Dense Disparity Estimation in Multiview Video Coding

I. Daribo, M. Kaaniche, W. Miled, M. Cagnazzo, B. Pesquet-Popescu

Telecom ParisTech
Signal and Image Proc. Dept.
37-39, rue Dareau, 75014 Paris, France
[daribo,kaaniche,miled,cagnazzo,pesquet]@telecom-paristech.fr

***Abstract*—Multiview video coding is an emerging application where, in addition to classical temporal prediction, an efficient disparity prediction should be performed in order to achieve the best compression performance. A popular coder is the multiview video coding (MVC) extension of H.264/AVC, which uses a block-based disparity estimation (just like temporal prediction in H.264/AVC). In this paper, we propose to improve the MVC extension by using a dense estimation method that generates a smooth disparity map with ideally infinite precision. The obtained disparity is then segmented and efficiently encoded by using a rate-distortion optimization technique. Experimental results show that significant gains can be obtained compared to the block-based disparity estimation technique used in the MVC extension.**

I. Introduction

A multiview video system consists in generating multiple views by capturing from different viewpoints the same scene via a set of multiple cameras. A set of slightly different views can be used to reproduce the scene in three dimensions.

The improvement of 3D technologies raised interest in 3D television (3DTV) [1] and in free viewpoint video (FVV) [2]. While 3DTV offers depth perception of program entertainments without wearing special additional glasses, FVV allows the user to freely change his viewpoint position and viewpoint direction around a 3D reconstructed scene. Other target fields are expected, like Digital Cinema, IMAX theaters, medicine, dentistry, air-traffic control, military technologies, computer games, etc.

In the meantime, the digital TV technology and 3D displays have largely improved recently, making even more relevant the problem of multiview applications. Capturing, processing and coding multiview video are now very active research topics. In particular, in sight of the huge amount of data concerned, compression assumes a paramount importance.

A straight method to compress multiview video is to encode each view independently using the state-of-the-art H.264/AVC encoder [**?**]. This approach is denoted, in the literature, as simulcast coding. However, since all the cameras capture the same scene through different viewpoints, there is an inter-view statistical expected dependencies between adjacent cameras, which is not exploited in the simulcast case.

MMSP'09, October 5-7, 2009, Rio de Janeiro, Brazil.
978-1-4244-4464-9/09/$25.00

In order to better deal with inter-view redundancy, the joint video team (JVT) is developing a multiview extension of H.264/AVC standard, known as multiview video coding (MVC) extension [3]. The aim of this extension is to provide new techniques improving coding efficiency taking advantage of both temporal and inter-view redundancies, thus leading to additional coding gain compared to the H.264/AVC simulcast solution.

On the other hand, adding the inter-view prediction to the temporal one requires more computational and memory resource. Nevertheless, it has been shown [4] that most of the coding gain of MVC comes from the inter-view prediction of the temporal *intra* picture, while for the *inter* pictures the temporal prediction is the most efficient prediction mode. As a consequence, limiting the inter-view prediction to the intra pictures is commonly reputed as a reasonable complexity/efficiency trade-off.

The inter-view correlation between adjacent cameras is removed via the so-called disparity estimation (DE) and compensation. Two main approaches, block-based and dense, have been used to estimate the disparity vectors (DV). A survey of the different techniques proposed in the literature can be found in [5]. The MVC extension employs a variable block-based disparity estimation, assuming that within each partition of the current macroblock the disparity vector is constant. However, this assumption does not always hold, especially around depth discontinuities and in textureless regions. Dense pixel-based approaches attempt to overcome this drawback by assigning one disparity vector to each pixel. Of course this means that the disparity map would require a very high bit-rate to be encoded: for this reason, dense DVs have rarely been consider for compression. The basic idea of this paper is to reduce the coding cost of the dense DV map by operating a RD-driven segmentation on it.

In particular, we propose to improve the disparity prediction unit in the MVC extension by using the dense DE (DDE) method described in [6] (because it achieves good results compared with the state-of-art methods, such as graph cuts and belief propagation based methods) followed by the segmentation step. Based on a set theoretic framework, this DDE approach incorporates various convex constraints corresponding to *a priori* information such as the range of DVs or the total variation regularization constraint which assures a smooth

disparity field while preserving discontinuities.

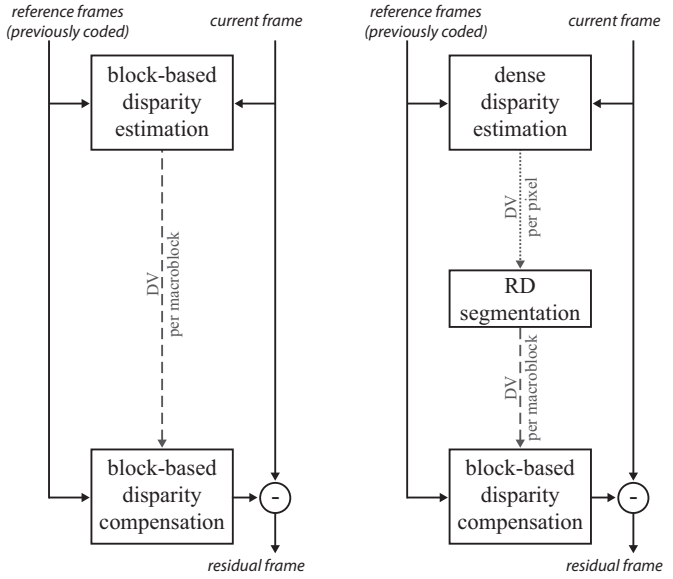


Fig. 1. Disparity prediction: (left) block-based estimation, (right) enhanced by a dense estimation.

The proposed scheme is summarized in Fig. 1: we replace the block-based disparity estimation (BDE) stage by a dense disparity estimation (DDE) one, and then, we apply a rate-distortion segmentation to the generated disparity map. This is performed by optimizing a Lagrangian cost function which takes into account the accuracy and the coding cost of the disparity map.

The remainder of this paper is organized as follows. Section II provides details about the DDE method. In Section III, we address the problem of the RD-optimized segmentation and encoding of the disparity field. Finally, in Section IV, we give experimental results confirming the effectiveness of the proposed method, while Section V draws conclusions and outlines future work.

II. Dense Disparity Estimation

A. Problem statement

Let $I^{(n,t)}$ and $I^{(n-1,t)}$ be two frames taken respectively by the n-th and $(n-1)$-th cameras at time t. We assume that cameras are rectified, so that the disparity vectors are restricted to the horizontal component, that will be denoted by d. DDE methods attempt to determine, for each pixel in the current frame $I^{(n,t)}$, the best corresponding pixel in the reference frame $I^{(n-1,t)}$. Generally, the estimation is obtained by minimizing a given cost functional, formulated in terms of the sum of squared differences:

$$\hat{d} = \arg\min_{d\in\Omega} \sum_{(x,y)\in\mathcal{D}} [I^{(n,t)}(x,y) - I^{(n-1,t)}(x+d,y)]^2 \quad (1)$$

where $\mathcal{D}$ is the picture support and Ω is the range of candidate disparity values. Generally, an initial estimate $\bar{d}$ of d is available, for example using a dense correlation-based method. Assuming that the magnitude difference of both fields is relatively small, the warped reference frame is approximated around $\bar{d}$ by a Taylor expansion:

$$I^{(n-1,t)}(x+d,y) \simeq I^{(n-1,t)}(x+\bar{d},y) + \nabla I_x^{(n-1,t)}(x+\bar{d},y)(d-\bar{d}) \quad (2)$$

where $\nabla I_x^{(n-1,t)}(x+\bar{d},y)$ is the horizontal gradient of the warped reference frame. Note that in (2), we have not made explicit that d and $\bar{d}$ are functions of $s=(x,y)$ for notation concision. Using the linearization (2), the criterion $\tilde{J}$ in (1) can be approximated by the quadratic convex functional J in d:

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

where

$$\begin{aligned} L(s) &= \nabla I_x^{(n-1,t)}(x+\bar{d}(s),y) \\ r(s) &= I^{(n,t)}(s) - I^{(n-1,t)}(x+\bar{d}(s),y) + \bar{d}(s)\ L(s) \end{aligned}$$

The minimization of this quadratic functional is an ill-posed problem as the components of L may locally vanish. Thus, to convert this problem to a well-posed one, we incorporate additional constraints reflecting the prior knowledge about the disparity field. In this work, we address the problem through a set theoretic framework [6]. Firstly, each constraint is represented by a closed convex set S_m with $m \in \{1,\dots,M\}$, in a Hilbert space $\mathcal{H}$. The intersection S of all the M sets S_m constitutes the family of possible solutions. Therefore, the constrained problem amounts to find the solution in S which minimizes the functional J:

$$\text{Find } \hat{d} \in S = \bigcap_{m=1}^{M} S_m \text{ such that } J(\hat{d}) = \min_{d\in S} J(d). \quad (4)$$

The constraint sets are modeled as level sets:

$$\forall m \in \{1,\dots,M\}, \qquad S_m = \{d \in \mathcal{H} \mid f_m(d) \le \delta_m\} \quad (5)$$

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

Hence, it is required to define the convex sets S_m to proceed to the DDE algorithm within the set theoretic framework. At this level, it is important to emphasize the great flexibility in incorporating any set of arbitrary convex constraints. In what follows, we will focus on $M=2$ constraints. The first one consists of restricting the variation of the disparity d within a specified range $[d_{\min}, d_{\max}]$. It can be expressed by the following constraint set S_1:

$$S_1 = \{d \in \mathcal{H} \mid d_{\min} \le d \le d_{\max}\} \quad (6)$$

Most importantly, a constraint can be incorporated in order to strengthen the smoothness of the disparity field in the homogeneous areas while preserving edges. Indeed, neighboring pixels belonging to the same object should have similar disparities. This can be achieved by considering the total variation $\mathrm{tv}(d)$ which can be defined as the sum over $\mathcal{D}$ of

the norm of the spatial gradient of d [7]. The total variation of the discrete disparity image $d = [d^{i,j}]$ is given by:

$$\begin{aligned}\mathsf{tv}(d) = &\sum_{i=0}^{W-2}\sum_{j=0}^{H-2}\sqrt{|d^{i+1,j} - d^{i,j}|^2 + |d^{i,j+1} - d^{i,j}|^2} \\ &+ \sum_{i=0}^{W-2}\sqrt{|d^{i+1,H-1} - d^{i,H-1}|} \\ &+ \sum_{j=0}^{H-2}\sqrt{|d^{W-1,j+1} - d^{W-1,j}|}\end{aligned}$$

where $W \times H$ is the support of the disparity image. Hence, a total variation based regularization constraint amounts to impose an upper bound τ on the tv of the image, leading to the following constraint set:

$$S_2 = \{d \in \mathcal{H} \mid \mathsf{tv}(d) \leq \tau\} \tag{7}$$

It is worth pointing out that the positive constant τ can be estimated for example through a learning procedure on image databases [8]. However, in our case we choose the value maximizing the quality of the disparity compensated picture, as shown in next section.

Finally, the disparity estimation problem is formulated by minimizing the quadratic objective function J in Eq. 3 under the mentioned constraint sets. To solve this problem, we used the efficient constrained quadratic minimization technique developed in [9], which is adapted to problems with quadratic convex objective functions. For the sake of brevity, we will not describe the algorithm here: for more details, the reader is referred to [9], [6]. In order for this algorithm to converge, the objective function J must be strictly convex. We introduce an additional term in order to assure this condition:

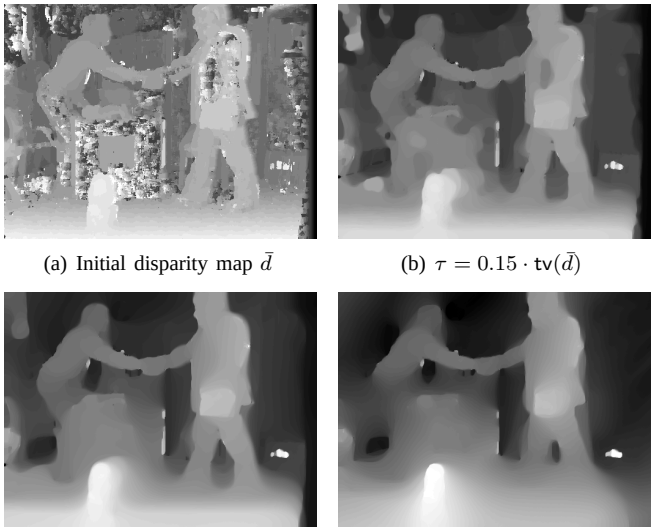
$$J(d) = \sum_{s\in\mathcal{D}}[r(s) - L(s)\ d(s)]^2 + \alpha\sum_{s\in\mathcal{D}}[d(s) - \bar{d}(s)]^2 \tag{8}$$

where $\bar{d}$ is an initial estimate and α is a positive real number: when it is large, we favor the regularization term and tend to have a final solution close to the initialization; on the contrary, when α is small, the data attachment term becomes dominant, and the solution can diverge from the initialization.

B. Influence of the parameters

In practice, the optimal value of the parameters $[d_{\min}, d_{\max}]$, τ and α may not be known exactly and it is, therefore, important to evaluate their impact in terms of coding rate and PSNR of the compensated picture. The choice of the range $[d_{\min}, d_{\max}]$ can be accurately found by matching certain points of interest selected manually in the two stereo frames. The upper bound τ, used to enforce the smoothness of the estimated disparity map, may be estimated from a scale value of the total variation of the initial disparity map $\bar{d}$, as shown in Fig. 2. A low scale value results in smoothing more the disparity map, and so, reducing the number of bit required for the transmission.

Table I and Table II show the impact of the parameters τ and α on the coding rate of the disparity map and on the quality of the disparity compensated picture, evaluated as PSNR between the original view and the disparity-compensated estimation. First, an arbitrary fixed value of α is used to determine the parameter τ. Then, the optimal value of α is determined. The value of both parameters is selected according on the highest PSNR value of the disparity compensated picture. Note that, in Table I and Table II, the bitrate of the dense disparity map has been computed using the dense disparity map as a picture with H.264/AVC in *intra* mode at a QP value of 0.



(a) Initial disparity map $\bar{d}$ (b) $\tau = 0.15 \cdot \mathsf{tv}(\bar{d})$

(c) $\tau = 0.10 \cdot \mathsf{tv}(\bar{d})$ (d) $\tau = 0.05 \cdot \mathsf{tv}(\bar{d})$

Fig. 2. Example of dense disparity maps at different values of the upper bound τ parameter (from "Book arrival" sequence, frame 36).

TABLE I
EXAMPLE OF THE INFLUENCE OF THE PARAMETER τ ON THE BITRATE AND THE PSNR OF THE DENSE DISPARITY COMPENSATED PICTURE (FROM "BOOK ARRIVAL" SEQUENCE, FRAME 36).

	bitrate (H.264/AVC intra, QP=0)	PSNR
$\tau = 50000$	0.6416 bpp	37.13 dB
$\tau = 40000$	0.5222 bpp	37.23 dB
$\tau = 30000$	0.4036 bpp	37.24 dB
$\tau = 20000$	0.3051 bpp	36.48 dB
$\tau = 10000$	0.2878 bpp	33.44 dB

TABLE II
EXAMPLE OF THE INFLUENCE OF THE PARAMETER α ON THE BITRATE AND THE PSNR OF THE DENSE DISPARITY COMPENSATED PICTURE (FROM "BOOK ARRIVAL" SEQUENCE, FRAME 36).

	bitrate (H.264/AVC intra, QP=0)	PSNR
$\alpha = 0.1$	0.4667 bpp	35.79 dB
$\alpha = 6$	0.4036 bpp	37.24 dB
$\alpha = 10$	0.3943 bpp	37.20 dB
$\alpha = 50$	0.3905 bpp	36.50 dB
$\alpha = 100$	0.3837 bpp	35.78 dB

III. RATE-DISTORTION-BASED SEGMENTATION

The purpose of the segmentation is to obtain, from the dense map produced by the algorithm described in the previous section, another map, which will be compatible with the rep-

resentation of motion vectors in H.264/AVC. This means that we have to segment the disparity map using the macroblock (MB), block and subblock shapes defined in the standard.

A. Block-based representation

As defined in the norm of H.264/AVC, the vector field can be described using a single vector per macroblock (16×16 pixels). However the MB can be partitioned in 16×8, 8×16 or 8×8 blocks, which can have a different vector each. Finally, block can be split into 8×4, 4×8 and 4×4 subblocks, which in turns can have a single vector.

B. Partition-based segmentation

As mentioned earlier, the dense disparity estimation method generates a map with real valued disparity vectors. A first approximation consists in truncating the precision with a quarter-pixel accuracy as in H.264/AVC standard. The disparity in the current MB is then represented by 256 quarter-pel vectors. Then, for any partition we have to choose a single vector from those of the dense representation. For example we have to choose one vector for the 16×16 partition, 2 for each of the 16×8 and 8×16 partitions and so on. In particular for any partition we consider the set of dense vectors that falls within it. Among them, we select 6 candidates: the average vector, the median (in the sense of the norm) vector, and the four vector whose norm is closest to the median one.

C. Rate-distortion optimization

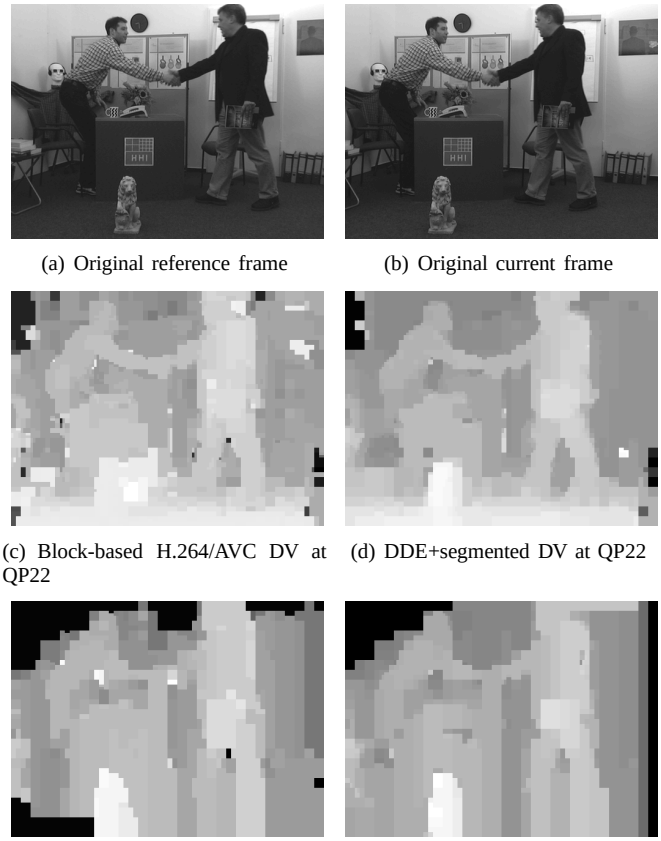
Let the quantization parameter QP be given. The distortion measure is the sum of squared intensity differences (SSD), and R the number of bits to be transmitted for the predictive disparity vector error. For the p^{th} partition B_k^p of the k^{th} macroblock B_k, the best approximated disparity vector $\hat{\mathbf{d}}$ is computed by minimizing the following:

$$\hat{\mathbf{d}} = \arg\min_{\mathbf{d}\in\Omega} J_{\mathbf{d}}(B_k^p|\mathrm{QP}) \tag{9}$$
$$\text{with} \quad J_{\mathbf{d}}(B_k^p|\mathrm{QP}) = \mathrm{SSD}(B_k^p, \mathbf{d}|\mathrm{QP}) + \lambda \cdot R(B_k^p, \mathbf{d}|\mathrm{QP})$$

The main relevance of the segmentation of the dense disparity map which finally ends in a block-based representation is to make good use of the smoothness of the dense disparity map. Indeed, despite the segmentation process reduces the quality of the solution provided by the dense disparity estimation algorithm, the resulting map, followed by a RD segmentation still can be consider as a good quality/bitrate trade-off representation over a direct block-based estimation. Furthermore, unlike H.264/AVC in which the disparity estimation is causal and local, our proposed disparity estimation has a global approach which favors the regularization of the disparity field. As a consequence, more MB will be coded in the SKIP mode, which is particularly efficient when the vector field is regular, since it consists in sending no side information nor residual: the vector is computed as the median of neighbors, and the block is copied from the compensated position of the original frame. The proposed method takes advantage from the augmented effectiveness of the SKIP mode which will be selected quite often, resulting in a remarkable rate reduction (see next section).

IV. Experimental Results

In this section, we provide some simulation results to evaluate the rate-distortion performance of the proposed structure. The experiments were run on three rectified multiview video sequences : “Book arrival”, “Door flowers” and “Outdoor” [10]. For all the video sequences, we use four views with a spatial resolution reduced to 512×384. We use the software JMVM 8.0 [11].



(a) Original reference frame (b) Original current frame

(c) Block-based H.264/AVC DV at QP22 (d) DDE+segmented DV at QP22

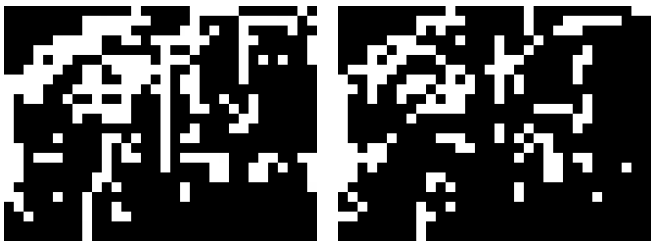
(e) Block-based H.264/AVC DV at QP42 (f) DDE+segmented DV at QP42

Fig. 3. Example of block-based disparity vectors (from “Book arrival” sequence, frame 36).

As seen in Section II-B, the parameters $[d_{\min}, d_{\max}]$, τ and α have to be chosen, and we have determined them heuristically. However they can be adjusted for each different sequence, by specifying their value in the Sequence Parameter Set (SPS). The increase in bitrate related to this side information is very small, such that we neglect this contribution when reporting experimental data about the coding rate.

Within the H.264/AVC framework, the rate-distortion estimation of the disparity vector generates different disparity fields at different QP values (Fig. 3). Disparity fields are usually smooth at low bit-rate which favors the selection of the SKIP mode. At high bit-rate, the distortion is privileged against the cost of the predictive disparity error which reduces

(a) H.264/AVC SKIP map at QP22 (b) DDE+segmented SKIP map at QP22



(c) H.264/AVC SKIP map at QP42 (d) DDE+segmented SKIP map at QP42

Fig. 4. Example of SKIP map correspondig to the disparity vectors field in Fig. 3. In black there are the SKIP macroblocks and in white the inter-macroblocks.

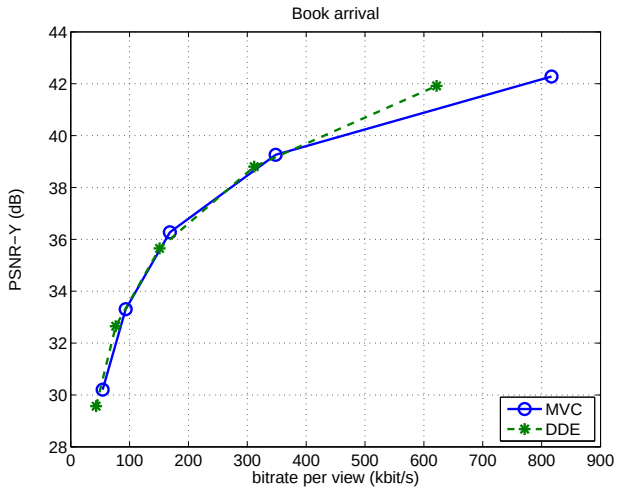
the number of SKIP macroblocks. We present a comparison in Fig. 4 at two QP points: 22 and 42. We can see in black the SKIP macroblocks. Our method has the benefit to generate a smooth block-based representation of the disparity vectors field at high bit-rate, which reduces the predictive disparity error, and subsequently uses more SKIP macroblocks. Especially at high bit-rate, when our method is used, the number of SKIP macroblocks increases, with a beneficial effect on the required coding rate. For example at QP=22 on the multiview video sequence "Book arrival" (Fig. 4), with the proposed method 58% of MB are coded in the SKIP mode, with respect to a mere 16% for the original encoder. At QP=42, we obtained a percentage of 78% against 71%.

Fig. 5 shows the results in terms of rate-distortion performance. Comparing the dense disparity estimation to the block-based reference H.264/AVC estimation clearly indicates the benefits of a dense estimation followed by a segmentation optimized for rate-distortion efficiency, especially for the "Outdoor" sequence, where a coding gain of 1.5 dB is achieved. The curve consists of 5 QP points which are 22, 27, 32, 37, 42.

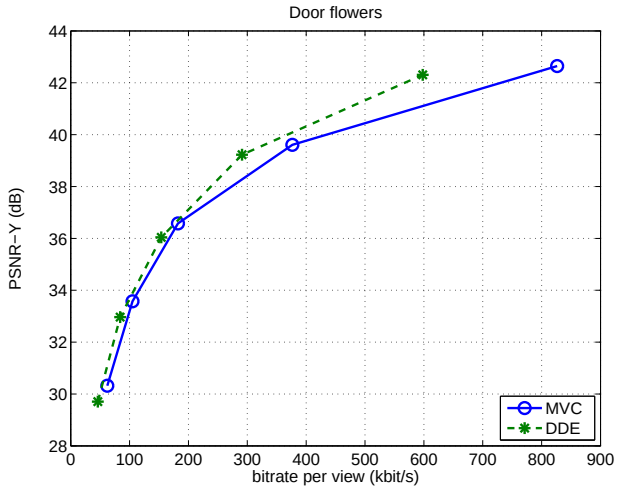
In addition, to measure the relative gain we used the Bjontegaard metric [12]. The results are shown in Table III for low bitrate and high bitrate corresponding respectively to the four QP points 27, 32, 37, 42 and 22, 27, 32, 37. We can see that our method works especially well on the sequence "Outdoor" (in which the disparity range is small, $[d_{\min}, d_{\max}] = [0, 8]$).

V. Conclusion

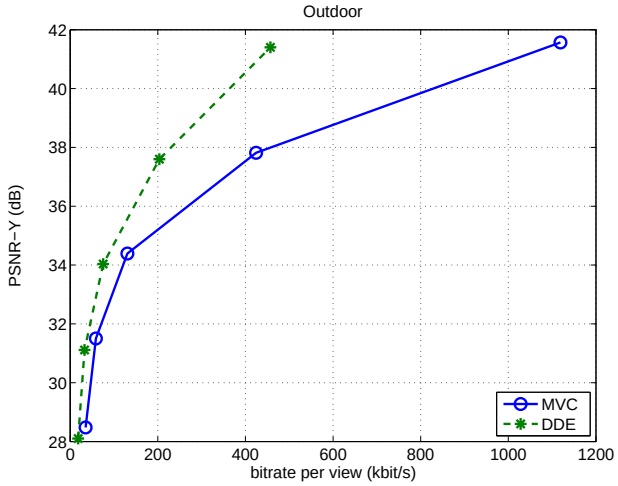
In this paper, we have presented the benefits of using a dense disparity estimation followed by a block-based segmentation



(a) Book arrival



(b) Door flowers



(c) Outdoor

Fig. 5. Rate-distortion coding results.

TABLE III
CALCULATION OF AVERAGE PSNR DIFFERENCES AND THE BITRATE SAVING.

	bitrate saving		**PSNR gain**	
	low	**high**	**low**	**high**
Book arrival	-1.49 %	-2.86 %	0.04 dB	0.10 dB
Door flowers	-12.83 %	-10.86 %	0.58 dB	0.52 dB
Outdoor	-60.03 %	-45.58 %	1.93 dB	1.59 dB

and coding of the disparity field in multiview video coding. As expected, a dense disparity estimation produces a smooth disparity field with an ideally infinite precision. This field is then presented with a quarter pixel precision and segmented based on an RD-optimized fashion. The smooth property of the estimated disparity vectors field allows a reduction of the bit-rate cost of the disparity vectors with a small reduction of the quality of the reconstructed picture.

Future work will focus on the RD selection of the disparity estimation parameters, and on the introduction of a new coding mode using the DDE and inspired to the inter-view DIRECT mode.

REFERENCES

[1] C. Fehn, E. Cooke, O. Schreer, and P. Kauff, "3D analysis and image-based rendering for immersive TV applications," *Signal Processing: Image Communication*, vol. 17, no. 9, pp. 705 – 715, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/B6V08-4717R8N-5/2/b912b9b1e7094297120ae3c9b9bb22f9

[2] J. Carranza, C. Theobalt, M. A. Magnor, and H. peter Seidel, "Free-viewpoint video of human actors," in *ACM Transactions on Graphics*, 2003, pp. 569–577.

[3] Y. Chen, Y.-K. Wang, K. Ugur, M. Hannuksela, J. Lainema, and M. Gabbouj, "The emerging MVC standard for 3D video services," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, 2009.

[4] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1461–1473, 2007.

[5] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV)*, 2001, pp. 131–140.

[6] W. Miled, J.-C. Pesquet, and M. Parent, "A convex optimization approach for depth estimation under illumination variation," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 813–830, 2009.

[7] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, no. 1-4, pp. 259–268, 1992.

[8] P. Combettes and J.-C. Pesquet, "Image restoration subject to a total variation constraint," *IEEE Trans. Image Process.*, vol. 13, no. 9, pp. 1213–1222, 2004.

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

[10] I. Feldmann, M. Müller, F. Zilly, R. Tanger, K. Muller, A. Smolic, P. Kauff, and T. Wiegand, "HHI test material for 3D video," Archamps, France, May 2008.

[11] A. Vetro, P. Pandit, H. Kimata, A. Smolic, and Y. Wang, "Joint multiview video model (JMVM) 8.0," JVT-AA207, Geneva, Apr. 2008.

[12] G. Bjontegaard, "Calculation of average PSNR differences between RD curves," Austin, TX, USA, Apr. 2001, iTU SG16 VCEG-M33.