ACCURACY-SCALABLE MOTION CODING FOR EFFICIENT SCALABLE VIDEO COMPRESSION

Guillaume Boisson [1], Edouard François [1] and Christine Guillemot [2]

[1]-THOMSON R&D France,
1 Avenue de Belle Fontaine, CS17616, 35576 Cesson-Sévigné, France
Email : guillaume.boisson / edouard.francois @thomson.net

[2]-IRISA, Campus Universitaire de Beaulieu, 35042 Rennes Cédex, France
Email : christine.guillemot@irisa.fr

ABSTRACT

For a scalable video coder to remain efficient over a wide range of bit-rates, covering e.g. both mobile video streaming and TV broadcasting, some form of scalability must exist in the motion information. In this paper we propose a new (t+2D) wavelet-based spatio-SNR-temporal-scalable video codec, coupled with an accuracy-scalable motion codec. It allows to decode a reduced amount of motion information at sub-resolutions, taking advantage that motion compensation requires less and less accuracy at lower spatial resolutions. This new motion codec proves its efficiency in our full-scalable framework, by improving significantly video quality at sub-resolutions without inducing any noticeable penalty at high bit-rates.

1. INTRODUCTION

During its 68th meeting, MPEG registered the responses to its Call for Proposals (CfP) on Scalable Video Coding (Cf. [1]), which can be seen as the starting point of scalable video coding standardization. Evidence had indeed been proved that scalable coding technologies can match single-layer coding performances, while addressing a number of applicative requirements that can not be easily met by non-scalable technologies (Cf. [2]). Several proposals achieved to fulfill CfP's main test, consisting in the decoding at various resolutions, frame-rates and bit-rates - from 6Mbps (high-quality TV) down to QCIF 64Kbps (mobile video streaming) - of once-encoded HD video material.

For a scalable video coder to remain efficient over such a wide range of bit-rates and resolutions, it is essential that motion information present some form of scalability. Since prior Call for Evidence on scalable video coding, several solutions have been proposed.

Hang et al. proposed in [3] a scalable motion coder coupled with famous (t+2D)WT scheme MC-EZBC. Each motion field was divided into a base layer (16x16 blocks and above) and an enhancement layer (smaller blocks). Although the adequate number of layers was empirically determined for each bit-rate, Hang et al. proved that their scalable motion codec can significantly improve MC-EZBC's low-rate performances.

As an alternative, so-called (2D+t) schemes naturally present motion scalability. Such schemes perform spatial transform first, then displacement can be estimated in each sub-band independently, before processing e.g. wavelet-domain temporal filtering [5], wavelet-domain prediction, or sophisticated contextual entropy coding [4]. In [5], Andreopoulos et al. proposed an in-band MCTF scheme, based on overcomplete wavelet transform, that outperformed spatial-domain MCTF (with full-pel accurate ME/MC and fixed block size).

In [6], Taubman and Secker showed that quantizing motion information induces a distortion roughly additive to texture quantization distortion. Using JPEG2000-like techniques (reversible wavelet transform and fractional bit-plane coding), the authors built a rate-scalable motion bit-stream and determined empirically an optimal balance between motion and texture bit-budgets. Nevertheless, performances may be limited by the amount of motion information in such a mesh-based 5/3 MCTF scheme.

Our proposed method is far less complex than previous approaches and gives promising results. The global scalable video coding framework is described in Section 2. In Section 3 we investigate the scalable motion coding issue, and propose a new layered motion representation according to spatial resolution. Experimental results in Section 4 show that significant quality improvement is perceived at lower spatial resolutions in comparison with non-scalable coding of motion. Furthermore, over-cost introduced by scalability is negligible at high bit-rates.

Simultaneously to this work, several scalable motion coding schemes have been proposed, some of those being similar to the original work presented here. The evaluation of their respective advantages has not been performed yet.

2. FRAMEWORK OVERVIEW

2.1. TWAVIX overall architecture

This work is an improvement of TWAVIX (for WAvelet-based VIdeo Coder with Scalability), whose performances have been tested comparable to state-of-the-art scalable solutions. TWAVIX is a (t+2D)WT coding scheme, briefly described in [7]. Like MC-EZBC, it performs first temporal analysis at full resolution, then spatial analysis and finally entropy coding of both motion and texture (see Fig. 1).

0-7803-8554-3/04/$20.00 ©2004 IEEE.

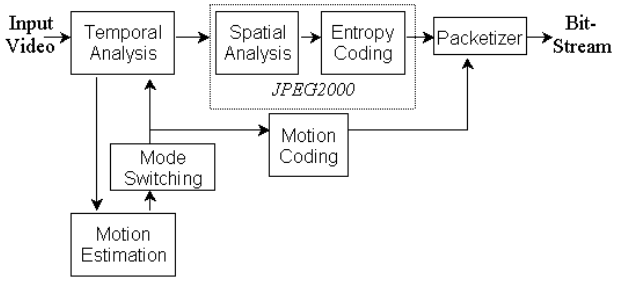

Figure 1: TWAVIX architecture overview

Motion is estimated thanks to a fast, eighth-pel accurate, hierarchical variable-size block-matching algorithm, from 256x256 blocks down to 4x4 blocks. Multi-level motion field's quad-trees are then pruned according to rate-distortion slopes of each node.

Depending on the sequence content, temporal analysis can consist in Haar MCTF, Backward/Forward Prediction, or Intra Coding. Regarding spatial analysis and texture coding, TWAVIX is coupled with JPEG2000 VM8.0 implementation.

Note that rather than adjust an average bit-rate over the entire encoded sequence, TWAVIX performs rate allocation independently for each GOF. This is far more realistic for applications involving varying bandwidth, such as "*Video streaming over heterogeneous IP networks*" and "*Mobile streaming video*" (Cf. [2]).

As regards motion coding, TWAVIX classically uses a non-scalable context-based adaptive arithmetic coder, inspired from [8]. Motion fields are computed and encoded at full spatial resolution into a single-layer bit-stream.

2.2. Spatial scalability and motion compensation

Let us stress that even if motion estimation is performed at original resolution, TWAVIX, unlike many scalable codecs in the literature, does not systematically reconstruct full resolution frames before performing temporal synthesis. All computations are processed at the real decoded resolution, by rescaling original motion fields quad-tree structures and vectors components. This choice is motivated by reality of applications (Cf. [2]) : we don't imagine a cellular phone or a PDA can afford to perform motion compensation at SD or HD resolution.

This means that motion compensation at decoder side will not systematically be processed at the same resolution as at encoder side. Consequently sub-pel interpolation demands a special care for each sub-resolution. Actually we use 8-tap FIR filters at original resolution and bilinear interpolation at lower resolutions.

3. SCALABLE MOTION CODING

In state-of-the-art coding schemes, motion parameters are usually coded losslessly as side-information. The tradeoff between the volume of motion information and the efficiency of energy compaction has been widely recognized. In non-scalable coders, various techniques have been used to optimize the amount of motion information according to a target bit-rate ; but in scalable coding there is an infinity of targeted bit-rates.

3.1. Natural motion scalabilities

Ideally, at decoder side, rate-adapted motion subsets should be available to optimize video quality. However, in a (t+2D)WT scheme, temporal filtering has been performed once using full-resolution motion fields (Cf. Fig. 2). The point is therefore to deduce subsets from these original-resolution motion fields, that will allow the decoder to preserve a reasonable motion/texture ratio, without penalizing too much motion compensation quality.

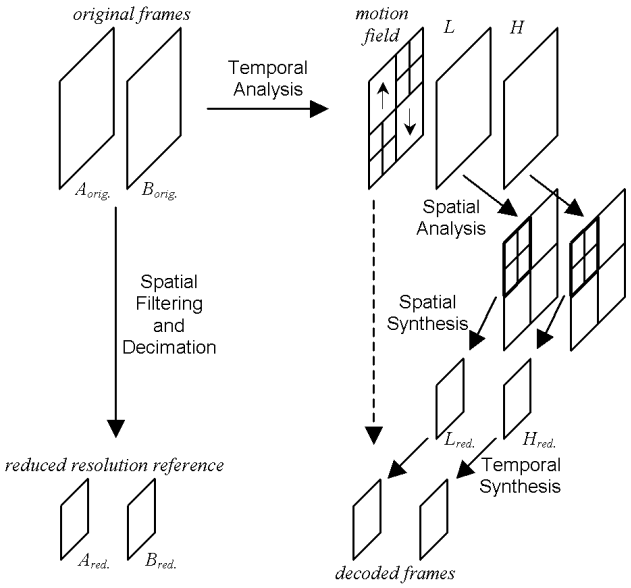

Figure 2: Spatial scalability in a (t+2D)WT framework

At low bit-rates, video is usually decoded at a reduced spatial resolution. So, on the one hand high-precision motion vectors are virtually useless, and on the other hand smallest blocks tend to vanish.

These statements could motivate to apply on motion information the same coding techniques as those used for texture (spatially-scalable transform and progressive coding), like in [6]. Let us first point out that unlike the triangular-mesh motion model used by Taubman & Secker, our variable-size block-based motion description does not suit a spatial transform, but presents inherent sparseness properties thanks to pruning.

Let us moreover note that a three-resolution scenario (e.g. QCIF-CIF-SD) is not sufficient to take advantage of block-size scalability. Even smallest (4x4) blocks of SD resolution do not disappear at QCIF resolution, at least for luminance. In addition, corrupting the block structure induces annoying visual artifacts. So in such a configuration, we can only rely on accuracy scalability.

3.2. Accuracy-scalable motion coding

Having investigated the impact of accuracy at decoder side, it appears that its usefulness decreases with spatial resolution. Once spatially filtered and decimated, temporal low and high frequencies do not benefit from the sub-pel accuracy which has been used at original resolution during temporal analysis.

This leads us to partitioning the bit-stream into accuracy layers. But unlike in [6] where each bit-plane of motion is

divided in several coding passes, we only introduce as many truncature points as decoded resolutions, in order to confine scalability over-cost.

For a three-level scenario, the optimal layering seems to consist in two enhancement layers of one-level accuracy, and a base layer of the corresponding approximate field. Figure 3 shows an example corresponding to ⅛ pixel-accurate motion estimation on SD video source.

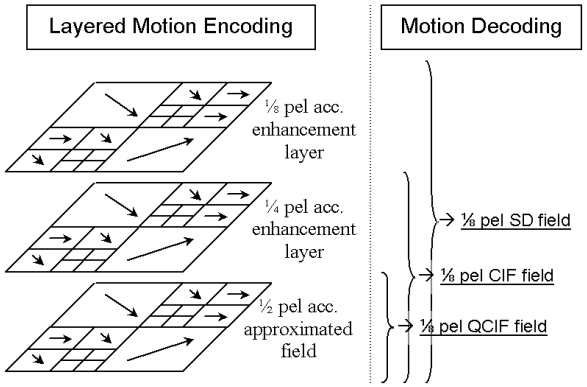

Figure 3: Example of three-level accuracy-scalable motion coding and decoding

After having encoded once the quad-tree structure, each vector's prediction residue of the base layer is encoded with our context-based adaptive arithmetic coder inspired from [8]. Then enhancement layers are encoded successively, without prediction coding since these layers can be assimilated with noise.

This simple and systematic technique allows to perform the same level of sub-pel interpolation through all decoded resolutions, while saving bit-budget for lower resolutions. At full resolution, a certain overcost is observed in comparison with non-scalable coding. This can legitimately be interpreted as the cost of scalability. Indeed, scalability inevitably lowers prediction and entropy coding efficiency.

Last, it is noteworthy that among the number of recently proposed scalable motion solutions, our work is similar to the AGP method independently developed by Wu, Golwelkar & Woods (Cf. [9]).

4. RESULTS

Results provided in this section correspond to CfP scenario1 (Cf. [1]). They are obtained by encoding once 704x576 60fps CITY, CREW and ICE sequences, then decoding them at the various bit-rates, frame-rates and resolutions described in Table 4.

width	height	frames/s	Kbit/s
176	144	15.0	64
176	144	15.0	128
352	288	15.0	192
352	288	30.0	384
352	288	30.0	750
704	576	30.0	1500
704	576	60.0	3000
704	576	60.0	6000

Table 4: CfP scenario 1 spatio-SNR-temporal scalability tests

For reasons of brevity, we present here average PSNRs on luminance component, with, for clarity over the wide bit-rate range, a logarithmic abscissa scale (see Figure 5).

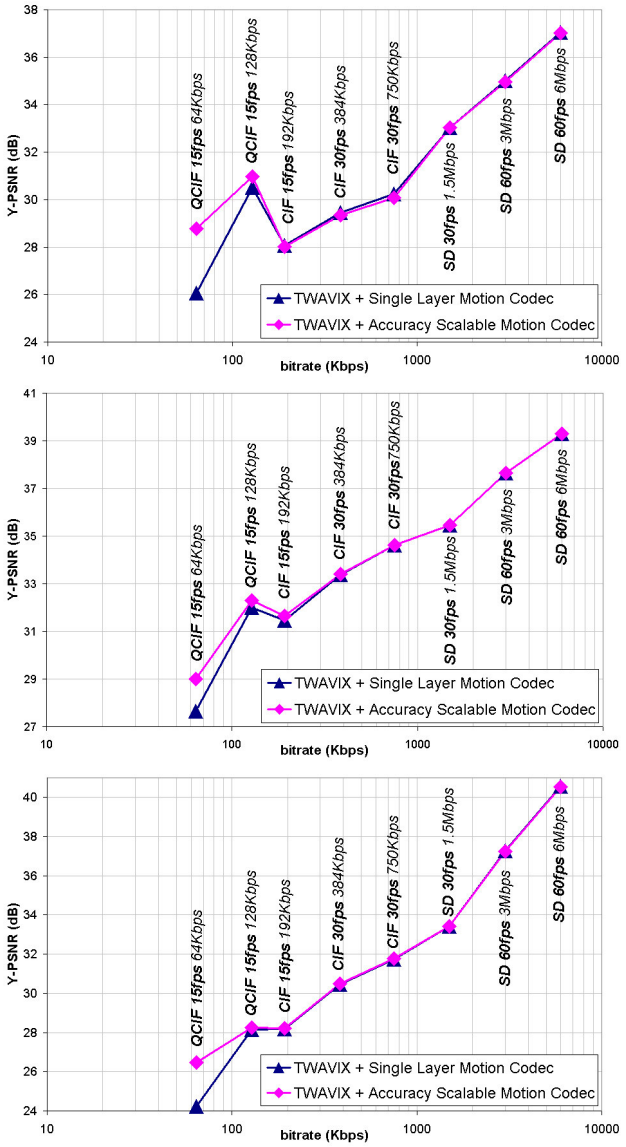

Figure 5: PSNR results for CITY (top), CREW (middle), and ICE (bottom) sequences

As one shall notice, there are five different spatio-temporal configurations in Table 4, namely QCIF-15fps, CIF-15fps, CIF-30fps, SD-30fps and SD-60fps. There are therefore five different reference sequences. These references have been defined in CfP procedure. For lower spatial resolutions, these "original" sequences are obtained by down-sampling using normative MPEG-4 filters. For lower frame-rates, reference sequences are obtained by frame-skipping, keeping even frames and discarding odd ones. Note that in terms of PSNR, these specifications do not favor (t+2D)WT solutions, which perform temporal filtering instead of rough decimation, and spatial low-pass filtering that does not correspond to MPEG-4 filters.

However, it can be observed that the curves corresponding to non-scalable motion coding exhibit satisfying results with regard

to state-of-the-art scalable coding. Indeed, in this configuration, TWAVIX can supply decoded video at all bit-rates while preserving SD high quality. In addition, our novel accuracy-scalable motion codec improves significantly performances at lowest (QCIF) resolution. At intermediate resolution (CIF), where the scalability over-cost is most critical, some improvement of the order of 0.1dB~0.2dB is perceptible, depending upon sequences. Last, at original resolution and high bit-rates, no penalty is observed.

As an example, Figure 6 compares motion budgets in percentage for ICE sequence, using our accuracy-scalable motion coder and our single-layer motion coder.

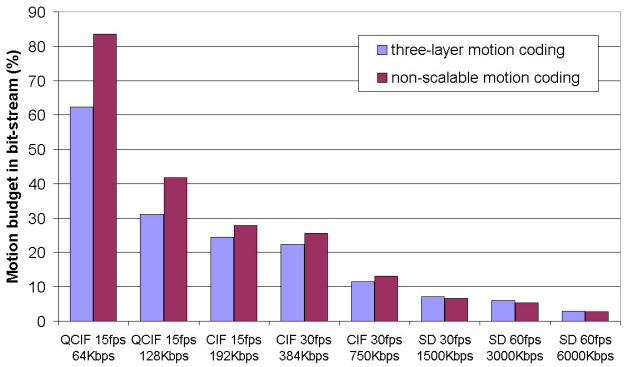

Figure 6: Motion bit-budget percentage in global bit-stream

Finally, Figure 7 illustrates for ICE and CITY sequences the visual quality gain that can be obtained with our scalable motion codec in comparison with a non-scalable one, at resolution QCIF 15fps, 64Kbps.

Figure 7: ICE (top) & CITY (bottom) QCIF 15fps 64Kbps, with non-scalable motion codec (left), and with accuracy-scalable motion codec (right)

5. CONCLUSION

A new scalable video coding scheme has been presented, that introduces scalable coding of motion in addition to full-scalable coding of texture. Motion codec principle consists of a layer partitioning according to accuracy and its adjustment to the level of spatial scalability. Although simple, this technique allows to cover a very wide range of bit-rates and improves significantly video quality at lower spatial resolutions, without any noticeable penalty at high bit-rates and full resolution.

6. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11, "Registered Responses to the Call for Proposals on Scalable Video Coding", *MPEG Document M10569*, Muenchen, March 2004.

[2] ISO/IEC JTC1/SC29/WG11, "Requirements and Applications for Scalable Video Coding", *MPEG Document N6052*, Gold Coast, October 2003.

[3] H.-M. Hang, S.S. Tsai, and T. Chiang, "Motion Information Scalability for MC-EZBC : Response to the Call for Evidence on Scalable Video Coding", ISO/IEC JTC1/SC29/WG11, *MPEG Document M9756*, Trondheim, July 2003.

[4] G. Boisson, E. François, D. Thoreau, and C. Guillemot, "Motion-Compensated Spatio-Temporal Context-Based Arithmetic Coding for Full Scalable Video Compression", *Picture Coding Symposium*, Saint Malo, France, April 2003.

[5] Y. Andreopoulos, M. Van der Schaar, A. Munteanu, J. Barbarien, P. Schelkens, and J. Cornelis, "Complete-To-Overcomplete Discrete Wavelet Transforms for Scalable Video Coding with MCTF", *Proc. SPIE/IEEE Visual Communication and Image Processing*, Lugano, Switzerland, July 2003.

[6] D. Taubman and A. Secker, "Highly Scalable Video Compression With Scalable Motion Coding", *Proc. IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003.

[7] G. Marquant, J. Viéron, G. Boisson, P. Robert, E. François and C. Guillemot, "Response to the Call for Evidence on Scalable Video Coding Advances", ISO/IEC JTC1/SC29/WG11, *MPEG Document M9784*, Trondheim, July 2003.

[8] D. Marpe, G. Blättermann, G. Heising, and T. Wiegand, "Video Compression Using Context-Based Adaptive Arithmetic Coding", *Proc. IEEE International Conference on Image Processing*, Thessaloniki, Greece, September 2001.

[9] Y. Wu, A. Golwelkar, and J.W. Woods, ISO/IEC JTC1/SC29/WG11, "MC-EZBC Video Proposal from Renssealer Polytechnic Institute", *MPEG Document M10569/S15*, Muenchen, March 2004.