Encoder and Decoder Side Global and Local Motion Estimation for Distributed Video Coding

Frederic Dufaux [#*], Touradj Ebrahimi [*]

[#] *Laboratoire de Traitement et Communication de l'Information (LTCI) – UMR 5141*
Télécom ParisTech and CNRS
F-75634 Paris Cedex 13, France
frederic.dufaux@telecom-paristech.fr

[*] *MultiMedia Signal Processing Group*
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
frederic.dufaux@epfl.ch, touradj.ebrahimi@epfl.ch

***Abstract*—In this paper, we propose a new Distributed Video Coding (DVC) architecture where motion estimation is performed both at the encoder and decoder, effectively combining global and local motion models. We show that the proposed approach improves significantly the quality of Side Information (SI), especially for sequences with complex motion patterns. In turn, it leads to rate-distortion gains of up to 1 dB when compared to the state-of-the-art DISCOVER DVC codec.**

I. INTRODUCTION

Digital media content is omnipresent in nowadays information society where fast and efficient access to information is paramount. This evolution has been possible thanks to the rapid and remarkable progresses in video coding technologies.

Over the last two decades, standardization efforts in MPEG and ITU-T have led to new technologies with ever improving coding performance. The resulting standards adopted throughout the years have been a cornerstone of the digital age experience, and the state-of-the-art H.264/AVC [1] is the latest outcome of this process.

In all MPEG and ITU-T video coding schemes, the encoder is in charge of exploiting the source statistics to achieve the most efficient compression, resulting in an asymmetric computational load where the encoder is significantly more complex than the decoder. For instance, the compression gains achieved by H.264/AVC are the result of an extensive analysis at the encoder in order to better represent the video signal. Namely, the encoder can choose from an ever growing number of coding modes, improving coding efficiency at the cost of much increased encoding complexity. While this incremental progression has brought significant achievements in the past, one may wonder for how long this path will continue to produce improved performance.

It is therefore a good time to reflect on the field of video compression. More specifically, it is legitimate to ask whether we are reaching a plateau in performance with the current standard architecture, or whether they are new promising theories and technologies which may bring significant breakthroughs in the near future.

Distributed Video Coding (DVC) has more recently emerged as a new coding paradigm. DVC finds its theoretical foundation in the Slepian-Wolf [2] and Wyner-Ziv [3] theorems. These results remained unexploited until the early 2000's when the first practical DVC schemes have been proposed [4][5]. The DVC paradigm presents a number of advantages: flexible partitioning of the computational complexity between the encoder and decoder, error resilience, codec-independent scalability and multi-view coding. DVC has gained a lot of interest over the last few years. Overviews of recent developments are presented in [6][7].

Most of the research activities on DVC have so far focused on outperforming conventional coding solutions under the specific constraint of very low encoding complexity, this feature being appealing for a number of upcoming up-link applications. For instance, the DISCOVER DVC codec [8] has reported some of best coding performance results at the moment. In particular, it consistently outperforms H.264/AVC Intra. For scenes with simple and uniform motion, it even outperforms H.264/AVC No Motion. However, for more complex scenes, its performance typically remains lower than H.264/AVC No Motion. Nonetheless, the encoding complexity advantage offered by DVC may be very short-lived due to exponentially increasing computing power as predicted by Moore's law.

In summary, substantial gains have been obtained with conventional coding by continuously adding more efficient analysis at the encoder, and hence resulting in a complex encoder – simple decoder arrangement. Conversely, new DVC codec designs have mainly focused on the opposite extreme, proposing advanced tools at the decoder, creating a simple encoder – complex decoder framework.

Challenging current thinking and as an avenue towards further coding advances, it has been proposed in [9] to develop a new class of codec architecture where both encoder and decoder are peers of equal importance and share the

MMSP'10, October 4-6, 2010, Saint-Malo, France.
978-1-4244-8112-5/10/$26.00 ©2010 IEEE.

workload burden. Thanks to Moore's law, which has been a driving force of technological changes in the late 20th and early 21st centuries, such a complex encoder – complex decoder architecture may quickly become acceptable for various application domains, as long as it provides a compelling value proposition and enable new products and services.

A similar path is taken in [10], which considers H.264/AVC and proposes to perform motion estimation both at the encoder and decoder for the coding of B frames. The scheme saves on the transmission of motion vectors and often achieves better prediction, leading to coding gains.

In [11][12], DVC schemes are presented discarding the low encoding complexity constraint and performing motion estimation both at the encoder and decoder. This paper follows a similar direction. Given that the effectiveness of DVC strongly depends on the correlation between the Side Information (SI) and the Wyner-Ziv (WZ) frame, we propose to perform motion estimation both at the encoder and decoder, combining global and local motion models. We show that the proposed architecture leads to better SI, resulting in coding gains. Although we do not explore the subject in this paper, it is important to underline that the proposed architecture also preserves the strong error resilience feature of DVC.

This paper is structured as follow. Related works are first reviewed in Sec. II. We then present the overall proposed system in Sec. III. The process to generate SI is described in more details in Sec. IV. Next, the performance of the proposed approach is assessed in Sec. V. Finally, Sec. VI draws conclusion and outlines future perspectives.

II. Related Work

In DVC, the quality of the SI to approximate the WZ frame has a great impact on coding efficiency. Most commonly, SI is estimated at the decoder by linear interpolation of motion vectors between consecutive reference frames. Various techniques have been proposed in order to improve SI generation.

Spatial smoothing and motion vectors refinement has been proposed in [13]. By providing motion fields closer to the true motion in the scene, the method results in better prediction. Motion compensated forward and backward extrapolation is introduced in [14], generating two SI which are exploited in the decoding process. Motion estimation with sub-pixel accuracy is considered in [15].

In [16], decoded bitplanes of the WZ frame are exploited to refine motion vectors. Several interpolation modes are also introduced. Motion compensated temporal interpolation is iteratively improved in [17] based on a partially decoded WZ frame. In [18], an iterative technique is proposed based on multiple SI and motion refinement. Based on error probability, the turbo decoder then determine which SI to select for each block. In the same way, a partially decoded WZ frame is used to improve SI generation in [19]. An enhanced motion compensated temporal interpolation method is also introduced.

In [11], a pixel domain DVC scheme is introduced, combining low-complexity encoder-side bitplane motion estimation with decoder-side motion compensated frame interpolation. Improvements are shown for sequences with fast and complex motion. Finally, [12] presents a DVC scheme where both the encoder and decoder cooperate to perform motion estimation. Results show that it reduces overall computational complexity while improving coding efficiency.

III. Proposed System

In this paper, we more specifically consider the DISCOVER DVC codec [8]. Input frames are first split into Group of Pictures (GOP). Key frames, corresponding to the first frame of each GOP, are conventionally encoded using H.264/AVC [1]. In turn, WZ frames undergo a DCT transform followed by uniform quantization. The quantized values are then split into bitplanes which go through a Low-Density Parity-Check Accumulate (LDPCA) encoder. At the decoder, SI approximating the WZ frames is generated from the previously decoded reference frames. SI is then used in the LDPCA decoder, along with the parity bits of the WZ frames requested via a feedback channel, in order to reconstruct the bitplanes, and subsequently the decoded video sequence. Hereafter, without loss of generality, we more specifically consider a GOP size of 2. Namely, odd and even frames are coded as key and WZ frames respectively.

In most DVC schemes, SI is generated at the decoder by motion compensated interpolation or extrapolation of previously decoded reference frames. Henceforth, it is very challenging to estimate an SI closely approximating the WZ frame, especially for scenes with complex motion. Moreover, such schemes assume that motion remains uniform in-between the reference frames, although this hypothesis often does not hold for complicated scenes.

In contrast, in the proposed DVC architecture, motion estimation and compensation is performed both at the encoder and decoder sides. Moreover, we consider the combination of global and local motion representations. The resulting codec is illustrated in Fig. 1.

A. Overview of SI Processes at the Encoder and Decoder

The following operations are performed at the encoder side. First, frame-based global motion estimation and compensation is performed between the current WZ frame and previous and next decoded key frames. It results in a first SI, denoted GMC SI. In parallel, block-based local motion estimation and compensation is applied between the previous and next decoded key frames, resulting in the MCTI SI. Finally, for each 16x16 MacroBlock (MB) of the WZ frame, the optimal prediction mode is determined by selecting the best approximation between GMC and MCTI SI.

The global motion parameters used in GMC, as well as the optimal MB mode selections, are transmitted to the decoder as supplementary information.

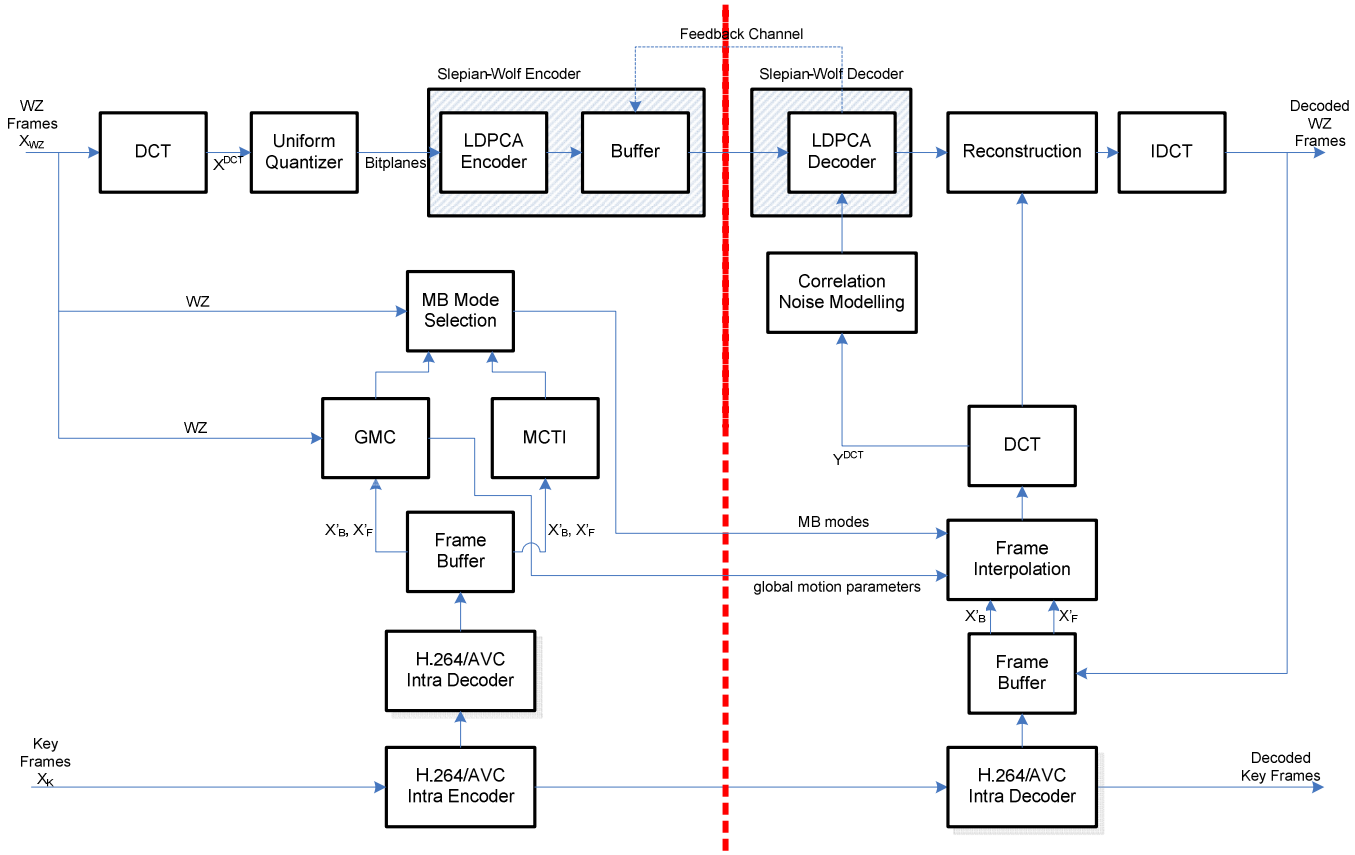

Fig. 1. Proposed DVC architecture.

At the decoder side, the global motion parameters are used to generate the GMC SI. Conversely, local motion estimation and compensation is performed again in order to produce the MCTI SI. Both SI are then fused using the optimal MB mode decisions in order to create the final SI.

As we will show hereafter, this DVC architecture allows improving the SI generation process by exploiting both global and local motion representations and selecting optimal MB mode decisions. Moreover, the bitrate needed to transmit supplementary information, namely the global motion parameters and the MB mode selections, remains marginal. Additionally, the proposed architecture preserves one of the essential features of DVC, namely the absence of a prediction loop. This prevents drifts in the presence of transmission errors and, along with the built-in joint source-channel coding structure, implies strong error resilience.

However, the rate-distortion performance gain is obtained at the expense of increased computational complexity at the encoder side, hence breaking the 'low encoding complexity' target commonly assumed in DVC research works.

IV. Side Information Generation

We now describe in more details the SI generation process taking place both in the encoder and decoder.

A. Global Motion Compensation

Using GMC, the global motion within the scene is modeled by a perspective transform, also known as homography. This model is valid whenever the scene can be approximated by a planar surface.

More specifically, two transforms H_1 and H_2 are computed between the WZ and previous key frame on the one hand, and between the WZ and next key frame on the other hand, as illustrated in Fig. 2. The GMC SI is then obtained by averaging both backward and forward predictions.

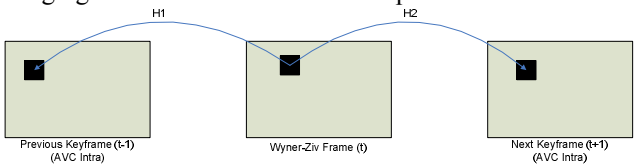

Fig. 2. Side information with the proposed GMC.

Note that, as the process is performed at the encoder, the WZ frame to be approximated is available and can be directly exploited in the minimization process, as explicitly detailed hereafter.

More precisely, the perspective transform is defined as

$$u_i = \frac{a_0 + a_2 x_i + a_3 y_i}{a_6 x_i + a_7 y_i + 1} \qquad v_i = \frac{a_1 + a_4 x_i + a_5 y_i}{a_6 x_i + a_7 y_i + 1}$$

where (u_i,v_i) denotes the pixel location in the WZ frame, (x_i,y_i) the corresponding position in the previous or next key frame, and a_0, a_1, …, a_7 the parameters of the transform.

The motion parameters are estimated by the global motion estimation technique introduced in [20]. More specifically, they are obtained by minimizing the expression

$$E = \sum_{i=1}^{N} \rho(\hat{I}(u_i, v_i) - I(x_i, y_i))$$

where $\hat{I}(u_i,v_i)$ and $I(x_i,y_i)$ represent the image pixel values of the WZ frame and previous or next key frame. In order to increase robustness to outliers, a truncated quadratic robust estimator is chosen for the metric ρ. The summation is carried over N pairs of pixels within the image boundaries. This non-linear problem is solved using the Levenberg-Marquardt gradient descent algorithm to iteratively estimate the parameters.

For each WZ frame, two set of parameters are calculated at the encoder, defining the backward and forward transform respectively. These global motion parameters are then transmitted to the decoder. An efficient way to represent this information consists in transmitting the differential coordinates of four points in the image defining the perspective transform [21]. Hence, the associated bitrate overhead remains marginal.

B. Motion Compensated Temporal Interpolation

An alternative SI is also generated by the conventional MCTI [13]. In this case, motion estimation is first performed between the previous and next decoded key frames. More specifically, block-based motion vectors are computed by block matching. Spatial motion smoothing [13] is then applied in order to further improve performance.

Unlike most DVC schemes, in the proposed approach, MCTI is applied both in the encoder and the decoder (see Fig. 1). On the one hand, this architecture allows selecting optimal MB modes at the encoder side for SI generation. On the other hand, motion vectors, which would otherwise represent a significant bitrate overhead, do not need to be transmitted.

C. Mode Selection

In previously proposed DVC codecs, the SI is most commonly generated solely at the decoder side, greatly limiting its effectiveness.

In the proposed architecture, two SI are available, based on GMC and MCTI respectively. By performing mode selection at the encoder, the WZ frame to be approximated is available and can be exploited to determine the optimal predictor. Straightforwardly, the latter is determined by selecting the best approximation between GMC and MCTI SI.

More precisely, the mode selection is made on a MB basis, where each MB corresponds to 16x16 pixels. One bit per MB is transmitted in order to signal the mode. On the one hand, this allows for local adaptation of the prediction, tailored to the scene content. On the other hand, the resulting bitrate overhead to transmit the MB decisions is negligible. For instance, for a video at QCIF resolution and 15 fps with a GOP of 2, the supplementary bitrate is only 99 bits per WZ frame or 0.742 kbps.

V. Experimental Results

Performance of the proposed approach is now assessed, taking the DISCOVER DVC scheme as reference [8]. The four test sequences "Foreman", "Soccer", "Coastguard" and "Hall Monitor", at QCIF resolution and 15 fps, are used for simulations. Note that only the luminance component is processed. We adopt the same test conditions as in [22].

A. Side Information

The quality of the SI resulting from the proposed approach is first assessed. For this purpose, we compare the SI obtained using three approaches: GMC, MCTI, and combined GMC-MCTI, as described in Sec. IV.A, IV.B and IV.C respectively. Fig. 3 shows PSNR as a function of the frame number for "Foreman" (corresponding to the rate-distortion point with quantization matrix Qi=8 [22]). Clearly, the proposed combined GMC-MCTI leads to higher SI PSNR. Straightforwardly, the gain is larger on the part of the scene exhibiting swift camera motion.

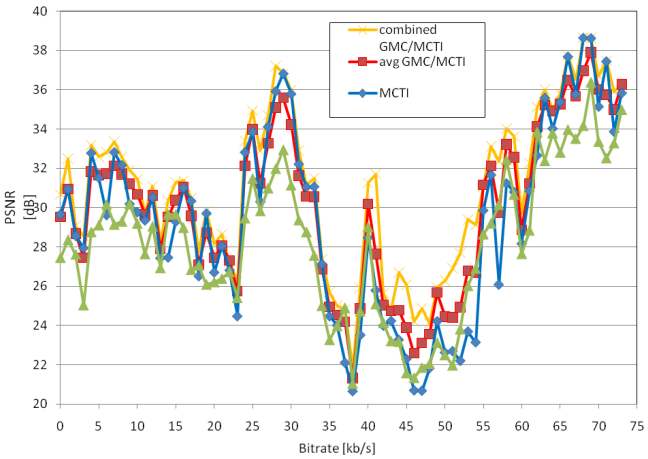

Fig. 3. Side Information PSNR for "Foreman".

Table I shows the corresponding SI average PSNR values for the four test sequences (corresponding to the rate-distortion point with quantization matrix Qi=8 [22]). We observe that the proposed combined GMC-MCTI consistently outperforms conventional MCTI. The gain is more significant for sequences with complex motion such as "Foreman" and "Soccer". In opposition, the gain is negligible for "Hall Monitor", as the scene is mainly static and in this case MCTI is performing very well. It can also be noticed that GMC alone is usually less efficient than MCTI, except for "Soccer". Clearly, Table I shows that the performance improvement is the result of the optimal combination of GMC and MCTI.

TABLE I
SIDE INFORMATION AVERAGE PSNR

Sequence	GMC	MCTI	combined GMC-MCTI
Foreman	28.18	29.31	30.97
Soccer	23.17	22.05	24.43
Coastguard	29.58	31.43	32.22
Hall	35.21	35.77	35.88

Fig. 4 shows the visual quality of the SI obtained by MCTI and the proposed combined GMC-MCTI for a sample frame of "Soccer". In this example, MCTI leads to very obvious distortions, due to the complexity of the motion in the scene which makes truthful interpolation very challenging. In contrast, the visual quality is dramatically enhanced when using the proposed technique. This improvement results from the encoder side motion estimation which is capable of accurately modeling the motion in the scene.

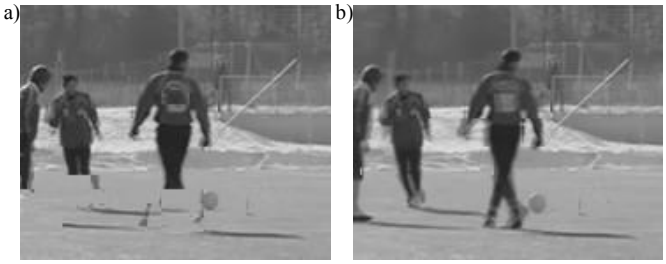

Fig. 4. Side Information visual quality for "Soccer": a) MCTI (PSNR = 21.92 dB), b) combined GMC-MCTI (PSNR = 25.94 dB).

B. Rate Distortion

Finally, we evaluate the rate-distortion performance of the proposed SI generation scheme. More precisely, we compare the two cases: DISCOVER [8] using common MCTI [13], and the proposed DVC architecture with combined GMC-MCTI. We also show the performance of two H.264/AVC schemes, H.264/AVC Intra and H.264/AVC no motion, commonly used as reference in DVC. In the former variant, all frames are Intra coded and no temporal correlation is exploited. In the latter variant, an IB…BI structure is used to exploit temporal redundancy, but without performing motion estimation (i.e. all motion vectors are zero).

The rate-distortion results are shown in Fig. 5 for the three sequences "Foreman", "Soccer" and "Coastguard". The proposed technique consistently outperforms DISCOVER. Again, the improvement is more important for "Foreman" and "Soccer" which exhibit fast motion. For these two sequences, the gain reaches up to 1 dB in the higher bitrate range, but remains around 0.5 dB in the lower bitrate range. For "Foreman", the proposed approach is now always outperforming H.264/AVC Intra, which is not the case with DISCOVER, and is getting closer to H.264/AVC no motion. For "Soccer", despite the performance improvement, the proposed scheme remains notably inferior to both H.264/AVC variants. Finally, both DVC approaches outperform conventional schemes for "Coastguard", as the uniform motion in this scene is well captured in the SI generation process.

VI. Conclusion

Most of the research activities on DVC have so far focused on outperforming conventional coding solutions under the specific constraint of very low encoding complexity. In this paper, we discard this constraint and put forward a complex encoder – complex decoder design.

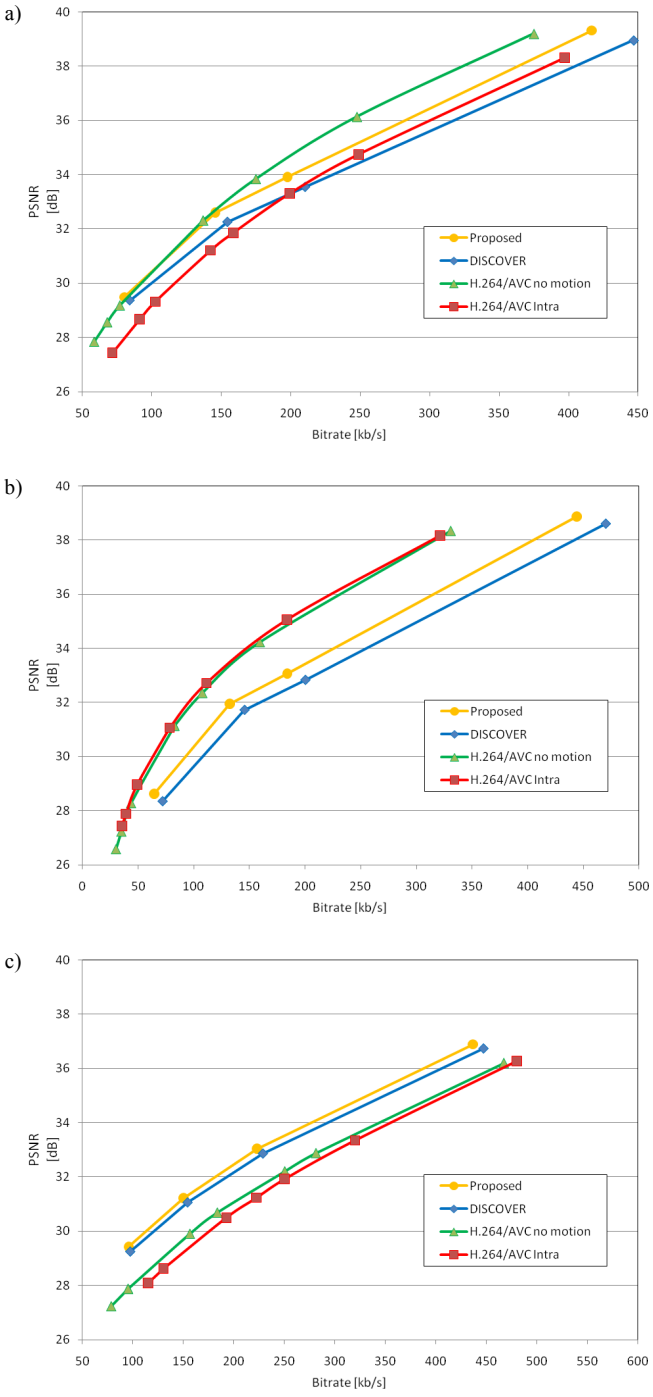

Fig. 5. Rate-distortion performance evaluation comparing the proposed combined GMC-MCTI, DISOCVER with MCTI, as well as H.264/AVC Intra and H.264/AVC no motion: a) "Foreman", b) "Soccer", c) "Coastguard".

More specifically, we propose a new DVC architecture where motion estimation is performed both at the encoder and decoder. Furthermore, we effectively combine global and local motion representations. Experimental results show that the proposed approach improves significantly the quality of SI when compared to common MCTI. In terms of rate-distortion, we report gains of up to 1 dB when compared to the state-of-

the-art DISCOVER DVC codec. Improvements are more important for sequences with complex and fast motion.

As future research activities, we will further explore DVC schemes with complex encoder – complex decoder characteristics, with the goal to outperform conventional MPEG or ITU-T video coding schemes.

ACKNOWLEDGMENT

The authors would like to acknowledge the use of the DISCOVER codec, a software which started from the IST WZ software developed at the Image Group from Instituto Superior Técnico (IST) of Lisbon by Catarina Brites, João Ascenso and Fernando Pereira.

REFERENCES

[1] T. Wiegand, G.J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard", *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 7, July 2003.

[2] J. Slepian and J. Wolf, "Noiseless Coding of Correlated Information Sources", *IEEE Trans. on Information Theory*, vol. 19, no. 4, July 1973.

[3] A. Wyner and J. Ziv, "The Rate-Distortion Function for Source Coding with Side Information at the Decoder", *IEEE Trans. on Information Theory*, vol. 22, no. 1, January 1976.

[4] R. Purit and K. Ramchandran, "PRISM: A new robust video coding architecture based on distributed compression principles", in *Proc. Allerton Conference on Communication, Control and Computing*, Allerton, IL, USA, October 2002.

[5] A. Aaron, R. Thang, and B. Girod, "Wyner-Ziv coding of motion video", in *Proc. Asilomar Conference on Signals and Systems*, Pacific Grove, CA, USA, November 2002.

[6] C. Guillemot, F. Pereira, L. Torres, T. Ebrahimi, R. Leonardi and J. Ostermann, "Distributed Monoview and Multiview Video Coding", *IEEE Signal Processing Magazine*, vol. 24, no. 5, 2007.

[7] F. Dufaux, W. Gao, S. Tubaro, A. Vetro, Distributed Video Coding: Trends and Perspectives, *EURASIP Journal on Image and Video Processing (Special issue on DVC)*, vol. 2009, 2009.

[8] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, M. Ouaret, "The Discover Codec: Architecture, Techniques and Evaluation", in *Proc. of Picture Coding Symposium*, Lisboa, Portugal, November 2007.

[9] F. Pereira, "Video Compression: Still Evolution or Time for Revolution ?", Keynote talk at the 10th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), London, UK, May 2009.

[10] S. Klomp, M. Munderloh, Y. Vatis, J. Ostermann, "Decoder-Side Block Motion Estimation for H.264 / MPEG-4 AVC Based Video Coding", in *Proc. IEEE International Symposium on Circuits and Systems*, Taipei, Taiwan, May 2009.

[11] T. Clerckx, A. Munteanu, J. Cornelis, and P. Schelkens, "Distributed video coding with shared encoder/decoder complexity," in *Proc. IEEE International Conference on Image Processing*, San Antonio, TX, September 2007.

[12] H. Chen and E. Steinbach, "Flexible distribution of computational complexity between the encoder and the decoder in distributed video coding," in *Proc. IEEE International Conference on Multimedia & Expo*, Hannover, Germany, June 2008.

[13] J. Ascenso, C. Brites and F. Pereira, "Improving Frame Interpolation with Spatial Motion Smoothing for Pixel Domain Distributed Video Coding", in *Proc. 5th EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services*, Smolenice, Slovak Republic, July 2005.

[14] K. Misra, S. Karande, H. Radha, "Multi-hypothesis based Distributed Video Coding using LDPC codes", in *Proc. Allerton Conference on Commun, Control and Computing*, Allerton, IL, USA, September 2005.

[15] L. Wei, Y. Zhao, A. Wang, "Improved Side-Information in Distributed Video Coding", in *Proc. International Conference on Innovative Computing, Information and Control*, Beijing, China, August-September 2006.

[16] J. Ascenso, C. Brites, F. Pereira, "Motion Compensated Refinement for Low Complexity Pixel Based Distributed Video Coding", in *Proc. IEEE International Conference on Advanced Video and Signal Based Surveillance*, Como, Italy, September 2005.

[17] X. Artigas, L. Torres, "Iterative Generation of Motion-Compensated Side Information for Distributed Video Coding", in *Proc. IEEE International Conference on Image Processing*, Genova, Italy, September 2005.

[18] W. A. R. J. Weerakkody, W. A. C. Fernando, J. L. Martinez, P. Cuenca, F. Quiles, "An Iterative Refinement Technique for Side Information Generation in DVC", in *Proc. IEEE International Conference on Multimedia and Expo*, Beijing, China, July 2007.

[19] S. Ye, M. Ouaret, F. Dufaux and T. Ebrahimi, Improved Side Information Generation with Iterative Decoding and Frame Interpolation for Distributed Video Coding, in *Proc. IEEE International Conference on Image Processing*, San Diego, CA, October 2008.

[20] F. Dufaux and J. Konrad, "Efficient, Robust, and Fast Global Motion Estimation for Video Coding", *IEEE Trans. on Image Processing*, vol. 9, no.3, March 2000.

[21] T. Ebrahimi, F. Dufaux, and Y. Nakaya, MPEG-4 Natural Video – II, in A. Puri and T. Chen, editors, *Multimedia Systems, Standards, and Networks*, Marcel Dekker, 2000, ch. 9, pp. 245-269.

[22] http://www.img.lx.it.pt/~discover/home.html