RECENT ADVANCES IN MULTI-VIEW DISTRIBUTED VIDEO CODING

Frederic Dufaux, Mourad Ouaret and Touradj Ebrahimi

Institut de Traitement des Signaux
Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland
Frederic.Dufaux@epfl.ch, Mourad.Ouaret@epfl.ch, Touradj.Ebrahimi@epfl.ch

ABSTRACT

We consider dense networks of surveillance cameras capturing overlapped images of the same scene from different viewing directions, such a scenario being referred to as multi-view. Data compression is paramount in such a system due to the large amount of captured data. In this paper, we propose a Multi-view Distributed Video Coding approach. It allows for low complexity / low power consumption at the encoder side, and the exploitation of inter-view correlation without communications among the cameras. We introduce a combination of temporal intra-view side information and homography inter-view side information. Simulation results show both the improvement of the side information, as well as a significant gain in terms of coding efficiency.

Keywords: distributed video coding, multi-view, motion compensated temporal interpolation, inter-view interpolation, homography, network of surveillance cameras

1. INTRODUCTION

Video surveillance systems are becoming omnipresent nowadays, due to high criminality and terrorist threats. Large surveillance systems are deployed in strategic places such as airports, public transportation and downtown. Thanks to the improved performance and reducing cost of cameras, a trend towards dense networks of cameras is expected. The wireless camera sensor network is one example of such as system, which consists of a large number of nodes that are densely positioned, each node being an independent, low power, smart device with sensing, processing and wireless communication capabilities.

Hereafter, we consider a network of multiple cameras which are capturing overlapped images from the same scene with different viewing positions, referred to as multi-view. Multi-view is of interest as it may benefit many vision-based techniques such as object recognition, event detection, target tracking and view interpolation. The range of applications for multi-view systems is very wide and covers different areas such as homeland security and military, but also environment monitoring and healthcare.

Conversely, the amount of data captured in multi-view systems grows tremendously, making data compression a key feature. Due to the strong correlation between images acquired by different cameras, multi-view data compression has its own characteristic that differs significantly from traditional image/video compression. Furthermore, in many applications, it is desirable to have low power consumption in the camera. This puts a strong constraint on the complexity of the encoding process. Furthermore, it prevents a complex inter-node communication system across cameras. It is therefore necessary to develop compression algorithms that are able to exploit the inter-view correlation without requiring any cooperation amongst the cameras.

MPEG is conducting work on Multi-view Video Coding (MVC) [1]. MVC is an extension of the recent Advanced Video Coding (AVC) standard [2]. MVC essentially performs block-based predictive coding across the cameras in addition to predictive coding along the time axis of each camera, hence achieving high compression efficiency. However, the encoder requires high computational power to perform predictive coding. In addition, it calls for communication between the cameras, which is often not feasible in practice.

Distributed video coding (DVC) is a new paradigm in video coding [3][4]. It is based on the Slepian-Wolf [5] and Wyner-Ziv [5] theorems. Basically, the optimal rate achieved when performing joint encoding and decoding of two or more correlated sources can be theoretically reached by doing separate encoding and joint decoding. DVC offers a number of potential advantages: flexible partitioning of the complexity between the encoder and decoder, robustness to channel errors due to intrinsic joint source-channel coding, codec independent scalability, and multi-view coding without communications between the cameras. In a scenario such as a network of surveillance cameras, DVC allows for low power / low complexity cameras as well as no communication between the cameras, which are major advantages.

In Multi-view Distributed Video Coding (MDVC) [7][8][9], side information can be generated either by intra-view Motion Compensated Temporal Interpolation (MCTI) within a camera, or by inter-view interpolation from the side cameras. In [7], view synthesis prediction is used to generate side information from the side cameras. However, no rate-distortion results are reported. In [8], a fusion technique is proposed based on pixel-difference and motion vectors amplitude. Nonetheless, the approach fails to show a coding gain when compared to MCTI. Finally in [9], side information is generated either from MCTI or from Homography-Compensated Inter-view Interpolation (HCII). A fusion technique is proposed to combine both predictions, leading to a coding gain of 0.2 to 0.5 dB when compared to conventional DVC with MCTI.

In this paper, we explore how to improve side information in MDVC. In particular, we introduce more complex modes to perform HCII when compared to [9]. More precisely, the inter-view prediction can be performed either from the left side view, the right side view, or an average of both. We present results in terms of the quality of the side information, as well as rate-distortion performance results.

This paper is structured as follow. We first introduce the concept of DVC in Sec. 2. In Sec. 3, we address multi-view video coding and MDVC. Our proposed MDVC system is presented in Sec. 4. Simulation results are given in Sec. 5 in order to assess its performance. Finally, we draw some conclusions in Sec. 6.

2. DISTRIBUTED VIDEO CODING

Distributed source coding is a new paradigm based on two Information Theory theorems: Slepian-Wolf [5] and Wyner-Ziv [5]. Basically, it states that the optimal rate achieved when performing joint encoding and decoding of two or more correlated sources can theoretically be reached by doing separate encoding and joint decoding. Based on this paradigm, a new video coding model is defined, referred to as DVC [3][4]. In particular, DVC relies on a new statistical framework, instead of the past deterministic approach of conventional coding techniques such as JPEG and MPEG schemes.

DVC offers a number of potential advantages. It first allows for a flexible partitioning of the complexity between the encoder and decoder. Furthermore, due to its intrinsic joint source-channel coding framework, DVC is robust to channel errors. Because it does no longer rely on a prediction loop, DVC provides with codec independent scalability. Finally, DVC is well-suited for multi-view coding by exploiting correlation between views without requiring communications between the cameras.

In this paper, we more specifically consider the DVC codec developed with the European project DISCOVER [10] and illustrated in Figure 1. This codec is partly based on the approach in [11]. Key frames are encoded using the conventional Intra coding of Advanced Video Coding (AVC) [2]. In turn, Wyner-Ziv (WZ) frames are encoded either in the pixel domain, or preferably in the DCT domain. The resulting pixels or transform coefficients undergo quantization. The quantized values are then split into bitplanes which go through a turbo encoder. At the decoder, side information approximating the WZ frames is generated by MCTI of the decoded key frames. The side information is used in the turbo decoder, along with the parity bits of the WZ frames, in order to reconstruct the bitplanes, and subsequently the decoded video sequence.

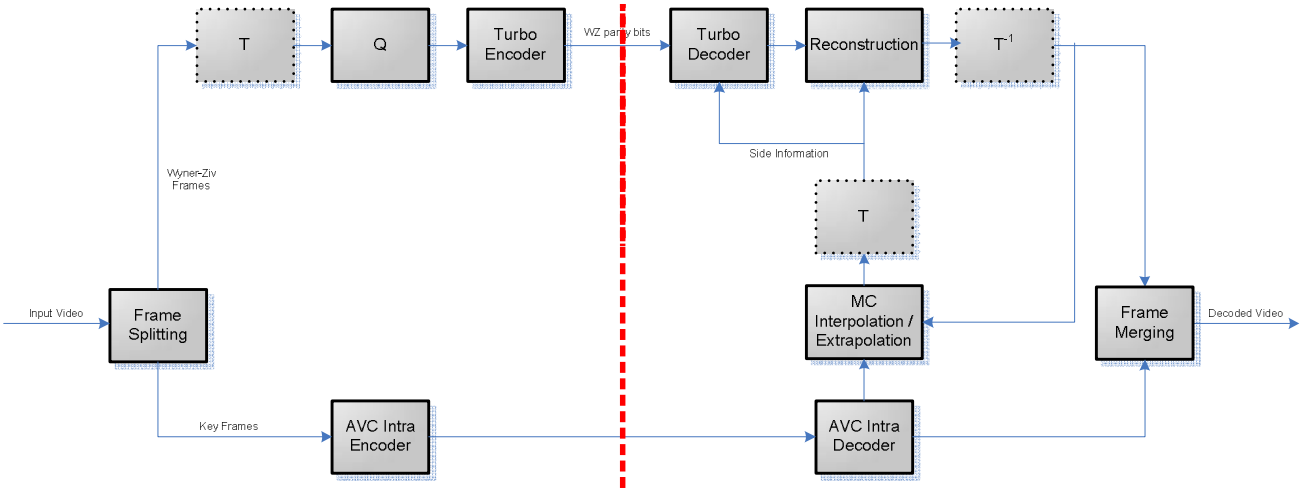

Figure 1 – Distributed Video Coding.

3. MULTI-VIEW VIDEO CODING

In this section, we consider the case of multi-view video coding. By multi-view, we refer to the case of multiple cameras which are capturing overlapped images from the same scene with different viewing positions. This configuration is of interest in video surveillance applications, as many vision-based techniques can benefit from multi-view, such as object recognition, event detection, and target tracking.

Multi-view imaging systems are often generating tremendous amount of data. Hence compression is paramount for their successful deployment. Thankfully, the different views exhibit a strong correlation which can be exploited during compression. However, camera sensor networks are severely constrained in terms of power consumption. Per consequent, it is desirable to have a low complexity encoder and to avoid a complex inter-node communication system.

3.1. Independent Coding

A first configuration is to merely encode the video stream of each camera independently, as illustrated in Figure 2. To achieve state-of-the-art coding performance, AVC can be used for this purpose. However, this configuration does not allow exploiting the correlation between the views. Furthermore, AVC entails a heavy burden on the encoder in terms of complexity.

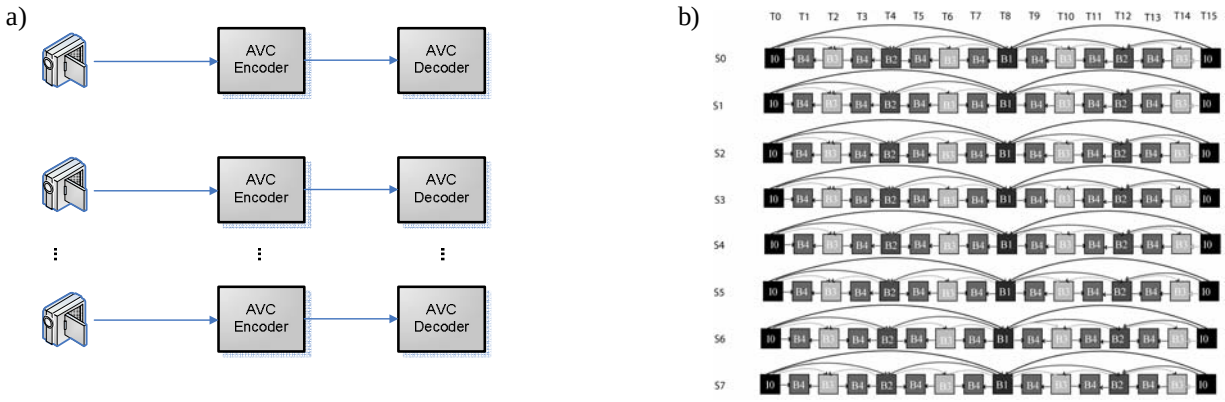

Figure 2 – Independent coding of each view: a) system configuration; b) frame coding structure.

3.2. Multi-view Video Coding (MVC)

MPEG is conducting work on Multi-view Video Coding (MVC), as an extension of AVC. In addition to predictive coding along the time axis of each camera, MVC performs block-based predictive coding across the cameras [1], as depicted in Figure 3. Hence, this coding scheme achieves high coding efficiency. However, this performance is obtained at the cost of a very high computational complexity at the encoder. Furthermore, it calls for communication between the cameras, which is often impractical.

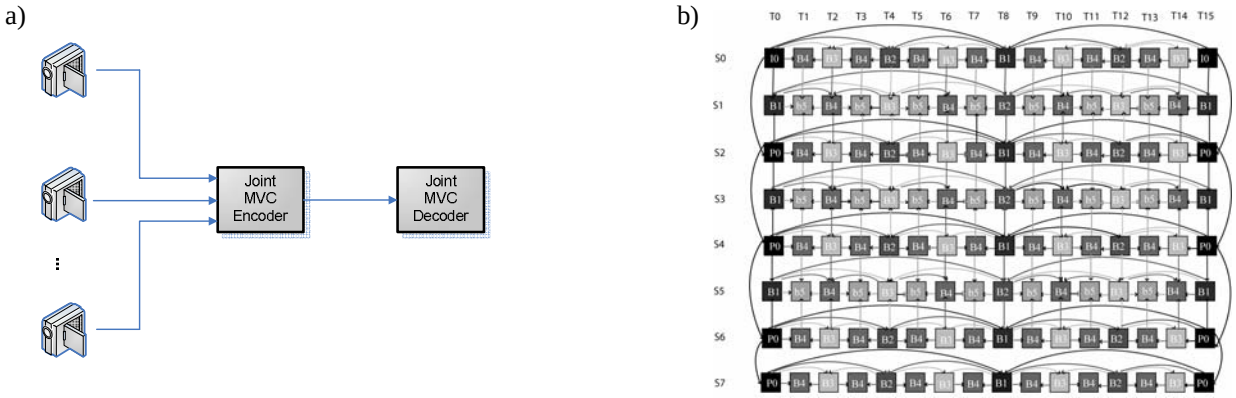

Figure 3 – Multi-view Video Coding (MVC): a) system configuration; b) frame coding structure.

3.3. Multi-view Distributed Video Coding (MDVC)

Given the limitations of the two configurations outlines in Sec. 3.1 and 3.2, MDVC appears as an attractive alternative. Using MDVC, the optimal coding rate can theoretically be reached by doing separate encoding and joint decoding, as shown in Figure 4. In this scenario, side information can be generated either by temporal interpolation within a camera sequence, or by inter-view interpolation from the side views. In other words, the correlation between the views can be exploited at the decoder side, even though the cameras do not communicate. In a practical scenario such as a network of surveillance cameras, this approach allows for low power / low complexity cameras and requires no communication between the cameras, which are major advantages.

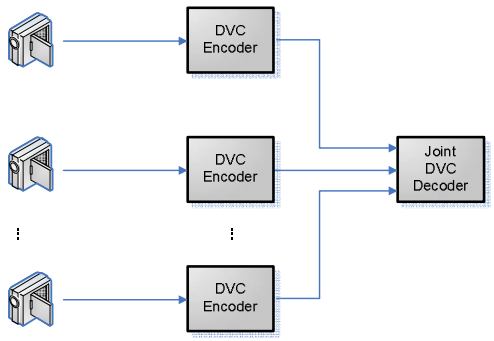

Figure 4 – Multi-view Distributed Video Coding (MDVC).

4. PROPOSED SYSTEM

We now described in more details the proposed MDVC approach. For the sake of simplicity, we consider the particular case illustrated in Figure 5. This set-up is composed of three cameras which are assumed to be static. On the one hand,

the two side views are coded using a conventional AVC Intra encoder [2]. On the other hand, the central view is coded using MDVC, with a codec similar to the one described in Sec. 2 and Figure 1, but which differs in the way to generate the side information. For the central view, we further consider that the odd frames are coded as key frames using AVC and the even frames as WZ.

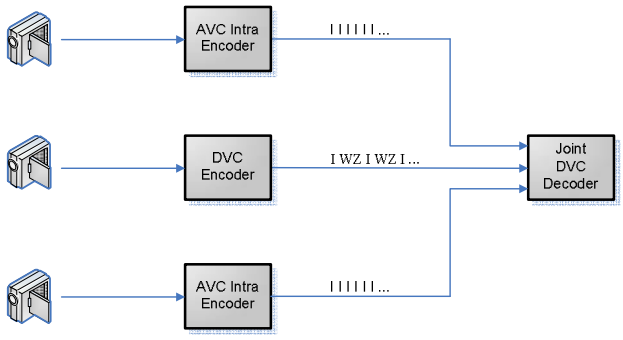

Figure 5 – Configuration under study for Multi-view Distributed Video Coding (MDVC).

4.1. Motion Compensated Temporal Interpolation (MCTI) for intra-view side information

Intra-view side information is generated by MCTI. More precisely, motion estimation is first performed between the previous and next decoded key frames. Commonly, block-based motion vectors are computed by block matching. The side information for the block in the WZ frame is then interpolated by a weighted sum of the motion compensated blocks in the previous and next frames. MCTI is illustrated in Figure 6.

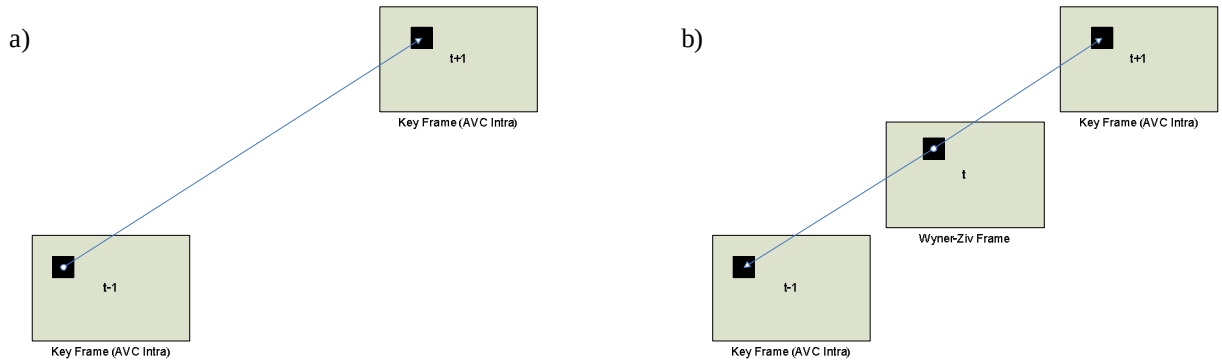

Figure 6 – Intra-view side information by Motion Compensated Temporal Interpolation (MCTI): a) block-based motion estimation, b) motion compensated interpolation.

4.2. Homography Compensated Inter-view Interpolation (HCII) for side information

We now discuss the inter-view side information generation using HCII. More specifically, the disparity between the central and side view is modeled by a homography, also known as a perspective transform

$$x_i' = \frac{a_0 + a_2 x_i + a_3 y_i}{a_6 x_i + a_7 y_i + 1}$$
$$y_i' = \frac{a_1 + a_4 x_i + a_5 y_i}{a_6 x_i + a_7 y_i + 1}$$
,

where (x_i', y_i') denotes the pixel location in the central view, (x_i, y_i) the corresponding position in the side view, and $a_0, a_1, ..., a_7$ the parameters of the transform. This model is valid whenever the scene can be approximated by a planar surface. Two transforms H1 and H2 are computed between the central and left views, respectively the central and right views. The parameters are computed once for the whole sequence, using the first decoded key frame of each view.

The parameters of the homography are estimated by the global motion estimation technique introduced in [12]. More specifically, this is done by minimizing the expression

$$E = \sum_{i=1}^{N} e_i^2 \text{ with } e_i = I'(x_i', y_i') - I(x_i, y_i),$$

where $I'(x_i', y_i')$ and $I(x_i, y_i)$ represent the image pixel values of the central and side views, and 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

$$a^{(t+1)} = a^{(t)} + H^{-1} b \text{ with } H_{kl} = \frac{1}{2} \sum_{i=1}^{N} \frac{\partial^2 e_i^2}{\partial a_k \partial a_l} \cong \sum_{i=1}^{N} \frac{\partial e_i}{\partial a_k} \frac{\partial e_i}{\partial a_l} \text{ and } b_k = -\frac{1}{2} \sum_{i=1}^{N} \frac{\partial e_i^2}{\partial a_k} = -\sum_{i=1}^{N} e_i \frac{\partial e_i}{\partial a_k} .$$

In order to increase robustness to outliers, a truncated quadratic robust estimator can be used

$$\sum_{i=1}^{N} \rho(e_i) \text{ with } \rho(e_i) = \begin{cases} e_i^2 & \text{if } |e_i| \le T \\ 0 & \text{if } |e_i| > T \end{cases},$$

where *T* is a threshold.

HCII can be applied in three distinct modes, as illustrated in Figure 7: by taking the transformed pixel in the left view using the H1 homography (refer to as HCII-left), by taking the transformed pixel in the right view using the H2 homography (refer to as HCII-right), or by taking the average of the two (refer to as HCII-avg).

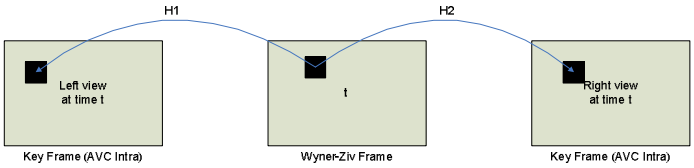

Figure 7 – Inter-view side information by Homography Compensated Inter-view Interpolation (HCII).

4.3. Fusion

In Sec. 4.1 and 4.2, we have presented four different modes to generate the side information:

- MCTI,

- HCII-left,
- HCII-right,
- HCII-avg.

It is possible to switch modes on a pixel by pixel basis. In this paper, we focus on the evaluation of the side information quality. For this purpose, we use an optimal fusion using the original frame to determine the optimal prediction, even though it is unworkable in practice. Note that some heuristics have been proposed in [8][9] to perform fusion.

5. RESULTS

Simulation results are performed using the DISCOVER [10] software codec. The "Breakdancing" multi-view video sequence [13] is used, with a spatial resolution of 256 x 192 pixels and a frame rate of 15 frames per second. The first frame of each view is shown in Figure 8.

Figure 8 – Breakdancing sequence, first frame from left, central, and right views.

We first evaluate the quality of the side information for the various prediction modes. More specifically, Figure 9 shows the PSNR as a function of the frame number, for the modes: HCII-left, HCII-right, HCII-avg, HCII (i.e. optimal fusion of HCII-left, HCII-right, and HCII-avg), MCTI and finally HCII-MCTI (i.e. optimal fusion of MCTI and all the HCII modes). Table 1 shows the corresponding average PSNR values. We observe that HCII-MCTI outperforms the conventional MCTI by more than 6 dB. While each of three HCII modes has a low PSNR, their combination results in a significant improvement of the side information with a gain of almost 3 dB.

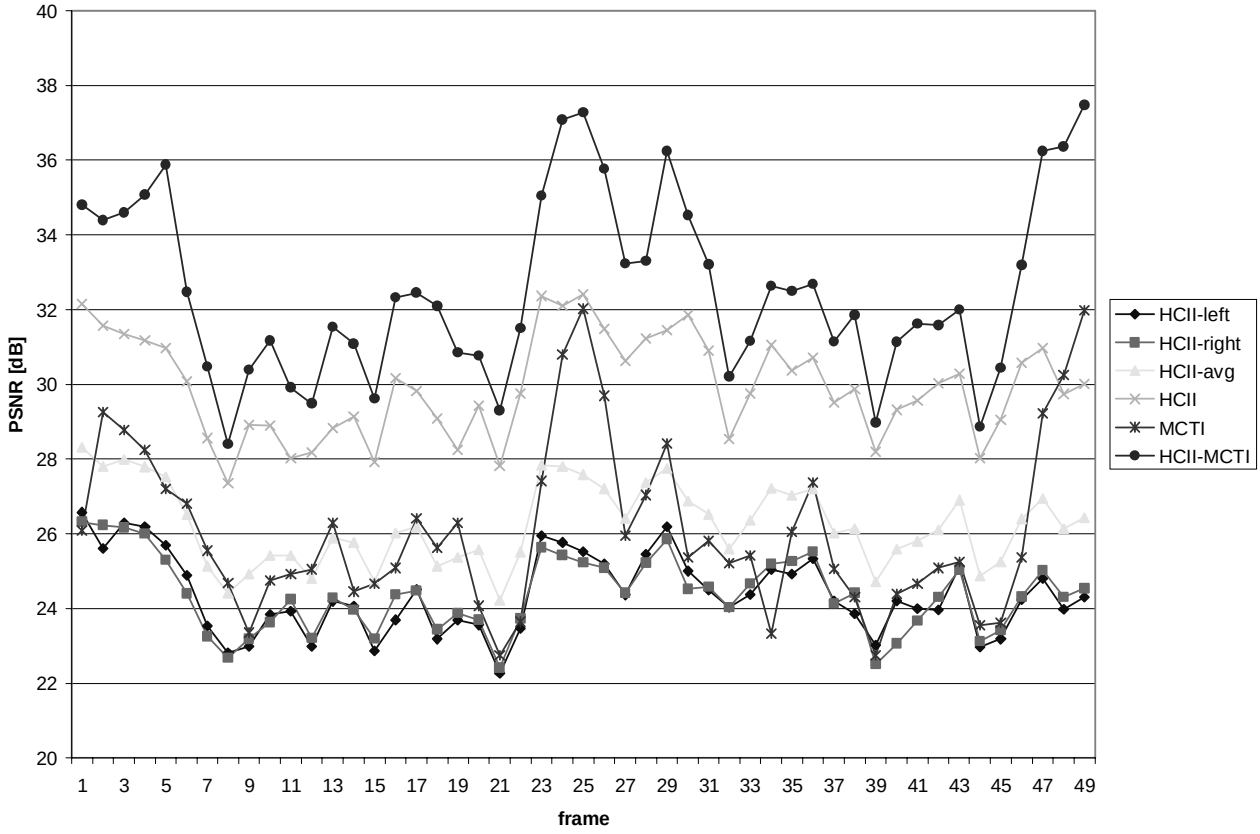

Figure 9 – Side information PSNR for Breakdancing: comparison of HCII-left, HCII-right, HCII-avg, HCII, MCTI, and HCII-MCTI.

	PSNR [dB]
HCII-left	24.37
HCII-right	24.38
HCII-avg	26.25
HCII	29.95
MCTI	26.11
HCII-MCTI	32.54

Table 1 – Average PSNR of side information for Breakdancing: comparison of HCII-left, HCII-right, HCII-avg, HCII, MCTI, and HCII-MCTI

Figure 10 shows the side information obtained with HCII-MCTI for the first WZ frame. When compared to the original frame, it appears that the prediction is good overall, except for the foreground dancer whose motion is not well captured neither by HCII nor by MCTI.

a) b)

Figure 10 – Side information:
a) original of first WZ frame, b) corresponding side information using combined HCII-MCTI.

Figure 11 shows an example of the modes used to generate the side information for the first WZ frame of the Breakdancing sequence. More precisely, four binary masks are shown, corresponding respectively to MCTI (used for 21732 pixels), HCII-left (used for 6574 pixels), HCII-right (used for 13584 pixels) and HCII-avg (used for 7262 pixels). We observe that for this example, the MCTI mode is the most frequently used. Indeed, MCTI is largely used on the static portion of the background, as expected. But the HCII modes are efficient in other regions.

a) b)

c) d)

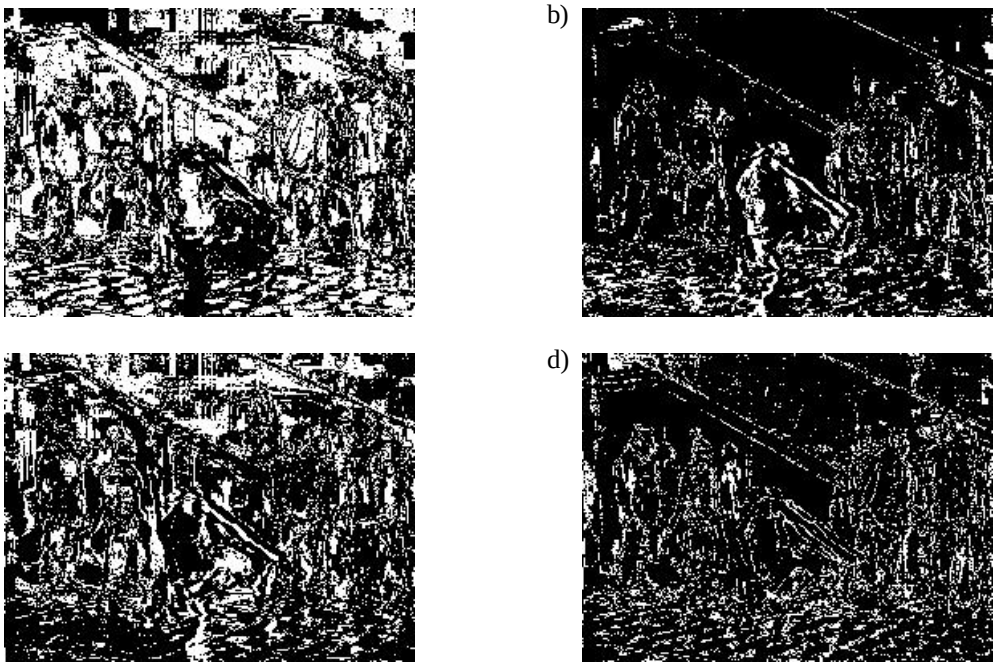

Figure 11 – Fusion mask for first WZ frame of Breakdancing:
a) MCTI, b) HCII-left, c) HCII-right, d) HCII-avg
(a pixel is white in one and only one mask, indicating that it is predicted using the corresponding mode).

Finally, we evaluate the rate-distortion performance in the DISCOVER codec. More precisely, we compare the three cases: MCTI, HCII and combined HCII-MCTI. The combined HCII-MCTI is significantly outperforming the two other modes, with gains ranging from approximately 1 dB to 2 dB. HCII alone is leading to noticeable gains when compared to MCTI, especially at the lower rates. Note that only the luminance component is encoded.

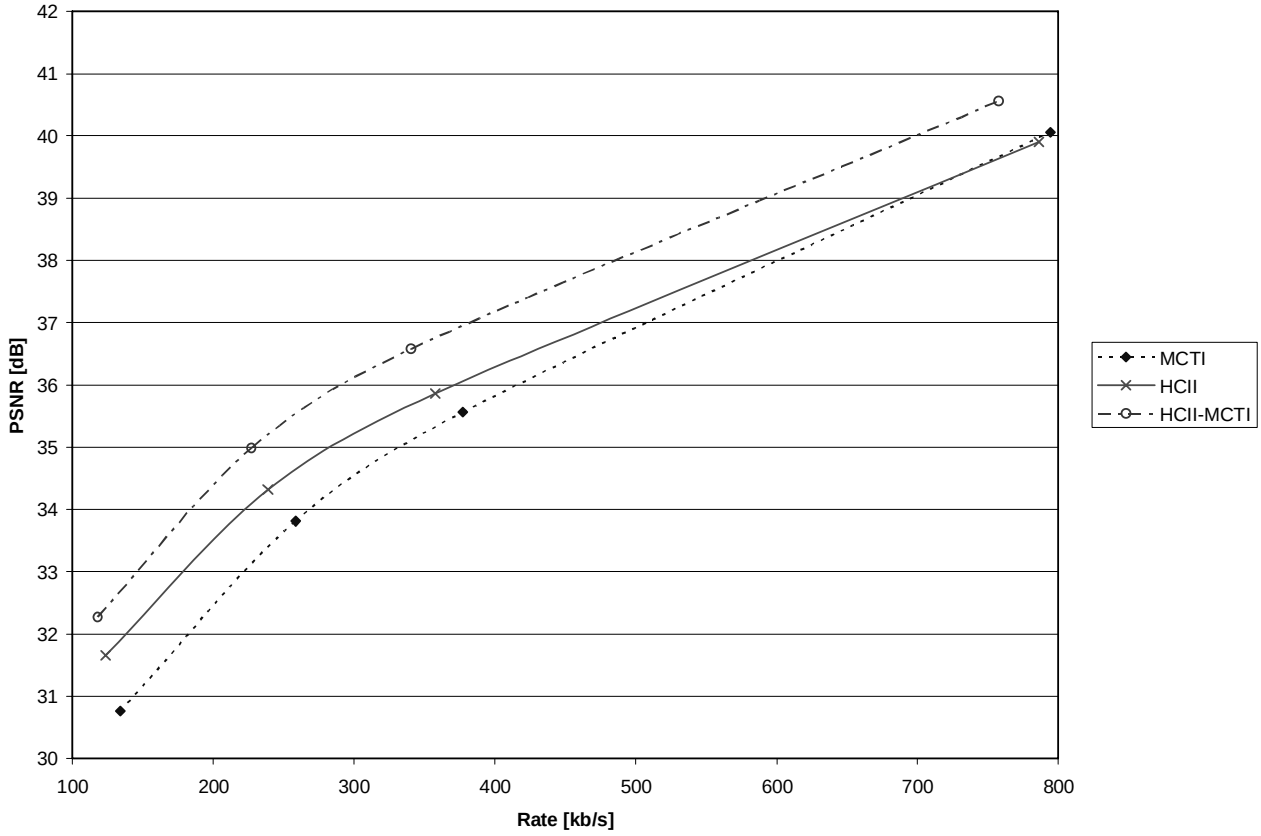

Figure 12 – Rate versus PSNR performance for Breakdancing: comparison of MCTI, HCII, and HCII-MCTI.

6. CONCLUSIONS

Distributed Video Coding (DVC) is a new paradigm in video coding. Besides being suited for multi-view coding, DVC also offers a: flexible partitioning of the complexity between the encoder and decoder, robustness to channel errors due to intrinsic joint source-channel coding, and codec independent scalability. In a network of surveillance cameras, Multi-view Distributed Video Coding (MDVC) allows for low power / low complexity cameras as well as no communication between the cameras.

In this paper, we explore how to improve side information in MDVC. In particular, we introduce more complex modes to perform HCII. Simulation results show that the side information is improved by 6 dB. In terms of rate-distortion, results using the DISCOVER codec show gains ranging from approximately 1 dB to 2 dB.

ACKNOWLEDGEMENT

This work was partially supported by the European Project DISCOVER [10] (IST Contract 015314) and the European Network of Excellence VISNET2 [14] (IST Contract 1-038398), both funded under the European Commission IST 6th Framework Program. 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. The authors would like also to acknowledge the Interactive Visual Media Group at Microsoft Research for the Breakdancing multi-view video sequence.

REFERENCES

[1] P. Merkle, K. Müller, A. Smolic, T. Wiegand, "Efficient Compression of Multi-View Video Exploiting Inter-View Dependencies Based on H.264/MPEG4-AVC", Proc. ICME 2006, International Conference on Multimedia and Expo, Toronto, Ontario, Canada, July 2006.
[2] 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.
[3] B. Girod, A. Aaron, S. Rane and D. Rebollo-Monedero, "Distributed video coding", Proceedings of the IEEE, vol. 93, no. 1, January 2005.
[4] R. Purit and K. Ramchandran, "PRISM: A new robust video coding architecture based on distributed compression principles", Proc. Allerton Conference on Communication, Control and Computing, October 2002.
[5] J. Slepian and J. Wolf, "Noiseless Coding of Correlated Information Sources", IEEE Trans. on Information Theory, vol. 19, no. 4, July 1973.
[6] 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.
[7] X. Artigas, E. Angeli, and L. Torres, "Side Information Generation for Multiview Distributed Video Coding Using a Fusion Approach", 7th Nordic Signal Processing Symposium (NORSIG), Reykjavik, Iceland, June 2006.
[8] X. Guo, Y. Lu, F. Wu, W. Gao, and S. Li, "Distributed Multi-view Video Coding", Visual Communications and Image Processing 2006, San Jose, CA, January 2006.
[9] M.Ouaret, F. Dufaux, and T.Ebrahimi, "Fusion-based Multiview Distributed Video Coding", 4th ACM international workshop on video surveillance and sensor networks 2006, Santa Barbara, CA, October 2006.
[10] http://www.discoverdvc.org
[11] C. Brites, J. Ascenso, F. Pereira, "Improving Transform Domain Wyner-Ziv Video Coding Performance", IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, France, May 2006.
[12] 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.
[13] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," ACM SIGGRAPH and ACM Trans. on Graphics, Los Angeles, CA, Aug. 2004.
[14] http://www.visnet-noe.org