FUSION SCHEMES FOR MULTIVIEW DISTRIBUTED VIDEO CODING

Thomas Maugey, Wided Miled, Marco Cagnazzo and Béatrice Pesquet-Popescu

TELECOM ParisTech, CNRS LTCI, Signal and Image Processing Department
46 rue Barrault, 75634 Paris Cedex 13, France
Email: {maugey, miled, cagnazzo, pesquet}@telecom-paristech.fr

ABSTRACT

Distributed video coding performances strongly depend on the side information quality, built at the decoder. In multi-view schemes, correlations in both time and view directions are exploited, obtaining in general two estimations that need to be merged. This step, called fusion, greatly affects the performance of the coding scheme; however, the existing methods do not achieve acceptable performances in all cases, especially when one of the estimations is not of good quality, since in this case they are not able to discard it. This paper provides a detailed review of existing fusion methods between temporal and inter-view side information, and proposes new promising techniques. Experimental results show that these methods have good performances in a variety of configurations.

1. INTRODUCTION

Distributed Video Coding (DVC) is a quite recent paradigm with high potential in many practical applications, in particular multiple camera video transmission. In the 1970's, Slepian-Wolf [1] and Wyner-Ziv [2] obtained interesting theoretical results in information theory, showing that a system can attain the same rate-distortion performances while encoding two correlated sources *jointly* or *independently* provided that the decoding is performed jointly. In multi-view video coding, this means that the correlation between cameras can be exploited just at the decoder, without affecting the performances. In other words, encoding complexity can be reduced and communication between cameras avoided, while compression efficiency is preserved.

Even though theory was known for a long time, only quite recently the first practical solutions for DVC have appeared. One popular solution takes as starting point the Stanford coding scheme [3], which consists in first splitting the video sequence into two subsets: the *key frames* (KF) and the *Wyner-Ziv frames* (WZF). This step is critical especially for multi-view DVC, since the frame repartition strongly affects the rest of the scheme, as well as the employed prediction and fusion techniques. The main existing solutions are summarized in [4]. In this work, we adopt the so called symmetric scheme 1/2, which gives identical roles to all cameras: each of them produces alternatively one KF followed by one WZF. A shift is introduced between cameras, in order to obtain a quincunx frame repartition in the time-view domain (see Fig. 1). The KFs are intra encoded/decoded using an H.264 Intra codec. The WZFs are transformed (using DCT) and quantized, and the resulting bit-planes are turbo-encoded. However, the systematic bits are not transmitted, and instead they are replaced, at the decoder side, by an estimation of the WZF, called *side information* (SI), generated from the decoded KFs. This SI is corrected by the turbo-decoder with the parity bits sent by the encoder, and finally the corrected DCT coefficients undergo an inverse transform to produce the decoded WZF.

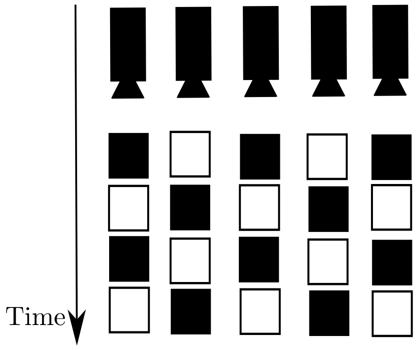

Figure 1: Time-space frame repartition. KFs are in black and WZFs in white.

As shown in many works [5], the global coding performances strongly depend on the quality of the SI. The better the side information, the fewer bits are required to encode the WZF. The proposed methods deal with the generation of the SI through interpolation. In the symmetric 1/2 frame disposition, for a given WZF two interpolations can be used, a temporal estimation (using the backward and the forward KFs) and an inter-view estimation (using KFs from the left and right views). The two estimations need to be combined, or fused, in order to build a unique SI for the turbo decoder, while improving the rate distortion performance. Many fusion techniques have been proposed in the literature. However, the presented performances do not show, for all cases, the gain with respect to just using temporal or inter-view side information.

In this paper, we first state the fusion problem in Sec. 2. In order to explore temporal and spatial redundancies, temporal and inter-view interpolations use respectively motion and disparity vectors estimation and compensation. We consider in this work a convex variational approach allowing to obtain dense and accurate displacement vectors. This approach is briefly described in Sec. 2.1, but reader is referred to [6] for more details. Then, we review in Sec. 3 some of the state-of-the-art fusion methods proposed in the multi-view DVC context and, in Sec. 4, we present three novel efficient fusion techniques. Experimental results are provided in Sec. 5. The paper ends with a conclusion in Sec. 6.

2. SIDE INFORMATION CONSTRUCTION

2.1 Motion/disparity estimation

In this paper, we consider only rectified multi-view sequences [7]. This is a very common assumption according to which, at a given time t, the disparity between the frames of the different cameras is only horizontal. We denote the t^{th} frame of the n^{th} camera by $I_{n,t}$. The goal of the interpolation is to generate an estimation of a frame from two available reference frames I_{n_1,t_1} and I_{n_2,t_2}. If $t_1 = t_2$, the interpolation uses only information from the left and right frames and is based on disparity estimation. Whereas, if $n_1 = n_2$, interpolation performs motion estimation between the forward and the backward frames. To estimate both disparity and motion vector fields, we use in this work a dense variational estimation method [6], consisting in minimizing an appropriate convex objective function under various convex constraints. A total variation based regularization constraint is considered in order to output a smooth disparity or motion field while preserving discontinuities. The resulting convex optimization problem is solved using a parallel block iterative algorithm based on recently developed convex analysis tools [6]. We use this very efficient algorithm to perform both temporal and interview estimations, leading finally to dense and accurate displacement fields with ideally infinite precision, which are used to compensate the reference frames.

2.2 Fusion problem statement

This section states the fusion problem and defines the notations for the SI generation of a WZF $W_{n,t}$. The fusion problem springs up since in the multi-view DVC context, one ends up with having two different estimations of the current WZF, $W_{n,t}$ coming from the temporal and the inter-view interpolations. This is illustrated in Fig. 2: motion estimation produces two motion vector fields, $\mathbf{v}_b$ and $\mathbf{v}_f$, which in turn are used to provide temporal estimations of $W_{n,t}$ from $I_{n,t-1}$ and $I_{n,t+1}$. Therefore, we note with $\tilde{I}_{n,t^-} = I_{n,t-1}(\mathbf{v}_b)$ the prediction obtained by compensating the image $I_{n,t-1}$ with vector $\mathbf{v}_b$. Likewise, we have $\tilde{I}_{n,t^+} = I_{n,t+1}(\mathbf{v}_f)$. As far as disparity estimation is concerned, we note the disparity fields as $\mathbf{v}_l$ and $\mathbf{v}_r$ (which have quite different characteristics from motion vector fields), and the corresponding estimations as $\tilde{I}_{n^-,t}$ and $\tilde{I}_{n^+,t}$. Finally, the two temporal (or interview) estimations are combined in order to obtain a single estimation, respectively $\tilde{I}_T = \frac{1}{2}\left(\tilde{I}_{n,t^-} + \tilde{I}_{n,t^+}\right)$ and $\tilde{I}_N = \frac{1}{2}\left(\tilde{I}_{n^-,t} + \tilde{I}_{n^+,t}\right)$. The fusion problem amounts to produce an estimation of $W_{n,t}$ from $\tilde{I}_T$ and $\tilde{I}_N$ with the target of minimizing the mean square error with respect to the actual WZF. In particular, an efficient fusion technique should produce a smaller MSE than both the “non-fusion” estimations $\tilde{I}_T$ and $\tilde{I}_N$.

3. EXISTING FUSION SOLUTIONS

In this section, we review the existing solutions for the fusion problem in the case of a quincunx frame repartition. Some existing solutions are not studied here since they are based on other configurations. For exemple, some techniques [8,9] use an hybrid frame repartition in the time-view space and are not available in the symmetric scheme adopted in this paper. In another solution, reported in [10, 11] the fusion is performed on the basis of a frame analysis to be carried out at the encoder side. However this method would not be coherent with our intended DVC framework, where no joint processing of images is allowed at the encoder.

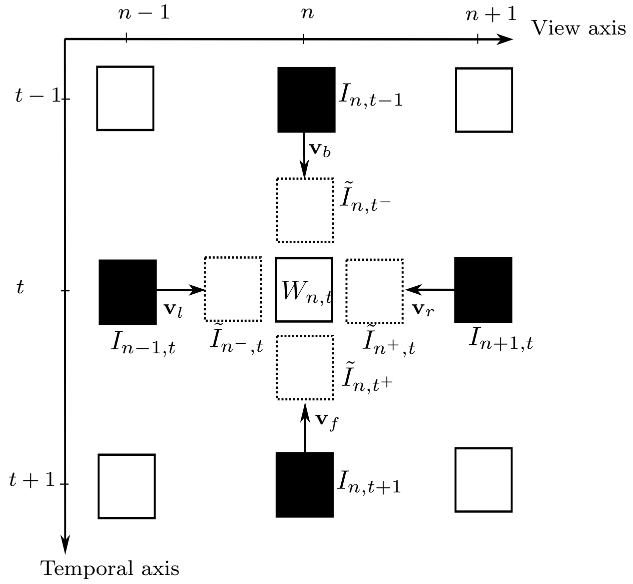

Figure 2: Fusion problem: I_x are the available KFs and $\tilde{I}_x$ their motion compensated version, estimating the WZF $W_{n,t}$. $\mathbf{v}_x$ are the vector fields.

The **ideal fusion** (Id), studied in [4,12] is the upper bound one can achieve when performing a fusion. Pixel by pixel, the true estimation error, taking into account the original WZF, is computed and used as an oracle in order to decide what is the best value for the SI. The equation of the ideal fusion is for each pixel $\mathbf{s}$:

$$\tilde{I}(\mathbf{s}) = \begin{cases} \tilde{I}_N(\mathbf{s}), & \text{if } |\tilde{I}_N(\mathbf{s}) - W_{n,t}(\mathbf{s})| < |\tilde{I}_T(\mathbf{s}) - W_{n,t}(\mathbf{s})| \\ \tilde{I}_T(\mathbf{s}), & \text{otherwise.} \end{cases}$$

The **pixel difference fusion** (PD) was proposed by Ouaret et al. in [8]. The interpolation error is estimated using the backward and forward frames of the same view. Two estimation errors are computed for the inter-view interpolation $E_N^b = |\tilde{I}_N - I_{n,t-1}|$ and $E_N^f = |\tilde{I}_N - I_{n,t+1}|$ and, similarly, for temporal interpolation $E_T^b = |\tilde{I}_T - I_{n,t-1}|$ and $E_T^f = |\tilde{I}_T - I_{n,t+1}|$. The equation of the PD fusion is therefore:

$$\tilde{I}(\mathbf{s}) = \begin{cases} \tilde{I}_N(\mathbf{s}), & \text{if } E_N^b(\mathbf{s}) < E_T^b(\mathbf{s}) \text{ and } E_N^f(\mathbf{s}) < E_T^f(\mathbf{s}) \\ \tilde{I}_T(\mathbf{s}), & \text{otherwise.} \end{cases}$$

The **motion compensated difference fusion** (MCD) was proposed by Guo et al. in [13]. In this fusion algorithm, the absolute value of the difference between $\tilde{I}_{n,t^-}$ and $\tilde{I}_{n,t^+}$ is thresholded by T_1 and the motion vector values are also thresholded by T_2. The equation of the MCD fusion process is:

$$\tilde{I}(\mathbf{s}) = \begin{cases} \tilde{I}_N(\mathbf{s}), & \text{if } |\tilde{I}_{n,t^-}(\mathbf{s}) - \tilde{I}_{n,t^+}(\mathbf{s})| > T_1 \\ & \quad \text{or } \|\mathbf{v}_b(\mathbf{s})\| > T_2 \\ & \quad \text{or } \|\mathbf{v}_f(\mathbf{s})\| > T_2 \\ \tilde{I}_T(\mathbf{s}), & \text{otherwise.} \end{cases}$$

The **view projection fusion** (Vproj) was proposed by Ferré et al. in [14]. In this case, the estimation $\tilde{I}_T$ is projected onto $I_{n-1,t}$ and $I_{n+1,t}$. This projection consists in disparity compensations ($dc_l(\cdot)$ and $dc_r(\cdot)$) based on a simple block matching disparity estimation. The error images $E_l = I_{n-1,t} - dc_l(\tilde{I}_T)$ and $E_r = I_{n+1,t} - dc_r(\tilde{I}_T)$ are thresholded, leading to two masks which are projected back onto the WZF, with disparity compensations ($dc_l^{-1}(\cdot)$ and $dc_r^{-1}(\cdot)$) based on $\mathbf{v}_r$ and $\mathbf{v}_l$. The equation of the Vproj fusion process is:

$$\tilde{I}(\mathbf{s}) = \begin{cases} \tilde{I}_N(\mathbf{s}), & \text{if } |dc_l^{-1}(E_l)(\mathbf{s})| > T \text{ or } |dc_r^{-1}(E_r)(\mathbf{s})| > T \\ \tilde{I}_T(\mathbf{s}), & \text{otherwise.} \end{cases}$$

The **temporal projection fusion** (Tproj) was proposed by Ferré et al. in [14]. It is the equivalent of the Vproj fusion in the temporal direction. The estimation $\tilde{I}_N$ is first projected on $I_{n,t-1}$ and $I_{n,t+1}$ by motion compensation. Two error images, $E_b = I_{n,t-1} - mc_l(\tilde{I}_N)$ and $E_f = I_{n,t+1} - mc_r(\tilde{I}_N)$, are then thresholded and the obtained masks are projected back onto the original position. The equation of the Tproj fusion process is:

$$\tilde{I}(\mathbf{s}) = \begin{cases} \tilde{I}_N(\mathbf{s}), & \text{if } mc_b^{-1}(E_b) < T \text{ or } mc_f^{-1}(E_f) < T \\ \tilde{I}_T(\mathbf{s}), & \text{otherwise.} \end{cases}$$

4. PROPOSED FUSION METHODS

The fusion solutions presented in the previous section achieve good performances in some cases. For example, the PD fusion is quite efficient when the temporal motion activity is low. On the contrary, non-fusion estimation qualities strongly depend on the sequence. In this section, we propose three new methods aiming at more robustness. The first two use the residual (*i.e.* the difference between the two compensated reference frames), like the MCD fusion does. The residual is commonly used to approximate the estimation error in DVC, for example for the distribution model analysis at the turbo decoder.

The **motion and disparity compensated difference binary fusion** (MDCDBin) compares the temporal and inter-view residuals, and uses for the estimation the one having the smallest one. Similarly to the existing solutions, the decision is binary. The residuals are defined as $E_T(\mathbf{s}) = |\tilde{I}_{n,t^-}(\mathbf{s}) - \tilde{I}_{n,t^+}(\mathbf{s})|$ and $E_N(\mathbf{s}) = |\tilde{I}_{n^-,t}(\mathbf{s}) - \tilde{I}_{n^+,t}(\mathbf{s})|$. Therefore, the equation of MDCDBin is:

$$\tilde{I}(\mathbf{s}) = \begin{cases} \tilde{I}_N(\mathbf{s}), & \text{if } E_N(\mathbf{s}) < E_T(\mathbf{s}) \\ \tilde{I}_T(\mathbf{s}), & \text{otherwise.} \end{cases}$$

The two following proposed methods adopt a quite different approach. Instead of a binary decision, the fusion process is now based a linear combination between the available values. Note that a linear fusion can in principle outperform the ideal binary fusion (Id) and no upper bound can be found.

In the case of **motion and disparity compensated difference linear fusion** (MDCDLin), the criterion of MDCDBin is improved by computing a linear combination of inter-view and temporal estimation, as follows:

$$\tilde{I}(\mathbf{s}) = \frac{E_T(\mathbf{s})}{E_T(\mathbf{s}) + E_N(\mathbf{s})}\tilde{I}_N(\mathbf{s}) + \frac{E_N(\mathbf{s})}{E_T(\mathbf{s}) + E_N(\mathbf{s})}\tilde{I}_T(\mathbf{s})$$

Finally, in the case of **Estimation-error and vector-norm based linear fusion** (ErrNorm), we build on the consideration that often, the larger are the motion vectors, the less reliable is the estimation. Therefore, we use the motion vector norms as weights in computing a linear combination between $\tilde{I}_T$ and $\tilde{I}_N$. The resulting image is then averaged with the one produced by MDCDLin to obtain the new estimation. More precisely, in the ErrNorm case we have the following equations:

$$\tilde{I}(\mathbf{s}) = \frac{\tilde{I}_{\text{err}}(\mathbf{s}) + \tilde{I}_{\text{norm}}(\mathbf{s})}{2} \qquad \text{where}$$

$$\tilde{I}_{\text{norm}}(\mathbf{s}) = \frac{(\|\mathbf{v}_b\| + \|\mathbf{v}_f\|)\tilde{I}_N(\mathbf{s}) + (\|\mathbf{v}_l\| + \|\mathbf{v}_r\|)\tilde{I}_T(\mathbf{s})}{\|\mathbf{v}_b\| + \|\mathbf{v}_f\| + \|\mathbf{v}_l\| + \|\mathbf{v}_r\|}$$

$$\tilde{I}_{\text{err}}(\mathbf{s}) = \frac{E_T(\mathbf{s})\tilde{I}_N(\mathbf{s})}{E_T(\mathbf{s}) + E_N(\mathbf{s})} + \frac{E_N(\mathbf{s})\tilde{I}_T(\mathbf{s})}{E_T(\mathbf{s}) + E_N(\mathbf{s})}$$

5. EXPERIMENTAL RESULTS

We compared the state-of-the-art fusion techniques presented above with the proposed ones, by running them on two multi-view test sequences, “Book Arrival” and “Outdoor”, from [15]. For both sequences, the spatial resolution was halved from 1024×772 to 512×386, and only the first 8 cameras were used. We performed the displacement estimation algorithm presented in Sec. 2.1 in order to produce the vector fields for both temporal and inter-view interpolations. We considered lossy coded KFs and four quantization steps (QP= 31, 34, 36 and 40), in order to observe the behavior of fusion methods in a relatively wide range of bit-rates.

The performance of all the methods are shown in Fig. 3 where we give the PSNR of the SI with respect to the original WZF. Note that the mean square error is commonly used to measure the SI quality, and the conclusions one can draw from it will be further confirmed with the final rate distortion performances. Gray bars correspond to simple cases, where only temporal or inter-view estimation are considered, the white bar correspond to the ideal (i.e. oracle-driven) fusion, the blue bars are the state-of-the-art methods explained in Section 3, and the red ones are the proposed techniques. We notice that for “Book Arrival” test sequence, the temporal estimation is slightly better than the inter-view one, while the opposite is true for the second sequence, “Outdoor”. In both cases, the comparison between the

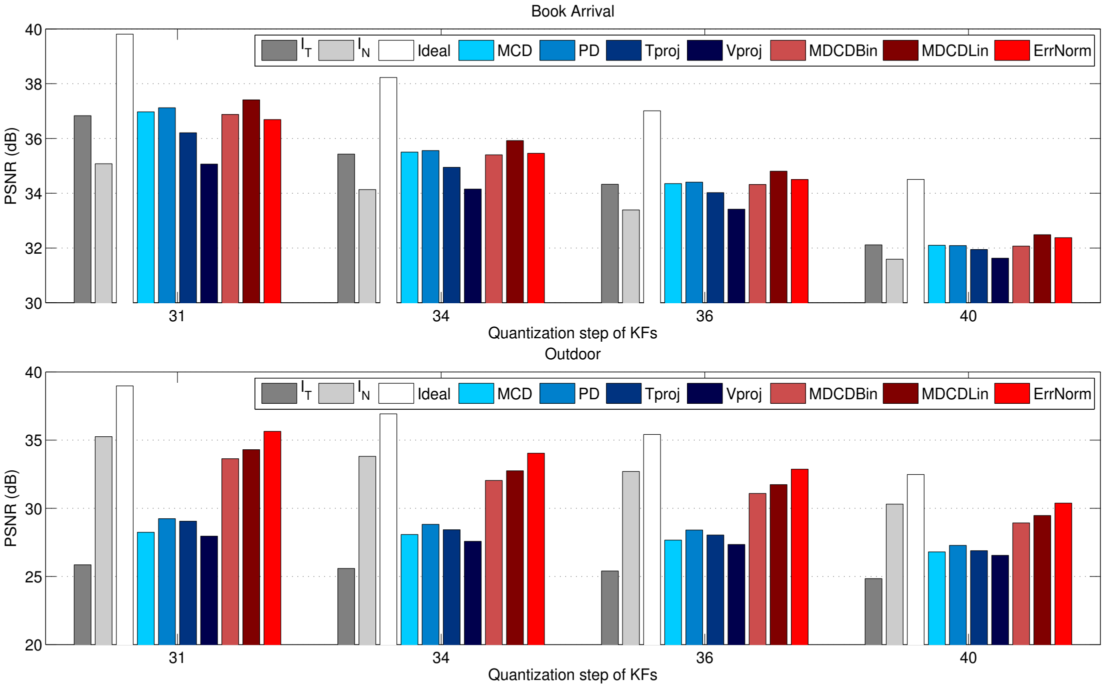

Figure 3: SI quality for different fusion methods, at different KF quantization levels, and for two test sequences "Book Arrival" and "Outdoor".

ideal fusion (which can be seen as an upper bound for fusion method performances) and no-fusion cases, shows that fusion has the potential of largely improve the WZF estimation.

However state-of-the-art methods look like not able to adequately take advantage from fusion: while for the "Book Arrival" sequence, MCD and PD fusion obtain good performances, much better than the non-fusion predictions $\tilde{I}_T$ and $\tilde{I}_N$, this is no longer the case for the second sequence, where state-of-the-art methods perform worse than simple inter-view estimation. We conclude that these methods are not robust enough when there is a sensible gap of quality between the temporal and inter-view estimations.

Different observations can be made for the proposed methods (red bars in Fig. 3). The first remark is that MDCDLin outperforms MDCDBin, showing that a linear based fusion is more efficient than a binary decision based method. Moreover, for "Book Arrival" sequence, the MDCDBin method reaches better performances than the existing solutions. For "Outdoor" sequence, where the other solutions obtain a lower SI quality, the proposed methods achieve good results and ErrNorm fusion sensibly improves the $\tilde{I}_N$ prediction. Finally, for ease of comparison, some of the results of Fig. 3 are reported in Tab. 1 and 2, in terms of the difference between the best non-fusion estimation for each sequence and three fusion methods, PD (the best existing method), MDCDLin and ErrNorm (the best proposed methods).

In Fig. 4 we present the rate-distortion performance obtained when using PD, MDCDLin and ErrNorm within a complete DVC multiview coder as DISCOVER [12]. The results confirm that the proposed methods (red curves) outperform existing ones (blue curves). In order to facilitate the comparison, the average performances computed with the Bjontegaard metric [16] are shown in Tab. 3 and 4. We note that ErrNorm is consistently better than than the non-fusion techniques (obtaining a rate reduction up to 3.64%), while MDCDLin is always better than PD, which in turns, is much worse than the non-fusion method for the "Outdoor" sequence.

QP	31	34	36	40
PD	-6.0131	-4.9926	-4.2939	-3.0226
MCDLin	-0.9516	-1.0624	-0.9639	-0.8322
ErrNorm	0.3893	0.2253	0.1658	0.0740

Table 1: Δ_{PSNR} between different fusion method and the best non-fusion estimation (inter-view estimation in this case) for "Outdoor" sequence.

QP	31	34	36	40
PD	0.2901	0.1293	0.0807	-0.0244
MCDLin	0.5777	0.4926	0.4799	0.3709
ErrNorm	-0.1393	0.0271	0.1761	0.2636

Table 2: Δ_{PSNR} between different fusion method and the best non-fusion estimation (temporal estimation in this case) for "Book Arrival" sequence.

6. CONCLUSION

In this paper, a review of some existing fusion methods in the multi-view DVC framework have been first presented. Then, three new fusion techniques showing better robustness and improving SI quality have been introduced. Experimental results show that the proposed solutions achieve good results for different inter-

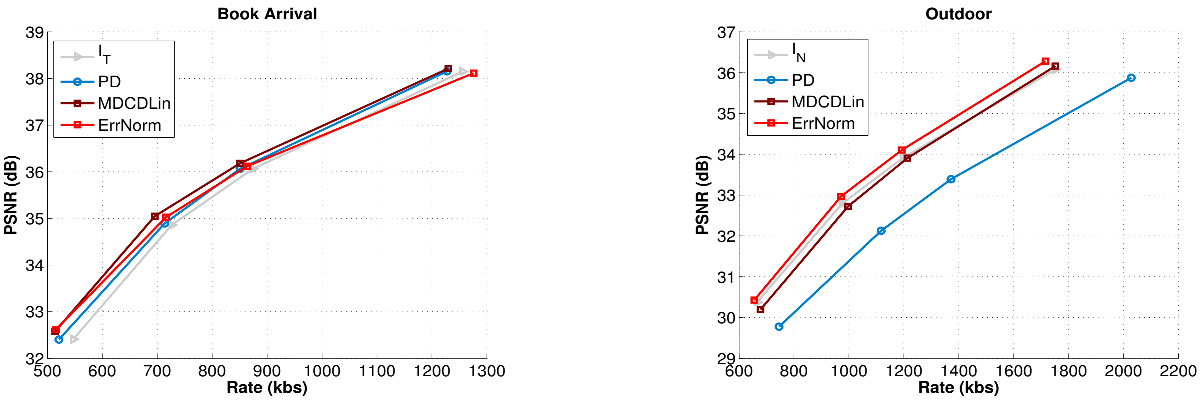

Figure 4: RD performances for 3 fusions methods and the best non-fusion estimation.

	Δ Rate (%)	Δ PSNR (dB)
PD	21.96	-0.84
MCDLin	2.24	-0.13
ErrNorm	-3.64	0.22

Table 3: Rate-distortion performance comparison between the different fusion methods and the inter-view non-fusion estimation for "Outdoor" sequence, obtained with the Bjontegaard metric [16].

	Δ Rate (%)	Δ PSNR (dB)
PD	-2.78	0.19
MCDLin	-6.07	0.37
ErrNorm	-3.13	0.20

Table 4: Rate-distortion performance comparison between the different fusion methods and the temporal non-fusion estimation for "Book Arrival" sequence, obtained with the Bjontegaard metric [16].

view and temporal estimation quality conditions, while existing methods loose their performances in the case of low temporal prediction quality. Future work will focus on comparing different approaches for dense field estimation in multi-view DVC framework [17].

REFERENCES

[1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. on Inf. Theory*, pp. 471–480, July 1973.

[2] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the receiver," *IEEE Trans. on Inf. Theory*, pp. 1–11, Jan. 1976.

[3] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. of the IEEE*, vol. 93, no. 71, pp. 71–83, Jan. 2005.

[4] T. Maugey and B. Pesquet-Popescu, "Side information estimation and new symmetric schemes for multi-view distributed video coding," *J. Vis. Comun. Image Represent.*, vol. 19, no. 8, pp. 589–599, 2008.

[5] A. Aaron, R. Zhang, and B. Girod, "Wyner-Ziv coding of motion video," in Proc. Asilomar Conference on Signals and Systems, Pacific Grove, California, Nov. 2002.

[6] T. Maugey, W. Miled, and B. Pesquet-Popescu, "Dense disparity estimation in a multi-view distributed video coding system," *Proc. of IEEE Intern. Conf. Acoust., Speech and Sign. Process.*, Apr. 2009, Taipei, Taiwan.

[7] A. Fusiello, E. Trucco, and A. Verri, "A compact algorithm for rectification of stereo pairs," *Int. J. Machine Vis. and Appl.*, vol. 12, no. 1, pp. 16–22, July 2000.

[8] M Ouaret, F Dufaux, and T Ebrahimi, "Fusion-based multiview distributed video coding," in *ACM Int. Workshop on Video Surveillance and Sensor Networks*, Santa Barbara, California, USA, Oct. 2006.

[9] X. Artigas, E. Angeli, and L. Torres, "Side information generation for multiview distributed video coding using a fusion approach," in *7th Nordic Signal Processing Symposium*, Iceland, June 2006.

[10] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Multiview Distributed Video Coding with Encoder Driven Fusion," Poznan, Poland, Sep. 2007.

[11] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Iterative multiview side information for enhanced reconstruction in distributed video coding," 2009, special issue on Distributed Video Coding.

[12] J.D. Areia, J. Ascenso, C. Brites, and F. Pereira, "Wyner-Ziv stereo video coding using a side information fusion approach," *IEEE MMSP*, pp. 453–456, Oct. Chania, Greece, 2007.

[13] X. Guo, Y. Lu, F. Wu, W. Gao, and S. Li, "Distributed multi-view video coding," *SPIE, Electr. Imaging*, Jan. 2006, San Jose, California, USA.

[14] P. Ferre, D. Agrafiotis, and D. Bull, "Fusion methods for side information generation in multi-view distributed video coding systems," *Proc. Inter. Conf. Image Process.*, vol. 6, pp. 409–412, Oct. 2007.

[15] I. Ingo Feldmann, P. Kauff, K. Mueller, M. Mueller, A. Smolic, R. Tanger, T. Wiegand, and F. Zilly, "HHI test material for 3D video," MPEG2008/M15413, April 2008, Airchamps.

[16] G. Bjontegaard, "Calculation of average PSNR differences between RD curves," Tech. Rep., 13th VCEG-M33 Meeting, Austin, TX, USA, Apr. 2001.

[17] M. Cagnazzo, T. Maugey, and B. Pesquet-Popescu, "A differential motion estimation method for image interpolation in distributed video coding," in *Proc. of IEEE Intern. Conf. Acoust., Speech and Sign. Process.*, Taipei, Taiwan, Apr. 2009.