DENSE DISPARITY ESTIMATION IN A MULTI-VIEW DISTRIBUTED VIDEO CODING SYSTEM

Thomas Maugey, Wided Miled and Béatrice Pesquet-Popescu

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

ABSTRACT

Distributed video coding (DVC) is a recent paradigm which aims at transferring part of the coding complexity from the encoder to the decoder. The performance of such a coding scheme strongly depends on the capacity to estimate correlation at the decoder and, consequently, on the side information quality. In this paper we consider a multi-view DVC framework and propose a very efficient dense disparity estimation technique for side information construction, based on a variational formulation. The simulation results show that our approach clearly outperforms the existing methods for inter-view side-information generation.

***Index Terms*—** Multi-view distributed video coding, disparity estimation, side information

1. INTRODUCTION

DVC is a promising technique based on results from information theory established in the 1970's [1, 2] stating that one can achieve the same coding performance by encoding two correlated sources independently and decoding them jointly, as by encoding *and* decoding them jointly. This would allow transferring part of the coding complexity from the encoder to the decoder. For multi-view video coding this is of particular interest, since cameras do not need to communicate with the each others. Many applications could benefit from it, such as video compression on mobile devices and video surveillance. Several works have been conducted to apply this theory to practical video coding, first for mono-view sequences, including [3], [4] and [5] and then for multi-view video [6, 7, 8, 9]. Our work fits within the framework defined in [3, 4]. When applied to a single video source, this method involves splitting the input video into two subsets, leading to two correlated sources: the key frames (KFs) and the Wyner-Ziv frames (WZFs), which are then separately encoded and jointly decoded. As presented in Fig.1, KFs are encoded and decoded using a classical Intra codec such as H.264 Intra. WZFs, on the other hand, are first transformed (typically using Discrete Cosinus Transform, DCT), quantized and then turbo-encoded. At the decoder side, WZFs are estimated from already decoded KFs, and the estimation, called Side Information (SI), is used as systematic bits and further refined using the parity bits sent by the turbo-encoder.

A similar coding strategy can be applied in the case of multi-view distributed video coding (MDVC). However, the presence of multiple video sources adds one dimension to the problem, thus raising new problems and potentially offering new solutions. Indeed, the construction of the SI now aims at exploiting both inter-view dependencies and temporal correlations. Many solutions in the literature propose specific methods for inter-view estimation, such as the homographic methods [7, 8], which estimate the homographic transformation between the views and use it to build the interpolation. Other techniques exist, [6], but they suppose the use of a particular frame disposition and they cannot be used with the scheme adopted in this work (a quincunx repartition of KFs and WZFs in the time-view space). Presently, one of the best and simplest techniques to generate inter-view interpolation uses block-based estimation techniques [6, 9], disparity vectors are estimated as the motion vectors, i.e. assuming that the disparity is blockwise constant and finding the best matching block. However, because this assumption does not always hold and the estimated disparity field does not provide a pixel-to-pixel mapping between left and right views, the interpolated images usually have visible artifacts. While in the classical multi-view video coding the cost of transmitting motion and disparity information prevented the expansion of dense (one vector per pixel) estimation methods, in MDVC this information is estimated only at the decoder, and therefore dense fields are not penalized compared with block-based ones. Obviously, the finer the disparity estimation, the better will be the generated SI.

In this paper, we adopt the previously described multi-view codec and we propose a dense disparity approach to generate a high quality inter-view estimation by using the disparity estimation method described in [10]. This method achieves good results compared with state-of-the-art methods, such as graph cuts and belief propagation based methods [10]. Based on a set theoretic framework, the proposed algorithm allows to incorporate various convex constraints, corresponding to a priori information, and yields disparity vectors with theoretically infinite precision. To obtain a smooth disparity field while preserving discontinuities, a total variation based regularization constraint is considered. The fusion of inter-view and temporal estimations is further studied and comparisons with state-of-the-art methods illustrate the gain of the proposed method.

The remaining of this paper is organized as follows: in Section 2, we explain the proposed side information generation method, and in Section 3, experiments ran on several multi-view test sequences show the performance of the proposed approach. Conclusions and future work directions are drawn in Section 4.

2. SIDE INFORMATION CONSTRUCTION

In this section, the side information estimation process for a WZF, denoted by W, is presented. We assume that four key frames are available for this estimation as presented in Fig. 2, which is possible in particular in the adopted frame repartition, i.e. a quincunx repartition of KFs and WZFs in the time-view space ([11]). Two estimations are computed: the temporal estimation is an interpolation between the previous KF, T_{-1}, and the next KF, T_{+1}. A second estimation is generated using the KF of the left view, V_l, and of the right view, V_r, as detailed in Section 2.2. The two estimations are then merged as explained in Section 2.3. The temporal estimation

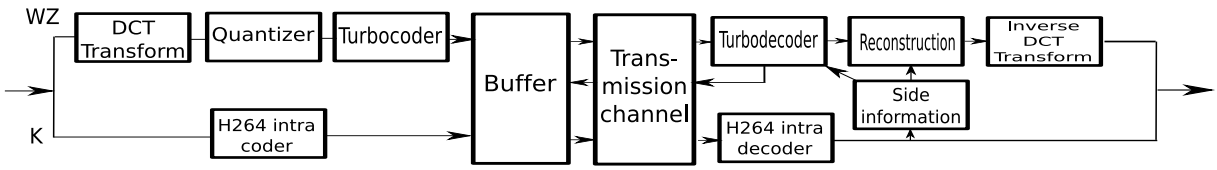


Fig. 1. DVC scheme.

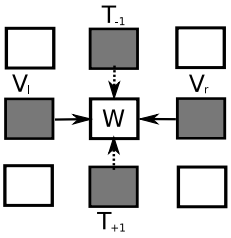


Fig. 2. Side information generation for the WZF (W). Dashed arrows: temporal estimation from the previous key frame, T_{-1}, and the next key frame, T_{+1}. Plain arrows: inter-view estimation from the left view, V_l, and the right view, V_r.

method used in this work is the one proposed by Ascenso et al. in [12]: after a temporal smoothing of the two reference frames, a classical block based forward motion estimation is performed between the two neighboring KFs. The obtained motion vector field is further refined by a bidirectional motion estimation to get symmetric motion predictions from the two KFs. The motion vector fields are then filtered with a weighted median filter, in order to eliminate the outliers and to get a smooth solution. Note that this method, which will be called MI in the sequel, has also been used for disparity estimation [9]. In the next section, we propose a dense disparity estimation method.

2.1. Dense disparity estimation method

Given a pair of stereo views, the dense disparity estimation problem is to estimate a 2D displacement field by searching for every pixel in the left view the corresponding pixel in the right view. When the two views are parallel, the vertical component of the disparity vector vanishes, so that only a scalar value has to be estimated. To estimate the disparity u, we minimize the following cost function, based on the sum of squared intensity differences:

$$J_0(u) = \sum_{(x,y)\in\mathcal{D}} [V_l(x,y) - V_r(x-u(x,y),y)]^2 \ , \qquad (1)$$

where $\mathcal{D} \subset \mathbb{N}^2$ is the image support. This expression is non-convex with respect to the displacement field u. Thus, in order to avoid a tough non-convex minimization problem, we consider a Taylor expansion of the non-linear term $V_r(x-u,y)$ around an initial estimate $\bar{u}$ (which, in our method, is obtained by a correlation method illustrated in Fig. 3(c)) as follows:

$$V_r(x-u,y) \simeq V_r(x-\bar{u},y) - (u-\bar{u})\ V_r^x(x-\bar{u},y) \ , \qquad (2)$$

where $V_r^x(x-\bar{u},y)$ is the horizontal gradient of the warped right image. Note that for notation concision, we have not made explicit that u and $\bar{u}$ are functions of (x,y) in the above expression.

With the approximation (2), the cost function J_0 under the minimization in (1) becomes quadratic in u. Thus, setting $\mathbf{s} = (x,y)$ the spatial position in either image, the objective function to be minimized can be rewritten as:

$$J(u) = \sum_{\mathbf{s}\in\mathcal{D}} [L(\mathbf{s})\ u(\mathbf{s}) - r(\mathbf{s})]^2 \qquad (3)$$

where

$$L(\mathbf{s}) = \nabla V_r^x(x-\bar{u}(\mathbf{s}),y)$$
$$r(\mathbf{s}) = V_r(x-\bar{u}(\mathbf{s}),y) + \bar{u}(\mathbf{s})\ L(\mathbf{s}) - V_l(\mathbf{s}).$$

Minimizing the objective function (3) aims at recovering the best estimate of the disparity image u from the observed fields L and r. This inverse problem is ill-posed due to the fact that the components of L may locally vanish. Thus, to convert this problem to a well-posed one, it is useful to incorporate additional constraints modelling prior knowledge and available information on the solution. In this work, we address the problem from a set theoretic formulation, where each constraint is represented by a convex set in the solution space and the intersection of these sets, the feasibility set, constitutes the family of possible solutions [10]. The aim then is to find an acceptable solution minimizing the given objective function. A formulation of this problem in a Hilbert image space $\mathcal{H}$ is therefore:

$$\text{Find } u \in S = \bigcap_{i=1}^{m} S_i \text{ such that } J(u) = \inf J(S) \ , \qquad (4)$$

where the objective $J : \mathcal{H} \to (-\infty, +\infty]$ is a convex function and the constraint sets $(S_i)_{1\le i\le m}$ are closed convex sets of $\mathcal{H}$. Constraint sets can generally be modelled as level sets:

$$\forall i \in \{1,\dots,m\}, \qquad S_i = \{u \in \mathcal{H} \mid f_i(u) \le \delta_i\} \ , \qquad (5)$$

where, for all $i \in \{1,\dots,m\}$, $f_i : \mathcal{H} \to \mathbb{R}$ is a continuous convex function and $(\delta_i)_{1\le i\le m}$ are real-valued parameters such that $S = \bigcap_{i=1}^m S_i \neq \emptyset$.

To solve the disparity estimation problem within the set theoretic framework described above, we incorporate, in what follows, the constraints modelling prior information on the estimated disparity field as closed convex sets of the form (5). The most common constraint on disparity is the knowledge of its range of possible values. Indeed, disparity values are nonnegative and often have known minimal and maximal amplitudes, denoted respectively by $u_{\min} \ge 0$ and $u_{\max}$. The associated set is

$$S_1 = \{u \in \mathcal{H} \mid u_{\min} \le u \le u_{\max}\} \ . \qquad (6)$$

Furthermore, in most stereo vision applications, the disparity map should be smooth in homogeneous areas while keeping sharp edges. This can be achieved with the help of a suitable regularization constraint. In this work, we make use of the total variation (tv) measure which recently emerged as an effective tool to recover smooth images in various image processing research fields. Practically, $\text{tv}(u)$ represents a measure of the lengths of the level lines in the image.

Hence, if u is known a priori to have a certain level of oscillation so that a bound τ is available on the total variation, controlling $\text{tv}(u)$, restricts the solutions to the convex set

$$S_2 = \{u \in \mathcal{H} \mid \text{tv}(u) \leq \tau\} \ . \tag{7}$$

The upper bound τ can be estimated with good accuracy from prior experiments and the considered minimization method is shown to be robust with respect to the choice of this bound [10].

In summary, we formulate the disparity estimation problem as the minimization of the quadratic objective function (3) over the feasibility set $S = \cap_{i=1}^{2} S_i$, where the constraint sets $(S_i)_{1 \leq i \leq 2}$ are given by equations (6) and (7). Many powerful optimization algorithms have been proposed to solve this convex feasibility problem. For the current work, we employ the constrained quadratic minimization method developed in [13] and particularly well adapted to our needs. However, due to space limitation, we will not describe the algorithm but the reader is referred to [10, 13] for more details. To illustrate this method, we show in Fig. 3 the original WZF, the initial disparity field $\bar{u}$ obtained by a correlation method, and the result of the algorithm. Note that the resulting disparity field u has real values with infinite precision. This figure also contains the result of the block-based estimation method proposed in [12].

2.2. Inter-view interpolation

After performing the disparity estimation between the two KFs of the neighboring views, the inter-view interpolation follows the same main steps as the temporal interpolation: the disparity fields corresponding to the WZF are estimated and used for a bidirectional disparity compensated prediction of W. The prediction $\widehat{W}_l$, respectively $\widehat{W}_r$, of W from the right, respectively, left view can thus be written, for the pixel $\mathbf{s}$, as:

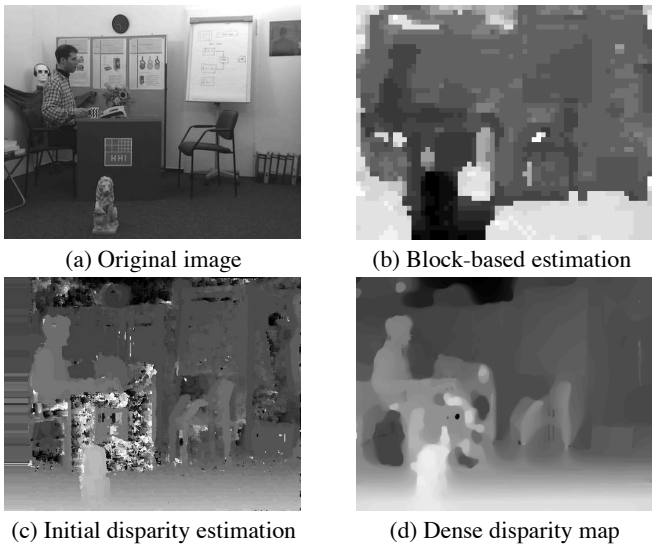
$$\widehat{W}_l(\mathbf{s}) = \widetilde{V}_r\left(x - \frac{1}{2}u(\mathbf{s}), y\right) \text{ and } \widehat{W}_r(\mathbf{s}) = \widetilde{V}_l\left(x + \frac{1}{2}u(\mathbf{s}), y\right) \tag{8}$$

where $\widetilde{V}_r$ and $\widetilde{V}_l$ are the disparity compensated right and left views, constructed by a B-spline interpolation. Then, the two interpolations $\widehat{W}_l$ and $\widehat{W}_r$ are merged. The fusion process consists in choosing for each pixel the best estimation according to a mean absolute error criterion. We denote in the following this disparity interpolation method by DI. In the ideal fusion (see Fig. 4(a)), the ideal fusion decision mask, d_I, using the original WZF, is computed as follows for each pixel $\mathbf{s}$:

$$d_I(\mathbf{s}) = \begin{cases} 0, \text{ if } |\widehat{W}_r(\mathbf{s}) - W(\mathbf{s})| \leq |\widehat{W}_l(\mathbf{s}) - W(\mathbf{s})| \\ 1, \text{ otherwise} \end{cases} \tag{9}$$

Obviously, this cannot be done in real conditions at the decoder but will serve as lower bound for our fusion process. In the real fusion, inspired by the work of Ouaret et al.[7], instead of the original frame, the decision mask, d_R, is estimated using the next frame T_{+1} (see Fig. 4(b)):

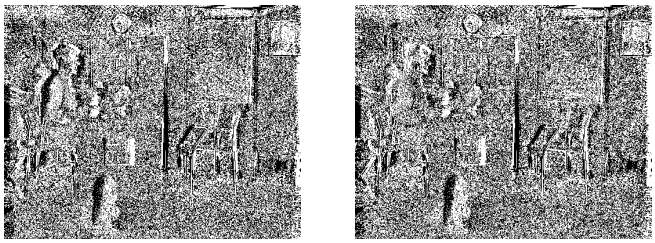
$$d_R(\mathbf{s}) = \begin{cases} 0, \text{ if } |\widehat{W}_r(\mathbf{s}) - T_{+1}(\mathbf{s})| \leq |\widehat{W}_l(\mathbf{s}) - T_{+1}(\mathbf{s})| \\ 1, \text{ otherwise} \end{cases} \tag{10}$$



(a) Original image (b) Block-based estimation

(c) Initial disparity estimation (d) Dense disparity map

Fig. 3. Disparity maps for the rectified "Book arrival" test sequence.



(a) Ideal fusion mask, d_I (b) Real fusion mask, d_R

Fig. 4. Fusion of the right and left disparity based interpolations for lossless reference frames. Black and white indicate the selection of the right, resp., left view.

The black and white values of the ideal mask in Fig. 4(a), which correspond to the pixels selected respectively from $\widehat{W}_r$ and $\widehat{W}_l$, clearly illustrate the occlusion zones around the objects of the scene. One can remark that the real fusion mask in Fig. 4(b) is quite similar to the ideal one, and in particular correctly finds the occlusion regions. Then, as presented in Tab. 1, even though the real disparity interpolation does not attain the performances of the ideal one, it outperforms the other existing methods by up to 2.5 dB. Indeed, the compensation with the dense disparity field provides a more precise SI for W and better follows the true geometry of the scene, as shown in Fig. 3.

Table 1. Inter-view estimation quality, in dB, for different methods (KFs are H.264 intra coded), "Book arrival" test sequence, 512×386.

QP of the Key Frames	Lossless	31	36	40
PSNR of KFs	∞	38.11	34.71	32.12
PSNR for homography	26.74	26.81	26.75	26.65
PSNR for MI	31.65	30.63	30.45	28.71
PSNR for DI (ideal fusion)	39.13	37.49	35.51	33.68
PSNR for DI (real fusion)	35.26	34.64	33.02	31.25

2.3. Fusion of temporal and inter-view estimations

At the decoder, we recall that two estimations are available for the side information construction: the temporal and the inter-view interpolation. The fusion problem is similar to the one previously described. The ideal fusion chooses the best estimation pixel by pixel

exploiting the original WZF. In this case, for the real fusion step, we did not choose the Ouaret method [7], which has quite low performance when motion activity is too important (as for "Outdoor" test sequence). The real fusion mask is build comparing $D_T = |T_{-1} - T_{+1}|$ and $D_V = |V_r - V_l|$. In other words, when the motion activity is too high, the corresponding pixel of the fusion is taken from the inter-view estimation, and when the disparity is too high, the pixel is taken from the temporal estimation. We have compared the fusion of the motion interpolation method used for both temporal and inter-view correlation, as in [9] (denoted by MI+MI), and our proposed method which consists of a fusion of the inter-view estimation presented in Section 2.2 and the motion interpolation presented above. This latter method is denoted by MI+DI and the comparison results are presented in Tab. 2. One can see that the MI+DI method leads to better side information than the classical MI+MI method.

Table 2. Side information quality, in dB, for different estimation methods (KFs are H.264 Intra coded), "Book arrival" test sequence, 512×386.

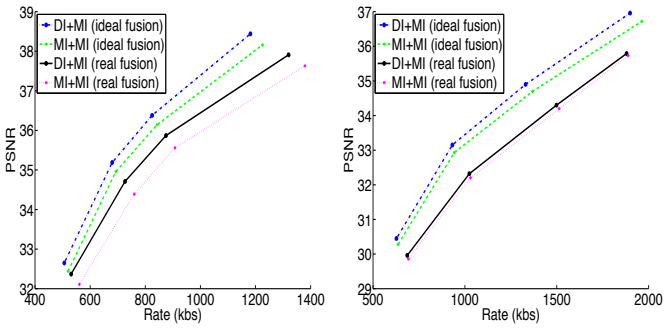
QP of the Key Frames	Lossless	31	36	40
PSNR for MI+MI (ideal fusion)	42.51	37.56	34.70	32.24
PSNR for MI+MI (real fusion)	36.63	33.27	32.53	30.66
PSNR for MI+DI (ideal fusion)	46.49	38.60	35.38	32.79
PSNR for MI+DI (real fusion)	36.27	35.05	33.27	31.34

3. EXPERIMENTAL RESULTS

In this section, we provide experimental results illustrating the coding performance of the proposed method. Experiments were run on two geometrically rectified multi-view test sequences: "Book arrival" and "Outdoor" [14]. For both sequences, the spatial resolution was reduced to 512×386, and only the first 7 cameras were used. The choice of minimal and maximal disparity amplitudes u_{min} and u_{max} was done by measuring the disparity at certain points of interest selected manually. The bound τ on the regularization constraint S_2 was fixed by calculating first the value of the associated convex function on the initial disparity field and then choosing a fixed ratio (20%) of this value. Rate-distortion curves for luminance components, shown in Fig.5, confirm the results obtained in Section 2. Note that, as all cameras play the same role and are equivalent (i.e., alternative coding of WZFs and KFs), the rate values correspond to the average rate per camera. One can see that the dense disparity estimation enables the DI+MI to outperform the MI+MI method. This shows that the dense disparity interpolation highly improves upon the block-based estimation, providing complementary information to the temporal estimation, in particular in the occlusion areas. However, for the "Outdoor" sequence, the real fusion mask is quite noisy, and the motion activity very important. Therefore, "bad" decisions will highly impact the final performance, which explains the RD curve closer to the MI+MI one in Fig. 5(b).

4. CONCLUSION

In this work we have proposed a new approach for multi-view DVC, which consists in using a dense disparity estimation method at the decoder. In comparison with the classical block based techniques, our method yields a better quality SI at the decoder, thus greatly improving the global coding performances. Future work will focus on proposing more efficient fusion methods in the multi-view framework, and on the extension of the disparity estimation method to non-rectified sequences.



(a) "Book arrival" sequence (b) "Outdoor" sequence

Fig. 5. Rate-distortion performance on two multi-view video test sequences (7 cameras, 384×512, 15 fps) for different side information estimation methods.

5. REFERENCES

[1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans. on Information Theory*, vol. 19, 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 Information Theory*, vol. 22, pp. 1–11, Jan. 1976.

[3] 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.

[4] 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.

[5] R. Puri and K. Ramchandran, "PRISM: A video coding architecture based on distributed compression principles," Tech. Rep. UCB/ERL M03/6, EECS Department, University of California, Berkeley, 2003.

[6] 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.

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

[8] X. Guo, Y. Lu, F. Wu, W. Gao, and S. Li, "Distributed multi-view video coding," *SPIE-IST Electronic Imaging, SPIE*, vol. 6077, pp. 15–19, Jan. 2006, San Jose, California, USA.

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

[10] W. Miled, J.-C. Pesquet, and M. Parent, "Disparity map estimation using a total variation bound," in *Proc. 3rd Canadian Conf. Comput. Robot Vis.*, Quebec, Canada, Jun. 2006, pp. 48–55.

[11] T. Maugey and B. Pesquet-Popescu, "Side information estimation and new symmetric schemes for multi-view distributed video coding," *J. Vis. Commun. Image Representation*, 2008.

[12] J. Ascenso, C. Brites, and F. Pereira, "Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding," in *EURASIP Conference on Speech and Image Processing, Multimedia Communications and Services*, Smolenice, Slovak Republic, June 2005.

[13] P.L. Combettes, "A block iterative surrogate constraint splitting method for quadratic signal recovery," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1771–1782, July 2003.

[14] 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.