

MOTION-COMPENSATED SPATIO-TEMPORAL CONTEXT-BASED ARITHMETIC CODING FOR FULL SCALABLE VIDEO COMPRESSION

Guillaume BOISSON¹, Edouard FRANCOIS¹, Dominique THOREAU¹, and Christine GUILLEMOT²

¹-THOMSON multimedia R&D France,

1 Avenue de Belle Fontaine, CS17616, 35511 Cesson-Sévigné Cédex, France
Email: guillaume.boisson / edouard.francois / dominique.thoreau @thomson.net

²-IRISA, Campus Universitaire de Beaulieu, 35042 Rennes Cédex, France

Email: christine.guillemot@irisa.fr

ABSTRACT

This paper is dedicated to the design of a fully scalable motion-compensated spatio-temporal subband video coding scheme. The main problem addressed is how to allow for a scalable capture and representation of the motion information and, at the same time, efficiently exploit the temporal correlation in the sequence. Two motion-compensated spatio-temporal decomposition structures are analyzed. Their respective amenability and limitations to provide an efficient and fully scalable representation of the information (both motion and texture) are investigated. This analysis led us to retain in the sequel a spatio-temporal signal decomposition approach proceeding first with a spatial decomposition. Two approaches for exploiting temporal dependencies in the different spatial bands are then studied. The first approach relies on classical motion-compensated temporal filtering, while the second uses motion-compensated spatio-temporal context-based arithmetic coding.

1. INTRODUCTION

The advent of heterogeneous communication infrastructures (such as the Internet and wireless networks) with time-varying capabilities and non guaranteed data delivery brings about new challenges in the design of compression systems. Low rate compression remains a largely sought capability, however this is not the only feature required. In order to optimize the end-to-end quality of service of the delivery chain, the compression scheme must allow for a flexible and dynamic adaptation of the compressed streams to network conditions that may vary in time. The concept of scalability is a key feature to fulfill these new requirements. Approaches based on motion-compensated spatio-temporal signal decomposition have thus gained attention as viable alternatives to classical predictive techniques for scalable video representation [1] [2] [3] [4] [5] [6] [7] [8].

The notion of full scalability refers here to the support of both spatial, SNR and temporal scalability. Most scalable approaches suffer however from limitations inherent to a non-adapted representation of motion information. E.g. in the solutions proceeding first with a temporal analysis followed by a spatial analysis, the motion is estimated on the full signal resolution. The resolution of the motion may be unduly too high if a lower texture resolution is used. Similar misadjustments are identified with SNR

scalability. In order to overcome the above limitations, we consider an alternative coding structure in which the spatial analysis is performed first, as already proposed in [9], [10], [11], and in [12] for digital cinema. This architecture inherently offers a multi-level capture of the motion that will be best adapted to the texture resolution retained. This scheme then brings about another issue: How to best exploit the remaining temporal correlation in the high frequency spatial bands. One solution consists in performing a dedicated motion-compensated temporal filtering on the successive frames of spatial details, as proposed e.g. in [11]. Besides, we aim to avoid the transmission of additional motion fields that would induce a penalizing overcost at low bit-rate. Therefore we investigate a mean of exploiting the temporal correlation remaining in the high spatial frequency bands by exploiting the motion fields estimated on low frequency bands. In this case, motion-compensated temporal filtering may not be the most appropriate solution, because of the inadequacy between the motion fields and the spatio-temporal variation of wavelet coefficients. Indeed, the discrete wavelet transform is not shift-invariant. As a result, coefficients corresponding to motion-connected spatial areas may have very different values, and significant energy may remain in the temporal high frequency band.

In order to overcome this difficulty, we consider a second solution based on motion-compensated spatio-temporal context-based arithmetic coding (MC-STAC). The contexts defined take into account both a spatial and a temporal neighborhood with integer and sub-pixel displacements. The capability of the approach to account for the shift-variance property of the spatial transform in presence of displacements has been investigated. The compression performances have been evaluated by considering only the spatial low horizontal - high vertical frequency band (s-LH band) in a GOP of 8 images. The approach, with and without motion compensation (using the motion field estimated on the spatial low frequency bands), has been first compared against a plain spatial context-based arithmetic coding. The interest of the motion-compensated spatio-temporal contexts is clearly shown against pure spatial contexts. A first rate-distortion evaluation on a group of 8 s-LH bands does not reveal real gain versus the approach based on motion-compensated temporal filtering (MCTF). However, the system does not incorporate yet any rate-distortion optimization. Also, the approach could be advantageously coupled with an MCTF filtering solution in order to handle show regions where MCTF leads to high residual energy in high frequency bands (e.g. unconnected

regions).

2. FULL SCALABILITY IN THREE DIMENSIONAL SUBBAND CODING

Classical video subband coding schemes, as the well-known MC-EZBC [7], follow the structure depicted in Fig.1. The video signal is first decomposed with a motion-compensated temporal filtering (MCTF), in the temporal analysis. Then the resulting temporal frequency bands are further decomposed into spatio-temporal subbands by a two-dimensional wavelet transform, during spatial analysis.

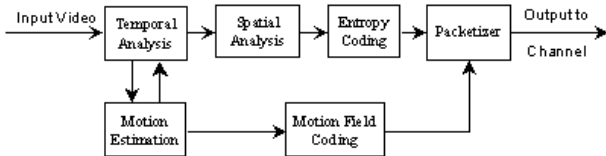


Fig. 1. Basic structure of MC-EZBC

Let us focus now on the scalability of such a scheme. The motion compensated temporal filtering applied on the video sequence produces temporal bands which can be regarded as embedded subsets corresponding to increasing frame-rates. The spatial analysis extracts from the texture different spatial frequency bands at successive resolutions. Last, the bit plane representation of the quantized symbols allows the construction of a progressive bit-stream. Thus, such a (2D+t) SBC scheme is often presented as being full-scalable. However the motion analysis raises specific issues with respect to the targeted full-scalable decoding of the video, issues that we try to address below.

2.1. Motion Estimation and Scalabilities in (2D+t)SBC

2.1.1. Temporal scalability

In a MCTF scheme, the motion estimation is performed at each iterative step, within a given temporal level (Cf. Fig.2). As a consequence the motion fields are inherently adapted to the temporal resolution considered. As such, one may consider that the motion information is “temporally scalable”.

2.1.2. Spatial scalability

Let us now consider spatial scalability. In (2D+t) SBC schemes, since the temporal analysis stage is first performed at the original resolution, the motion information turns out to be not “spatially scalable”. Indeed, in order to reconstruct the sequence at a reduced spatial resolution, one makes use of motion fields estimated on texture at higher (original) resolution. This over-precision of the motion fields induces an extra description cost in comparison with a single-layer coder working directly at the lower resolution. The problem is hence how to always guarantee a complete compatibility between motion resolution and texture resolution. This problem can be best addressed by considering instead the use of an alternative scheme for 3D-SBC, in which spatial analysis is performed first. In the sequel this architecture is referred to as the (t+2D) scheme. This is discussed in detail in section 2.3.

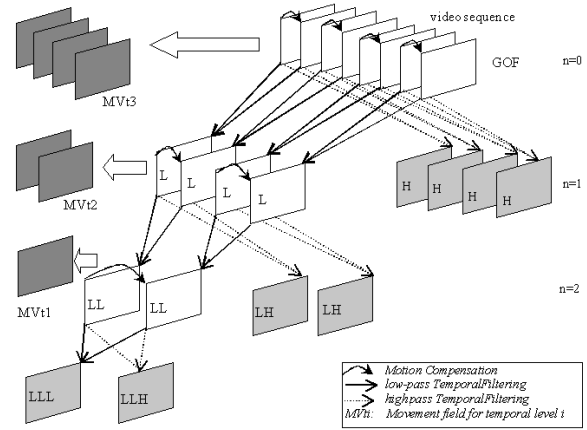


Fig. 2. Typical 3-level MCTF scheme

2.1.3. Fidelity scalability

Fidelity scalability commonly deals with the number of bit planes of texture to be decoded, which is generally independent from the motion estimation step. For that scalability, the motion information can be seen as an incompressible overhead, which becomes more and more prejudicial at lower bit-rates. This is the reason why a fine grain scalable coder, tested against a single-layer coder at several bit-rates, tends to be less competitive at lower bit-rates.

A fine-grain scalable (FGS) representation of the motion is not addressed here. However, further study may be dedicated to design a representation of motion fields as a multi-level tree whose vector density or accuracy would vary in function of the quantization step of a given quality layer.

2.2. (t+2D) Subband Coding approach

In order to solve this full - or at least spatial - scalability problem, one can modify the classical (2D+t) SBC scheme by simply swapping the temporal and spatial analysis stages, as illustrated in Fig.3.

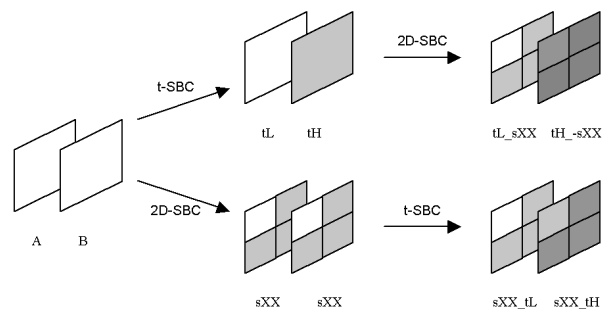


Fig. 3. Spatio-temporal decomposition in (2D+t) and (t+2D) SBC schemes

Each spatial band can be further decomposed with MCTF, using the motion field estimated on the low frequency band, or a specific motion field estimated separately on it. The first solution has been implemented and Fig.4 compares the compression

ratio obtained for the sLH-tH band of the described (t+2D)SBC scheme, with the compression ratio of the tH-sLH band of the classical symmetrical (2D+t)SBC scheme. Truncated Haar filtering is used, with no special handling for unconnected areas. The different levels of distortion are tuned by selecting a given number of bit planes.

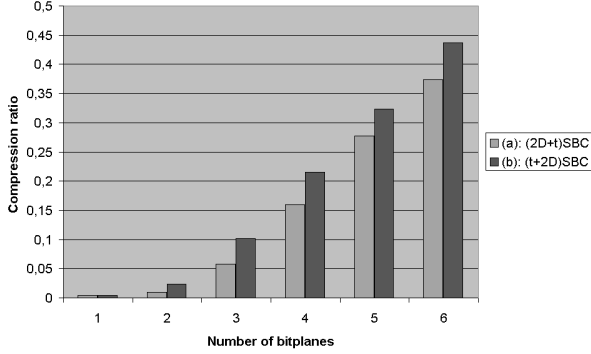


Fig. 4. Comparative compression performance for (a): (2D+t)SBC (temporally scalable), and (b): (t+2D)SBC (spatially and temporally scalable)

It can be observed that in these conditions MCTF is not efficient on the high spatial frequency bands in comparison with in classical (2D+t)SBC schemes. Indeed, the discrete wavelet transform is not shift-invariant. Two motion-connected pixels can have very different wavelet coefficients, according to the direction and the magnitude of displacement, the direction and the length of the spatial filter. Hence, in general, if a reasonable constraint of cost is taken into account, the DFD resulting from a matching motion estimation performed on high spatial frequency band remains energetic. Therefore, to avoid an additional overcost of motion information, a dedicated motion estimation performed on the high spatial frequency bands may not be the most appropriate solution. Neither is the reuse of a pixel-world motion field, because the filtering results in a significant amount of high frequency information. So classical ME-MCTF turns out to be dedicated to natural pixel domain - or low frequency band.

More interesting coding schemes have been proposed, that first apply a spatial transform to video frames, and exploit temporal correlation between successive frames of each spatial band by prediction [10] or by temporal filtering [9] [11], resorting to an over-complete transform to overcome the shift-variance problem. As an alternative to these efficient techniques, we rely instead on an extension of the classical context-based arithmetic coding techniques.

2.3. Exploiting temporal correlation in high spatial frequency subbands

Natural video frames usually comprise smooth regions, nearly motion-invariant, separated by edges that are typically smooth curves. In the wavelet domain, this is represented by large positive/negative coefficients varying along edges, among large regions of zeroes. So, in a given frame, most zeroes (or small values) have a neighborhood of zeroes (small values), whereas most large values have a neighborhood comprising other large values. The high efficiency

of still-image codec EBCOT [13] and video codec MC-EZBC [7] are mainly due to the exploitation of these local spatial dependencies by context-based adaptive arithmetic coding.

Similar temporal dependencies exist between successive frames of spatial details (i.e spatial high frequency bands) along motion trajectories. Fig.5 shows the spatial frequency band capturing the horizontal details, resulting from a wavelet decomposition of two successive frames. One can observe that the remaining energy and correlation are relatively high. In the sequel we present a motion-compensated spatio-temporal context-based arithmetic coder that takes benefit of all these dependencies.

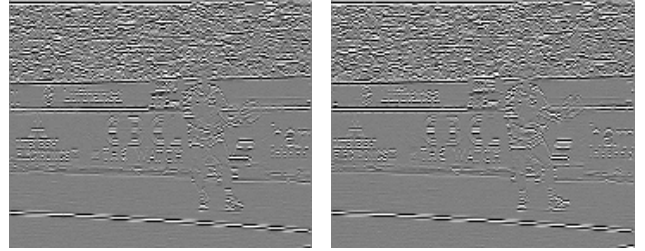


Fig. 5. Low-High spatial frequency bands extracted from two successive frames of the Stefan sequence

3. PROPOSED SCHEME

3.1. Overall architecture

In order to allow for a full-scalable signal representation, we consider an architecture in which spatial decomposition is performed first, preserving spatial scalability of motion fields as explained above. Let us consider the three-level scalable coder {progressive Standard Definition (SD) - CIF - QCIF}, depicted in Fig.6.

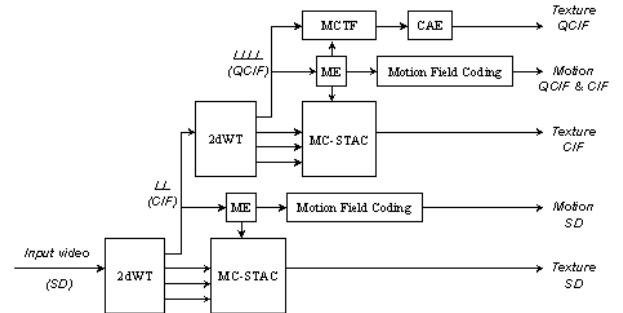


Fig. 6. Architecture of the scalable coder supporting QCIF, CIF, and progressive SD spatial resolutions

Motion estimation is performed in a classical manner on the Low-Low frequency band at each level of spatial analysis, with a motion rate constrained hierarchical variable size block-matching estimator. The classical MCTF scheme is only performed on the lowest spatial subband, here the LLLL band, corresponding to the QCIF resolution. Every high spatial frequency band is coded with a Motion Compensated Spatio-Temporal Arithmetic Coder (MC-STAC), with adaptive statistic gathering, using the adapted motion fields.

3.2. Motion Compensated Spatio-Temporal Arithmetic Coding (MC-STAC)

In order to deliver a fine grain scalable bit-stream, each high spatial frequency band is encoded per bit-plane, with three different coding primitives, called in function of the coefficients status :

- Significance Coding (or Zero Coding), for insignificant coefficients,
- Sign Coding, for coefficients just tested significant,
- Magnitude Refinement for significant coefficients.

Hence, the information involved in coding and context modeling is binary : significant/insignificant for Significance Coding and Magnitude Refinement, positive/negative for Sign Coding.

3.2.1. Motion-Compensated Spatio-Temporal Context Modeling

In addition to the conventional spatial neighborhood, we consider a motion-compensated temporal context, inspired from [6].

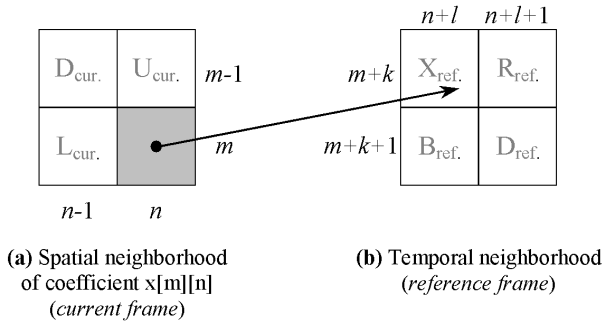


Fig. 7. Motion-compensated spatio-temporal context

Let x_t and $x_{t-\Delta t}$ be respectively the current and the reference frames of a given high spatial frequency subband. Let $x_t[m][n]$ be the coefficient to be encoded. For that coefficient, the spatial context consists in its left, upper, and diagonal upper-left neighbors in the current frame:

- $L = x_t[m][n-1]$,
- $U = x_t[m-1][n]$,
- $D_c = x_t[m-1][n-1]$.

Let us note $\mathbf{v}_{t \rightarrow t-\Delta t}[m][n]$ the displacement vector associated to that coefficient. For simplicity, we will name dx and dy the horizontal and vertical components of that vector. The displacements dx and dy can be a priori non-integer. Let l and k be respectively the integer parts of dx and dy , and let dl and dk be their fractional parts:

$$\mathbf{v}_{t \rightarrow t-\Delta t}[m][n] = \begin{pmatrix} dx \\ dy \end{pmatrix} = \begin{pmatrix} l + dl \\ k + dk \end{pmatrix} \quad l, k \in \mathbf{Z}^2$$

The motion-compensated temporal context contains information from the reference frame, as shown in Fig.7, and denoted:

- $X = x_{t-\Delta t}[m+k][n+l]$,
- $B = x_{t-\Delta t}[m+k+1][n+l]$,

- $R = x_{t-\Delta t}[m+k][n+l+1]$,
- $D_r = x_{t-\Delta t}[m+k+1][n+l+1]$.

The complete context, gathering spatial and motion-compensated neighborhood conditioning information, is depicted in Fig.8.

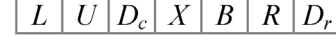


Fig. 8. Basic context indexing

3.2.2. Additional context parameters

The context can be augmented by considering sub-pixel displacements. The fractional parts of motion vector components can thus be inserted in the context under the form:

- $frac_{dx} = \frac{dl}{accuracy} = \frac{dx-l}{accuracy}$
- $frac_{dy} = \frac{dk}{accuracy} = \frac{dy-k}{accuracy}$, $accuracy \in \mathbf{Q}$

where the term *accuracy* refers to the precision of the estimated displacement vectors, under the form of a rational number. For example, in the case of a quarter-pixel accuracy, the set of possible values for $frac_{dx}$ and $frac_{dy}$ is $\{0,1,2,3\}$.

In addition, to take into account the characteristic of “shift variance” of the spatial discrete wavelet transform, the parity (i.e. parity of the integer part) of the motion vector components, can also be added to the context:

- $par_{dx} = l \pmod{2} = Ent(dx) \pmod{2}$
- $par_{dy} = k \pmod{2} = Ent(dy) \pmod{2}$

Actually, only the parity of a given displacement component needs to be considered for a given frequency band. E.g., only the parameter par_{dy} needs to be considered for the s-LH band. Similarly only the parameters par_{dx} and $par_{max(dx,dy)}$ need to be taken into account for respectively the s-HL and s-HH spatial frequency bands. These parameters are shown in Fig.9. The complete context model is depicted in Fig.10.

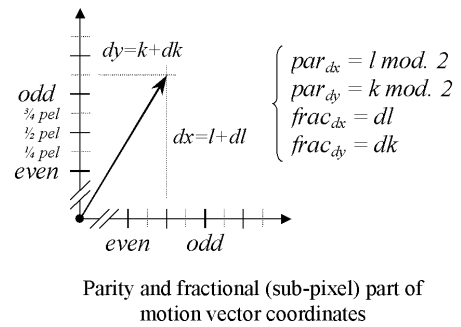


Fig. 9. Additional parameters of the motion-compensated spatio-temporal context

This definition leads to a large number of possible contexts. With a quarter-pixel accuracy, twelve bits are needed for context indexing, as shown in Fig.10. It was nevertheless observed that,

probably since we treat large subbands (176x144 for CIF high frequencies, 352x288 for SD high frequencies), and because we do not split each band into small independent blocks, the convergence of the statistics learning phase is relatively fast.

However, overall, it has been observed that the displacement parity parameters do not bring any significant gain. A small gain has only been revealed for sign coding. Also, the gain that one might expect from the coefficient phase information brought by motion sub-pixel accuracy is counterbalanced by the exponential increase of the number of contexts. This in turn leads to some context dilution with a corresponding impact on the rate. Notice that, to avoid context dilution, a fusion strategy could be envisaged.

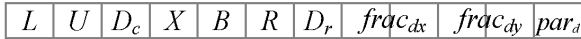


Fig. 10. Extended context indexing for quarter pixel accuracy

3.3. Results

The compression performances of the MC-STAC technique have been evaluated by considering only the spatial low horizontal - high vertical frequency band (s-LH band) in a GOP of 8 images, with one-level decomposition from CIF to QCIF (upper part of the Fig.6 scheme). The approach, with and without motion compensation (respectively referred to as MC-STAC and STAC), has been first compared against a plain spatial context-based arithmetic coding. When motion is considered, the motion fields estimated on the spatial low frequency bands are being used, in order to avoid sending extra motion information.

Fig.11 and 12 show the improvements in compression obtained with the STAC and MC-STAC techniques versus a plain spatial context-based arithmetic coding (intra-frame AC) for two sequences *foreman* and *stefan*. One can observe the impact of the motion in the contexts used in the arithmetic coding procedure, especially with the sequence *stefan* with high motion.

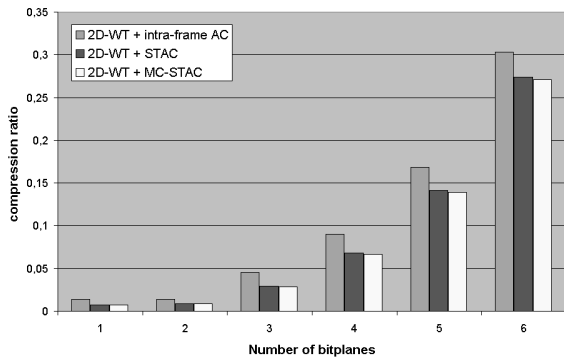


Fig. 11. Low-High band of Foreman sequence : compared compression ratio

The rate-distortion performance of the approach has been evaluated against a solution using MCTF on the S-LH frequency bands.

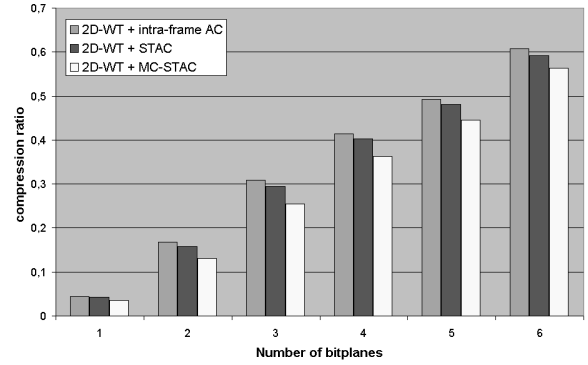


Fig. 12. Low-High band of Stefan sequence : compared compression ratio

In the tests, the rate has been adjusted by only selecting sets of bit planes in the different temporal frequency bands for the approach based on MCTF or in the consecutive s-LH bands of the GOP for the MC-STAC based approach. Fig.13 shows the MSE measured on the group of s-LH frequency bands. In this particular testing scenario, the MC-STAC based solution has not been shown to outperform the MCTF based solution. However, the results do not take into any local rate-distortion optimization of rate allocation. In addition both approaches can be advantageously combined by using an adaptive mode selection based on connected/unconnected pixel criteria, i.e. in regions where MCTF would lead to high residual energy in high frequency bands. The spatio-temporal arithmetic coder (without motion) can also be applied on temporal frequency bands resulting from the application of MCTF on the group of spatial frequency bands.

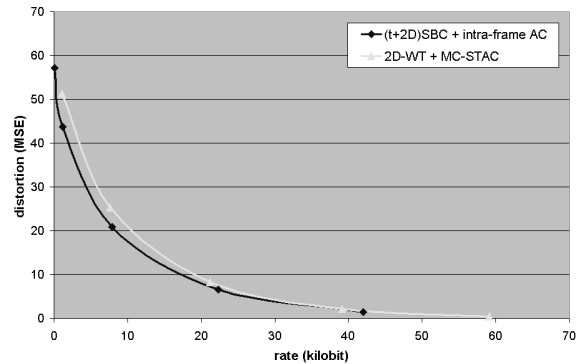


Fig. 13. Low-High band of Stefan sequence : rate-distortion curves

In addition, conditional distributions, as seen above, are updated along with bit-plane coding. However, the initialization of the distribution is crucial, especially if the number of contexts is large. In the results above, the initialization relies on a uniform distribution. This could be improved by using an average distribution estimated on a training set of sequences. The high number of contexts presents some risk of a so-called "context dilution" leading to a close to uniform conditional distribution with an obvious impact on the rate. Some heuristic context fusion could also

be introduced to avoid this phenomenon. Finally, spatial and temporal dependencies may be better modeled by considering sets of bit-planes rather than separate bit-planes (i.e. with N-ary coding instead of binary coding), at the expense however of a coarser grain scalability.

4. CONCLUSION

In this paper, we have investigated different spatio-temporal analysis structures for a fully scalable representation and coding of video signals. In this context, we have then privileged solutions based on a spatial analysis followed by different techniques to exploit temporal redundancy between consecutive temporal frequency bands. As such we have considered a motion compensated spatio-temporal arithmetic coder (MC-STAC) as an alternative solution to motion-compensated temporal filtering for processing spatial high frequency bands. The importance of the motion information in spatio-temporal context modeling has been evidenced. Although, the MC-STAC does not outperform approaches based on MCTF, both techniques could be coupled advantageously by using adaptive selection on the basis of high temporal frequency energy or of connected/unconnected region criteria.

5. REFERENCES

- [1] J.R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Trans. Image Proc.*, vol. 3, no. 5, Sept. 1994.
- [2] S.-J. Choi and J. W. Woods., "Motion compensated 3-d subband coding of video," *IEEE Trans. Image Proc.*, vol. 8, no. 2, Feb. 1999.
- [3] B. Pesquet-Pospescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," *Proc. ICASSP 2001*, May 2001.
- [4] A. Secker and D. Taubman, "Motion-compensated highly scalable video compression using an adaptive 3d wavelet transform based on lifting," *Proc. IEEE Intl Conf. on Image Processing 2001*, Oct. 2001.
- [5] L. Luo, J. Li, S. Li, Z. Zhuang, and Y.-Q. Zhang, "Motion-compensated lifting wavelet and its application in video coding," *Proc. IEEE Intl Conf. on Image Processing 2001*, Oct. 2001.
- [6] J. Xu, Z. Xiong, S. Li, and Y.-Q. Zhang, "Three-dimensional embedded subband coding with optimized truncation (3d-escot)," *Applied and Computational Harmonic Analysis*, vol. 10, pp. 290–315, 2001.
- [7] P. Chen and J. W. Woods, "Improved mc-ezbc with quarter-pixel motion vectors," *ISO/IEC JTC1/SC29/WG11, MPEG2002/8366*, May 2002.
- [8] D. S. Turaga and M. Van der Schaar, "Unconstrained motion compensated temporal filtering," *ISO/IEC JTC1/SC29/WG11, MPEG2002/8388*, May 2002.
- [9] H. W. Park and H. S. Kim, "Motion estimation using low-band-shift method for wavelet-based moving-picture coding," *IEEE Trans. Image Proc.*, vol. 9, no. 4, pp. 577–587, April 2000.
- [10] Y. Andreopoulos, A. Munteanu, and al., "Wavelet-based fine granularity scalable video coding with in-band prediction," *ISO/IEC JTC1/SC29/WG11, MPEG2002/7906*, March 2002.
- [11] M. Van der Schaar, J. Ye, Y. Andreopoulos, and A. Munteanu, "Fully scalable 3-d overcomplete wavelet video coding using adaptive motion-compensated temporal filtering," *ISO/IEC JTC1/SC29/WG11, MPEG2002/9037*, Oct. 2002.
- [12] P. Chen and J. W. Woods, "Comparison of mc-ezbc and h26l tml 8 on diggital cinema test sequences," *ISO/IEC JTC1/SC29/WG11, MPEG2002/8130*, March 2002.
- [13] D. Taubman, "High performance scalable image compression with ebcot," *IEEE Trans. Image Proc.*, vol. 9, no. 7, pp. 1158–1170, July 2000.