

MULTIVIEW DISTRIBUTED VIDEO CODING WITH ENCODER DRIVEN FUSION

Mourad Ouaret, Frederic Dufaux and Touradj Ebrahimi

Institut de Traitement des Signaux
Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland
{mourad.ouaret, frederic.dufaux, touradj.ebrahimi}@epfl.ch

ABSTRACT

In this paper, a scheme based on multiview and Distributed Video Coding (DVC) is introduced. More specifically, a new fusion technique between temporal, homography and Disparity Compensation View Prediction (DCVP) side informations is introduced. For this purpose, a binary mask is computed at the encoder based on knowledge of the original video. This mask is compressed and then transmitted to the decoder. The latter merges the different side informations with the help of the reconstructed binary mask. The simulation results show that the fusion improves the rate-distortion performance over monoview DVC by a maximum gap of around 1.0 dB.

1. INTRODUCTION

Due to the reducing cost of cameras and improved display technology, multi-camera systems are being used in many fields such as video surveillance and monitoring, 3D reconstruction and Free Viewpoint Television (FTV). However, the amount of captured video in such systems is often huge. This makes video compression important and a key issue in multiview video systems.

Multiview Video Coding (MVC) [1] is the consequence of work conducted by MPEG in 3D Audio-Video (3DAV). It is based on H.264/AVC [2] used for single camera encoding. It performs block-based predictive coding across the different views as well as the time axis of each camera. Predictive coding gives the best compression efficiency. On the other hand, the encoder requires high computational power in addition to communication between the cameras in a practical scenario. However, this is not feasible as it requires complex inter-camera communicating systems, which is time and power consuming and entails complex networking issues.

For the sake of having less complex encoders, work is conducted in the field of Distributed Video Coding (DVC) [4]. Theoretically, it states that the rate achieved when performing joint encoding and decoding of two sources can be reached by doing separate encoding and joint decoding. In a practical scenario, this implies low power / low complexity cameras as well as no communication between the cameras. On the other hand, DVC shift the complexity towards the decoder.

In monoview DVC schemes, side information is often generated temporally using frames from the same camera,

usually the previous and the forward ones. In multiview DVC [5, 6, 7], frames from the side cameras are also involved in generating the side information. The latter is combined, or fused, with the one generated temporally in order to improve the compression efficiency. In [5], View Synthesis Prediction (VSP) is used to generate side information from the side cameras. However, the rate-distortion performance of the approach is not investigated. In addition, VSP requires depth map estimation, which is a hard problem for real world scenes. A fusion technique is used with multiview DVC based on pixel-difference and motion vector thresholding in [6]. Finally in [7], a fusion technique is introduced between temporal and homography-based side informations to improve the overall rate-distortion performance. The homography is estimated using a robust gradient descent algorithm. The fusion consists of two merging algorithms, one is used at low bitrates and the other one at high bitrates. The simulations in [7] report that the fusion outperforms monoview DVC by around 0.2~0.5 dB.

In this paper, a fusion technique is introduced based on some prior knowledge of the original video. For this purpose, a binary mask is computed at the encoder. Then, it is compressed using JBIG [8] and transmitted to the decoder. The latter uses the recovered mask along with temporal, homography and DCVP side informations to construct a better side information.

The paper is structured as follows. First, the monoview and multiview DVC architectures are introduced in section 2 and 3 respectively. Then, the different side information generation techniques are presented in section 4. Then, the simulation results are presented in section 5. Finally, some concluding remarks are drawn in section 6.

2. MONOVIEW DISTRIBUTED VIDEO CODING

DVC is the consequence of information-theoretic bounds established by Slepian and Wolf [9] for distributed lossless coding, and by Wyner and Ziv [10] for lossy coding with decoder side information. In a practical scenario, lossy coding is used. In this paper, the DVC scheme from [11] is used. The latter is illustrated in Fig. 1.

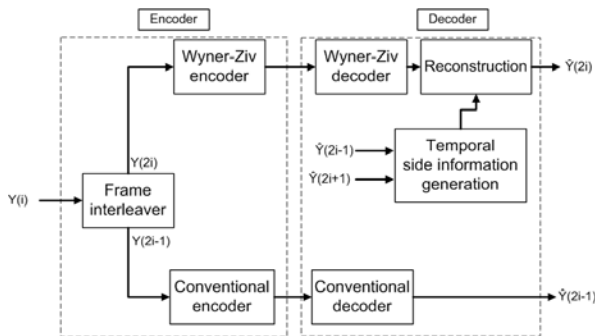


Figure 1. Conventional DVC scheme (GOP = 2).

The Wyner-Ziv encoder operates in the DCT domain. In other words, an interleaved turbo encoder is used to generate parity bits for the quantized DCT coefficients. Moreover, the quantized DCT coefficients are organized in bands, which are separated into bitplanes. The latter are organized from the most to the least significant one. The parity bits are generated for a certain number of bitplanes. As the number of bitplanes increases, the quality of the refined frame improves. For simplicity, the case where the Group Of Pictures (GOP) is equal to two is considered. The conventionally decoded previous and forward key frames are used to generate side information by motion compensated interpolation. To exploit the side information, the decoder assumes a statistical model, which is a Laplacian distribution of the difference between the individual DCT coefficients of the original Wyner-Ziv frame and the side information. The decoder combines the side information and the received parity bits to recover the original frame. For more details on the used DVC scheme, see [4], [10] and [11].

3. MULTIVIEW DISTRIBUTED VIDEO CODING

Multiview DVC differs from the monoview one in the frames involved in the side information generation process. In monoview DVC, frames within the same camera are used for that purpose. In multiview DVC, frames from other cameras can additionally be used to generate side information. In this work, DCVP, homography and temporal side informations are merged at the decoder based on some prior knowledge of the original video. For this purpose, the encoder transmits a binary mask to the decoder in order to define the reference used for each pixel. The mask is compressed using JBIG prior to transmission. The resulting multiview DVC scheme is shown in Figure 2.

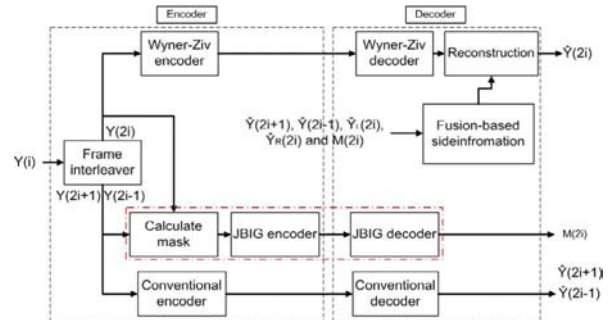


Figure 2. Multiview DVC scheme (GOP = 2).

4. SIDE INFORMATION GENERATION

4.1 Temporal Motion Compensated side information

Temporal side information is used in monoview DVC. It is generated by temporal motion estimation using the previous and the forward key frames. Block-based motion vectors from the previous frame towards the forward frame are computed. Then, the motion vectors are interpolated at mid point to generate the side information. This is done by considering the intersection point of each motion vector with a virtual frame at mid-distance from both key frames as shown in Figure 3. The virtual frame is filled at the block position, which is closest to the intersection point. The latter is filled by a weighted sum of both key frame blocks.

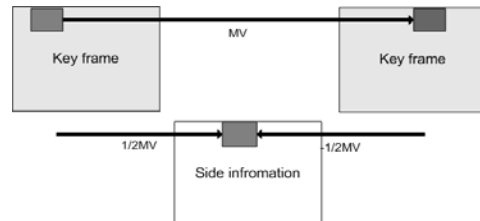


Figure 3. Temporal side information (GOP = 2).

4.2 Homography side information

The homography is a 3×3 matrix that relates one view to another one in the homogenous coordinates system. The matrix has 8 parameters a, b, c, d, e, f, g and h , such that each point from the first view (x_1, y_1) is mapped to a point (x_2, y_2) in the second view up to a scale λ such that :

$$\lambda \begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix}$$

$$x_2 = \frac{ax_1 + by_1 + c}{gx_1 + hy_1 + 1}$$

$$y_2 = \frac{dx_1 + ey_1 + f}{gx_1 + hy_1 + 1}$$

When $a=e=1$ and $b=d=g=h=0$ the model is a pure translation. When $g=h=0$ the model is called an affine transformation. Otherwise, it is called a perspective transformation. These models are suitable when the scene can be approximated by a

planar surface, or when the scene is static and the camera motion is a pure rotation around its optical center [12]. In our case, the first assumption applies.

Depending on the model used, the parameters are computed such that the sum of squared differences between the current frame and the warped frame is minimized. To compute the model parameters, a gradient descent method [12] is used. The latter minimizes a truncated quadratic error function to remove the influence of outliers.

Different side informations are computed using the homographies from the side cameras. More precisely, the side information is computed from the left, right or both cameras as shown in Figure 4.

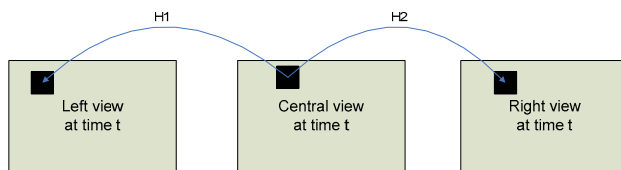


Figure 4. Homographies relating the central view to the side ones.

4.3 Disparity Compensation View Prediction with variable motion vector weighting

DCVP uses the same algorithm as the one used to generate temporal side information. But instead of using the previous and forward frames of the same camera, DCVP uses the left and right frames from the side cameras. Interpolating the motion vectors at mid-point means that middle camera is located exactly at equal distance from the two side cameras. This is not true in all cases.

To calculate the optimal motion vectors weight, the first frame of each camera is conventionally encoded and then decoded. The decoder performs block-based motion estimation between the left and the right camera frames. Then, the motion vectors are weighted with the weights 0.1, 0.2 ... until 0.9. For each weight the Peak Signal to Noise Ratio (PSNR) with respect to the central camera frame is computed. The weight with maximum PSNR is kept and used for the rest of the video.

4.4 Fusion-based side information

In general, the side information generated from the side cameras (either by homography or DCVP) has a poorer quality than the one generated temporally. This is due to the larger disparity between the side camera frames when compared to the one between the previous and forward frames.

The fusion merges the different side informations (temporal, homography and DCVP) in order to improve the quality of the final one.

The idea is to determine a very good estimate of the Wyner-Ziv frame, which is called the reference. The decision for each pixel (i.e. either DCVP, homography or temporal side information) is taken with respect to this reference, as detailed here after.

At the encoder:

Each pixel of the Wyner-Ziv frame is compared to the ones from the previous and the forward frames. If the one from the previous pixel has a closer value, the binary mask at the pixel position is set to one. On the other hand, if the forward pixel has a closer value, it is set to zero. This process is illustrated in Figure 5.

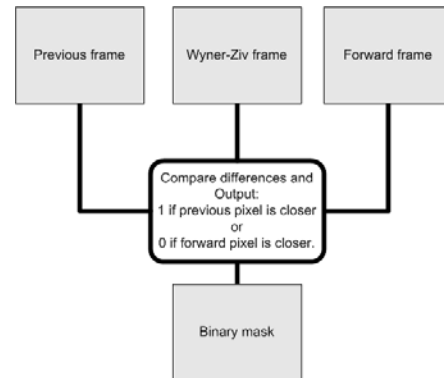


Figure 5. The binary mask generation at the encoder.

The binary mask is encoded using JBIG from the Joint Bi-level Image experts Group [8]. The encoding is preceded by a morphological closing and opening [13]. This eliminates isolated points in the binary mask that would compromise the compression efficiency of the JBIG software.

At the decoder:

The fusion-based side information generation is illustrated in Figure 6. The binary mask is recovered and the different side informations, temporal, homography and DCVP are computed. The binary mask defines for each pixel which reference to use. One in the binary mask means that the pixel values from both side informations should be compared with pixel from the previously decoded frame. On the other hand, a zero in the binary mask means that the comparison is made with respect to the forward one.

5. RATE-DISTORTION SIMULATIONS

5.1 Test material and conditions

The sequences Ballet and Breakdancers [14] are used to compute the rate-distortion curves for both monoview and multiview DVC schemes. The spatio-temporal resolution used is 256x192@15 frames per second. The GOP size is set to two. H.264/AVC Intra is used for encoding the keyframes. The camera setup shown in Figure 7 is used. The side cameras are conventionally encoded. Thus, Only the middle camera contains Wyner-Ziv frames.

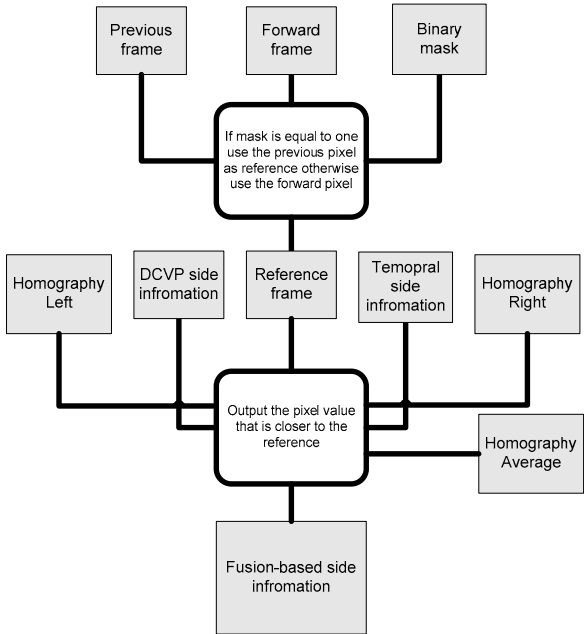


Figure 6. Fusion process at the decoder.

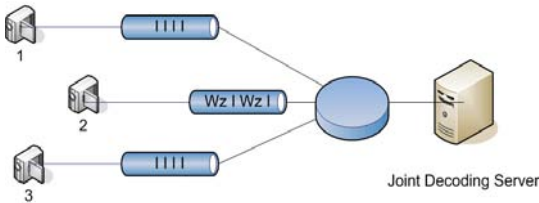


Figure 7. Multiview DVC camera setup. I stands for Intra frame and WZ for Wyner-Ziv frame.

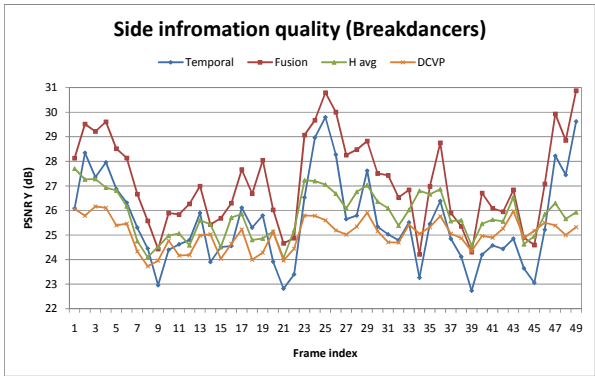


Figure 8. Side information quality for Breakdancers.

5.2 Rate-Distortion performance

In Figures 8 and 9, the PSNR of the different side informations is plotted. It is obvious that the temporal side information quality for the sequence Ballet is by far superior to homography and DCVP by around 7.0 and 8.0 dB respectively. This gap is much smaller for the sequence Breakdancers. The latter contains higher motion than the sequence Ballet. Moreover, the area occupied by the moving object is

greater in the Breakdancers case. This makes temporal interpolation less efficient.

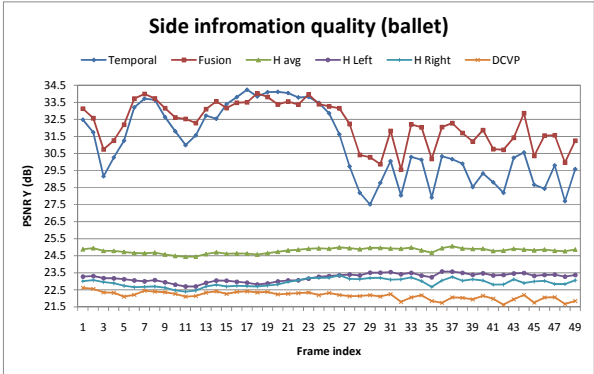


Figure 9. Side information quality for Ballet.

The R-D plots in Figure 10 show that fusion outperforms temporal side information by around 1.0 dB at low bit rates for the sequence Breakdancers. As the bit rate increases, the performance gap decreases. The plots show a 0.5 dB performance gap at average bit rates and a similar performance at very high video quality.

Further, the homography has a similar performance at low bit rates as temporal side information. The gap increases with bit rate to reach a maximum of around 1.0 dB at very high bit rates.

Finally, DCVP has the worst performance with a gap around 1.0-2.0 dB in favour of temporal side information.

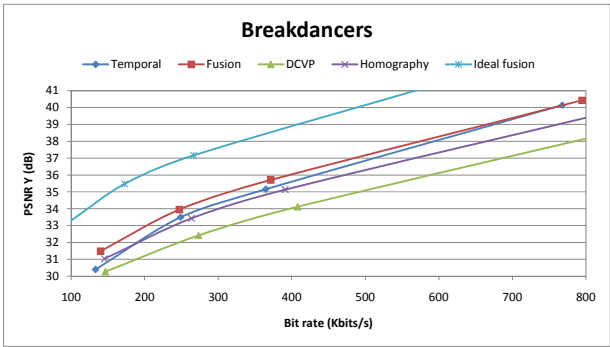


Figure 10. R-D plots for the sequence Breakdancers.

H.264 Intra + Wyner-Ziv rate (Kbits/s)	PSNR Y (dB)	Mask rate (Kbits/s)	Total rate (Kbits/s)	Mask rate (%)
128.44	31.49	11.92	140.36	8.492447991
234.38	33.95	12.9	247.28	5.216758331
358.08	35.71	13.38	371.46	3.602002907
780.05	40.44	14.7	794.75	1.849638251

Table 1. The four R-D points for the Breakdancers sequence.

For the sequence Ballet, The fusion behaves in a similar way with respect to temporal side information as for Breakdancers. It is illustrated in Figure 11. On the other hand, the

homography and DCVP have a poorer performance with respect to temporal side information when compared to Breakdancers.

When comparing these results to the ones in [7] (i.e. when the mask is entirely computed at the decoder), the gain is around 0.5 dB at low bit rates and 0.3 db at average bit rates.

For both sequences, an ideal fusion is computed by using the original Wyner-Ziv frame as a reference. It has a better performance than temporal side information by around 3.0 and 4.0 dB for the sequences Ballet and Breakdancers respectively.

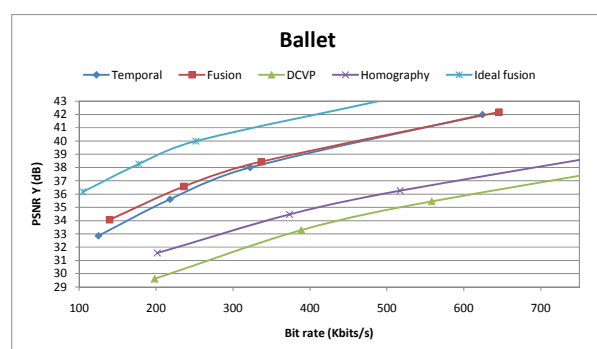


Figure 11. R-D plots for the sequence Ballet.

H.264 Intra + Wyner-Ziv rate (Kbits/s)	PSNR Y (dB)	Mask rate (Kbits/s)	Total rate (Kbits/s)	Mask rate (%)
125.79	34.06	13.8	139.59	9.886094992
221.27	36.57	15	236.27	6.348668896
321.44	38.45	15.4	336.84	4.571903574
629.53	42.19	16.15	645.68	2.501239004

Table 2. The four R-D points for the sequence Ballet.

The tables 1 and 2 show the four R-D points for both sequences. It is noticed that the mask's rate decreases as the video quality increases. Thus, the mask's rate becomes less significant with increased quality as the total rate increases as well.

6. CONCLUSION

In this paper, a novel multiview DVC scheme is introduced. In this scheme, a binary mask is calculated at the encoder based on knowledge of the original video. Then, the mask is compressed and transmitted to the decoder. The final side information is constructed from the temporal, homography and DCVP side informations using the binary mask. It is shown that the fusion improves the side information and the overall rate-distortion performance. The system introduced is interesting for applications requiring low complexity encoders such as distributed sensor networks.

This work can be extended by investigating other fusion techniques. The focus should be on the ones where the fusion mask is entirely calculated at the decoder. It is obvious

that the fusion has to improve the overall rate-distortion performance and close the gap on the ideal fusion as well.

7. ACKNOWLEDGEMENT

This work was developed within DISCOVER, a European Project (www.discoverdvc.org), funded under the European Commission IST FP6 programme.

We would like to thank Microsoft research for providing the video sequences Ballet and Breakdancers.

8. REFERENCES

- [1] <http://www.chiariiglione.org/mpeg>.
- [2] Thomas Wiegand, Gary J. Sullivan, Gisle Bjøntegaard, and Ajay 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] Martinian, E., Behrens, A., Xin, J., Vetro, A., "View Synthesis for Multiview Video Compression," *Picture Coding Symposium (PCS)*, April 2006.
- [4] Bernd Girod, Anne Aaron, Shantanu Rane and David Rebollo-Monedero, "Distributed Video Coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71-83, January 2005.
- [5] Xavi Artigas, Egon Angeli, and Luis Torres, "Side Information Generation for Multiview Distributed Video Coding Using a Fusion Approach," *7th Nordic Signal Processing Symposium (NORSIG)*, Reykjavik, Iceland, June 7-9, 2006.
- [6] Xun Guo, Yan Lu, Feng Wu, Wen Gao, Shipeng Li, "Distributed Multi-view Video Coding," *Visual Communications and Image Processing 2006*, San Jose, California, USA, January 17-19 2006.
- [7] M.Ouaret, F. Dufaux, T.Ebrahimi, "Fusion-based Multiview Distributed Video Coding," *fourth ACM international workshop on Video surveillance and sensor networks 2006*, Santa Barbara, California, October 27 - 27, 2006.
- [8] <http://www.jpeg.org/jbig/index.html>.
- [9] J. Slepian and J. Wolf, "Noiseless Coding of Correlated Information Sources," *IEEE Trans. on Information Theory*, vol. 19, no. 4, July 1973.
- [10] 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.
- [11] C. Brites, J. Ascenso, F. Pereira, "Improving Transform Domain Wyner-Ziv Video Coding Performance," *International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, May 2006.
- [12] Frederic Dufaux and Janusz Konrad, Efficient, "Robust, and Fast Global Motion Estimation for Video Coding," *IEEE transactions on image processing*, vol. 9, no.3, March 2000.
- [13] R. M. Haralick, S. R. Stemberg, and X. Zhuang, "Image analysis using mathematical morphology," *IEEE Trans. Pattern Anal Machine Intell.*, vol. 9, pp. 523-550, July 1987.
- [14] <http://research.microsoft.com/TVM/3DVideoDownload/>.