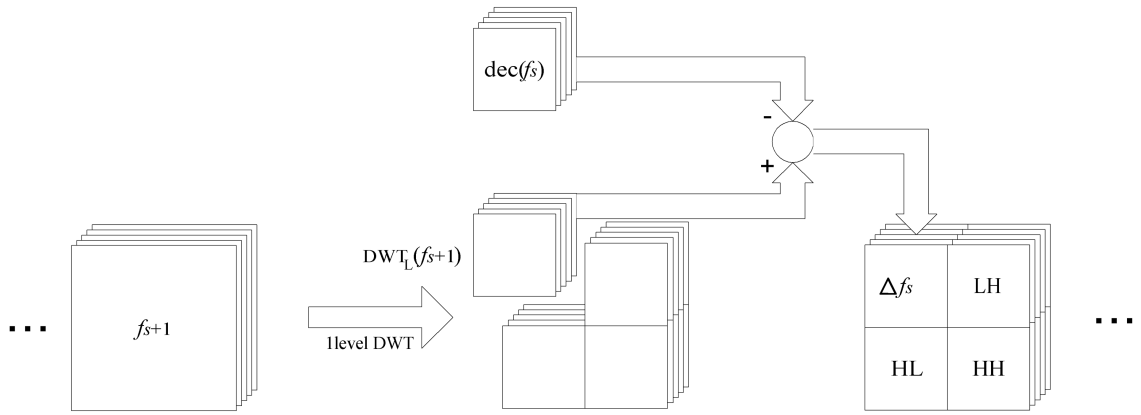

Figure 2 STool: detail

3 Subband Entropy Coding

3.1 System-1

In the System-1 implementation of STool we leave unchanged the entropy coding part of the MSRA software.

3.2 System-2

In the System-2 implementation of STool, we use an extension of our Embedded Morphological Dilation Coding algorithm [3]. EMDC has already been tested for 2D images and 3D volumes [4]. It is the most performing codec among the family of morphological coders, his performance are competitive with EBCOT and EZBC, his complexity is quite similar to the popular embedded zerotree based schemes. In fact, EMDC is an embedded progressive significance map coder integrated with a context based arithmetic coder; the coding is organized, as usual, according to a bit-plane based refinement of the quantization step. The most peculiar aspect of the scheme consists in the way the significance map is

coded: at each bit-plane the coefficient scanning order and the coding process follow the analysis activity of a "multiresolution dilation morphological operator" which directly explore the positions of significant coefficients.

In our scalable video coding System-2 we extended the EMDC to a GOF-EMDC (we prefer not to use the 3D term), where a GOF is a group of frames that in our case could be MCTF generated frames or *delta*-frames. We consider the $GOF_{i,t}$ of index i and made of same temporal level t frames. Low delay features can be traded with arithmetic coding efficiency because they could have an impact on the choice of the subband scanning order inside a GOF and/or on the size of the various $GOF_{i,t}$.

Details on GOF-EMDC and on its usage in STool System-2 can be found in [5].

4 Extractor and Decoder

4.1 System-1

As we generate 3 different bit-streams, one for each spatial level, we must produce 3 configuration files for the MSRA software. We split each global operation point into three points, one for each scale. Currently this is accomplished on a predetermined procedure in order to guarantee a good balance between prediction accuracy and detail information. The constraint we have in doing that is the maximum operating point which must be guaranteed for each spatial level.

4.2 System-2

Also in this case the total bit-rate amount is split in two levels (for Scenario 2) by keeping into account the maximum operating point for the base-layer. The use of GOF-EMDC as entropic coder, instead of the MRSA one, required the realisation of a new extractor which should work, at each spatial level, in terms temporal and SNR scalability. Also the bitstream structure has been completely re-designed. Temporal scalability means to keep all the generated $GOF_{i,t}$ bit-streams until the selected temporal decomposition deep *t*'. The bits assigned for each spatial level has to be partitioned among the selected GOF. This is a simple task because the GOF-EMDC generated bit-stream is progressive, thus SNR scalability only implies to cut the GOF bit-streams. In order to decide where to optimally cut the retained GOF bit-streams we use a "fractional bit-planes" procedure. The bit-plane fractioning is related here to the end of the various EMDC coding steps within each bit-plane. More details on the bit-stream structure and sub-stream extraction can be found in [5].

5 Requirements fulfilment

As the System-1 is mostly based on the MSRA software it inherits most of its functionalities and limits. Some of the limits have been overcome by the System-2 implementation and some additional functionalities introduced.

In the following table we declare and argument the requirements [6] fulfilment of our submitted proposals. In bold we highlight the issues which are directly related to our contribution (i.e. the STool architecture and the GOF-EMDC in System-2). We tried to be precise on the fact that STool doesn't prevent many functionalities in the cases where they depends on the properties of tools that we didn't directly implemented or modified (e.g. the MCTF or the MV estimation and coding).

Requirement		System-1	System-2
1	spatial scalability	**Dyadic** spatial scalability. As many levels as allowed by original data.	**Dyadic** spatial scalability. As many levels as allowed by original data.
2	temporal scalability	**Dyadic** temporal scalability. As many levels as allowed by original data.	**Dyadic** temporal scalability. As many levels as allowed by original data.
3	SNR scalability	**Layered quality levels.** The number of layers must be decided at the encoder side. If layer number increases performance slightly decreases.	**Progressive** bitwise truncation is possible at the decoder side, or if made by the extractor, no extra decoding directives must be generated.
4	Complexity scalability	Complexity decreases by decreasing spatio-temporal resolution and target quality.	Complexity decreases by decreasing spatio-temporal resolution and target quality.
5	Region of interest scalability	Not implemented, but can be easily provided.	Not implemented, but can be easily provided.
6	object based scalability	Not yet implemented.	Not yet implemented.
7	combined scalability	Spatio-temporal and quality level combination must be defined at the encoder side. Actually limited by some configuration problems of the MSRA software: STool base-layer generation sometimes conflicts with other needed combinations.	Spatial temporal and SNR levels can be **freely selected by the decoder** among the provided spatio-temporal levels. The maximum quality achievable at any spatial layer must be decided at the encoding side.
8	Robustness to different types of transmission errors	Not yet implemented.	Not yet implemented. Synchronization marks and the more common error resilience strategies can be used with the GOF-EMDC.
9	graceful degradation	Not yet implemented.	Not yet implemented. GOF-EMDC steps 2) and 3) can be disabled without prejudice the coding performance. This would make the coding of spatial subbands independent, thus the error propagation limited to the damaged subband and therefore error concealment much effective.
10	robustness under “best-effort” nets	Not yet implemented.	Not yet implemented.
11	10, with server and path diversity	Not yet implemented.	Not yet implemented.
12	colour depth	Can be supported.	Can be supported.
13	coding efficiency performance	Seems better than AVC at the base-layer level, **to be evaluated** at other spatial levels. Spatial aliasing kept under control using half-band DWT filters.	Seems better than AVC at the base-layer level, **to be evaluated** at other spatial levels. Spatial aliasing problems especially on some sequencies, due to spatial downsampling with DWT filters, not yet addressed.
14	base-layer compatibility	**Possible (not yet tested).** STool uses a base layer at the lower spatial resolution. If base layer is provided by another standard we just have to decode and generate temporal subband to be used as prediction for the higher resolution spatial subbands. To keep	**Possible (not yet tested).** STool uses a base layer at the lower spatial resolution. If base layer is provided by another standard we just have to decode and generate temporal subband to be used as prediction for the higher resolution spatial subbands. To keep

		the complexity low the same decoded MV could be reused for MCTF.	the complexity low the same decoded MV could be reused for MCTF.
15	Low complexity codecs	**Not implemented**.	**Not yet implemented**. **Computational complexity can be reduced** e.g. in the entropy coding part by omitting the steps 2),3) and 4) of the GOF-EMDC without prejudice coding performance. Memory complexity can be reduced by tailoring the GOF sizes to the device capability, without prejudice coding performance.
16	end-to-end delay	In the MSRA codec entropy coding is made on a GOP basis. The GOP size is actually defined by the number of temporal MCTF levels and it usually equals to 16 or 32. Then the **end-to-end delay, at the considered frame rate, is dominated by the GOP size**.	For entropy coding the GOF size at various temporal level can be adapted, without prejudice coding performance, in order to match delay-oriented constraints. Then the **end-to-end delay is limited by the MCTF** that is used. In this proposal the "Barbell Lifting" of the MSRA software has been used. However, others MCTF architectures can be used, for example low delay designed ones.
17	random access capability	Can be supported. A certain number of frames need to be decoded depending on the used MCTF and the GOP size.	**Can be supported**. A certain number of frames need to be decoded depending on the used MCTF.
18	support for coding interlaced material	Not yet implemented.	Not yet implemented.
19	System interface to support quality selection	Not yet implemented.	Not yet implemented.
20	multiple adaptations	Can be supported.	**Single and multiple adaptation are both supported**. The second simply consists in propagating into the extracted bit-streams the necessary information for subsequent extractions (see [5]), this doesn't prejudice coding performance.

6 Results

Subjective evaluation will be done during the meeting, while PSNR results only have the role to demonstrate graceful degradation when scaling down the bit-rate.

In the core experiment description [1] it is suggested to apply "method A" described in [7] in order to evaluate the PSNR performance of the proposed tools. In the case of STool, "method A" is equivalent to "method B". Then our references simply are the spatial low resolution subbands, at various resolution levels and rounded up to 8 bpp, produced by the first spatial DWT of Fig.1.

PSNR results has been produced by the cross-checker and can be found attached to the document [8]. All System-2 results are referred to the multiple adaptation extraction modality. Slightly better results should be obtained in single mode extraction.

7 Conclusions

Technical description of the University of Brescia contribution to the SVC-CE1 has been provided. Reference software used was the MSRA system. The STool architecture has been presented and submitted in two different system implementations. System-1 improve the performance of the MSRA software but inherits most of its limits. System-2 shows similar performance (our opinion in visual quality is not disclosed here in order to not bias the evaluation) but is much more flexible and provides free selection of operating points as well as arbitrary multiple adaptation paths. Low end-to-end delay as well as random access features are not prevented by STool and depends on features and tools which we didn't address in this proposal. In addition, especially for System-2, there are many aspects that, to be in time with the CE deadline, haven't been optimized; then we are confident in the possibility of further improve the visual results without having to change the main ideas presented in this document.

8 References

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

[2] Jizheng Xu, Ruiqin Xiong, Bo Feng, Gary Sullivan, Ming-Chieh Lee, Feng Wu, Shipeng Li, "3D Sub-band Video Coding using Barbell lifting," ISO/IEC JTC/WG11 M10569, S05.

[3] F. Lazzaroni, R. Leonardi and A. Signoroni, "High-performance embedded morphological wavelet coding," IEEE Signal Processing Letters, vol.10, n.10, pp. 293-295, Oct. 2003.

[4] F. Lazzaroni, A. Signoroni and R. Leonardi, "Embedded Morphological Dilation Coding for 2D and 3D Images," Visual Communication and Image Processing, VCIP 2002, SPIE vol. 4671, pp.923-934, San José, CA, USA, Jan. 2002.

[5] N. Adami, M. Brescianini, R. Leonardi and A. Signoroni, "Fully embedded entropy coding with arbitrary multiple adaptation capabilities," ISO/IEC JTC1/SC29/WG11, M11378, Palma de Mallorca, October 2004.

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

[7] ISO/IEC JTC1/SC29/WG11, "Results of SVC CE3 (Quality Evaluation)," M10931, Redmond, July 2004.

[8] W.-J. Han, "Samsung verification of Univ. of Brescia (System I and II)," ISO/IEC JTC1/SC29/WG11, M11275, Palma de Mallorca, October 2004.