DEPTH MAP CODING BY DENSE DISPARITY ESTIMATION FOR MVD COMPRESSION

Marco Cagnazzo, Béatrice Pesquet-Popescu

Institut TELECOM – TELECOM ParisTech
Signal and Image Processing Department
(Département de Traitement du Signal et Image, TSI)
46 rue Barrault 75634 Paris Cedex FRANCE

ABSTRACT

In multiview video representation, one of the most popular format is the so-called multiple view video plus depth. This representation is made up of N image sequences, each accompanied by a sequence of depth maps, telling the distance of each represented pixel from the observing camera. The depth maps are needed at the decoder side in order to generate intermediate views and therefore to enrich the user experience. This format is very flexible but also very demanding, in terms of storage space or and transmission bandwidth. Therefore, compression is needed.

At this end, one of the key steps is an efficient representation of depth maps. In this work we build over a proposed method for multiple view video coding, based on dense disparity estimation between views. This allows us to obtain a compact and high-quality depth map representation. In particular we explore the complex relationship between estimation and encoding parameters, showing that an optimal parameter set exist, that allows a fine-tuning of the estimation phase and an adaption of its results to the subsequent compression phase. Experiments are encouraging, showing remarkable gain over simple methods such as H.264/AVC simulcast, and even some gain with respect to more sophisticated techniques such as MVC.

Index Terms— Multiview video plus depth, depth map, multiview video coding, disparity estimation

1. INTRODUCTION

One of the most popular representations of multiview video and arguably the most adapted to free-viewpoint television [1] is the so-called multiple-views-plus-depth (MVD) format [2, 3]. When MVD is used, for each view we dispose of a texture video sequence and of a depth map sequence, representing, for each temporal instant, the distance of the current pixel from the point of observation. An example of MVD video is shown in Fig. 1.

Thanks to ANR agency for funding via the PERSEE project *Projet ANR-09-BLAN-0170.*

MVD is extremely demanding in terms of storage space and transmission bandwidth, therefore compression is mandatory in order to manage this representation. Several approaches exist for MVD compression [4]. A simple, first one, is to independently compress each texture and depth sequence from each view. This approach is commonly referred to as Simulcast, see Fig. 1(a). Simulcast has the advantage of being simply implementable, backward compatible, and of allowing to decode immediately a single view for 2D screens. It has been chosen as reference in the Call for Proposal issued by the MPEG committee for the standardization of MVD [5]. Of course, one expects that more sophisticated schemes, taking into account the redundancy between views and between texture and depth, would achieve far better compression performance than the multicast scheme (this is actually the rationale behind the CfP). For example, as shown in Fig. 1(b), one can apply H.264/MVC [6] over texture sequences and (separately) over depth maps. Since depth and texture have very different content, no coding gain is expected by jointly coding texture and depth with H.264/MVC. Nevertheless, some redundancy between texture and depth does exist, and this scheme does not exploit it. For example they partially share movement and disparity information, and above all, rate allocation between them should be jointly performed. However the latter is a quite difficult issue, and one of the key problems to be solved in order to achieve efficient coding [7].

In this paper we consider a MVD compression scheme inspired by our previous work on multiview video (without depth) [8]. We exploit dense disparity estimation to obtain RD efficient prediction not only between textures (belonging to different views) but also between depth maps. Moreover we explore the relationship between the estimation parameters and the compression ones. This study results in a coding paradigm providing competitive performances with respect to the state of the art.

The rest of the paper is organized as follows. Section 2 recalls the principles of the dense disparity estimation algorithm we use in the proposed method. This is useful to give insight about the relationships between estimation and compression parameters. Then, the proposed scheme is shown in Section 3.

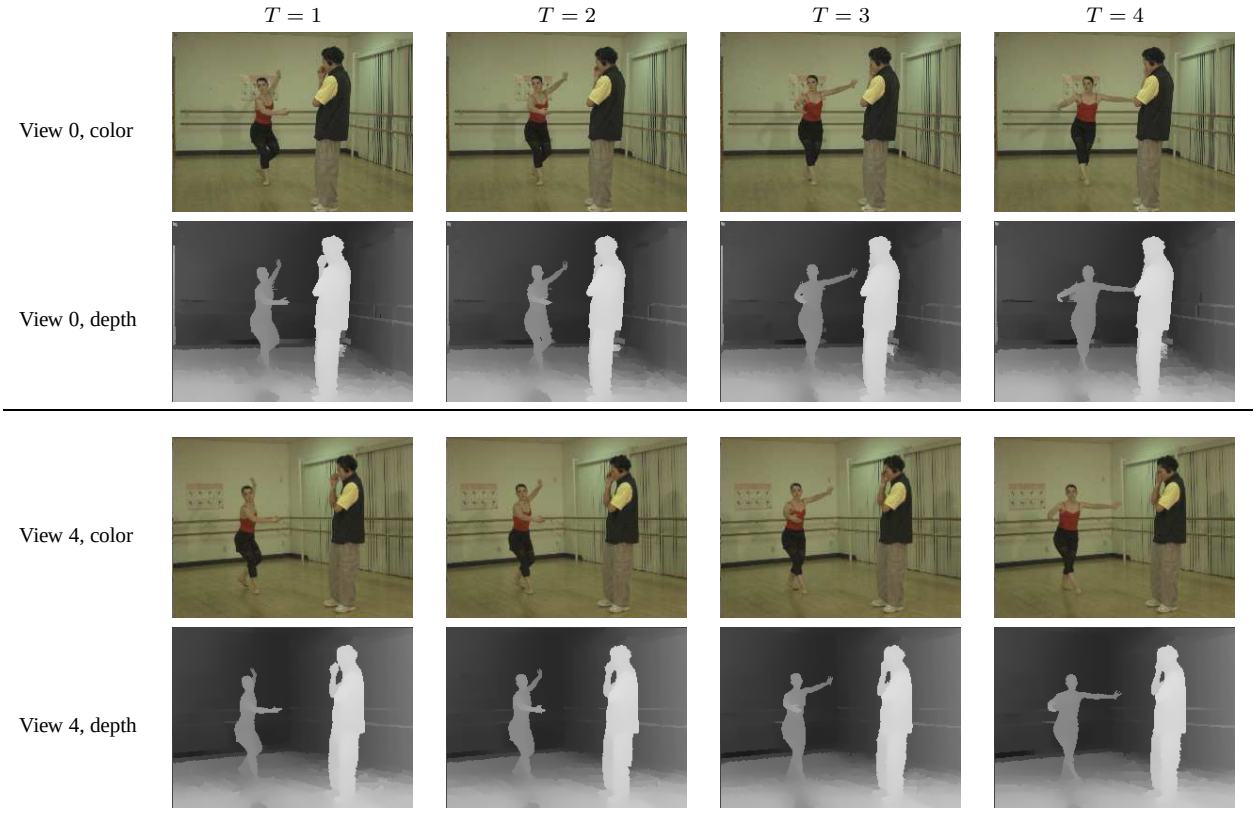

Fig. 1. Example of multiple-view-plus-depth video.

In particular we describe the reference encoder originally proposed in the case of multiview video, and then we show the proposed changes needed to efficiently take into account the depth information. The experimental results are reported in Section 4 where we show that a simple relationship can be inferred between the parameters of the dense estimation algorithm and of the compression algorithm. Thereafter, we provide the depth map coding performances and finally the global RD results, compared with some reference scheme. Finally, Section 5 draws conclusions and ends the paper.

2. DENSE DISPARITY ESTIMATION

In this section we describe a method for dense (*i.e.* one vector per pixel) disparity field estimation. Even though dense disparity fields cannot be directly used for compression because of their huge coding cost, we can use them to derive an R-D efficient representation of the disparity field, taking advantage of the global formulation of the disparity estimation problem. In other word we are able to produce a disparity field that standard approach would not take in consideration given their local and causal approach to the estimation problem. We have proved the effectiveness of this approach for multiple view video coding (without depth) in a previous work [8]. In the present paper we want to extend this concept to the MVD case and moreover to explore the critical issue of parameter tuning for the dense disparity estimation (DDE). At this end, it is necessary to recall the main ideas of DDE, which is the objective of this section.

Let I_t^n be the rectified frame taken by the n-th camera at time t. Therefore the disparity vectors can only have a the horizontal component, which we call d. Dense disparity estimation has the target of finding the disparity field $d(\mathbf{p}) = d(x, y)$ (that is the disparity vector for any pixel position $\mathbf{p} = (x, y)$) which best matches pixel $\mathbf{p}$ in current frame I_t^n in view n with pixel $\mathbf{p} + d(\mathbf{p})$ in the reference frame I_t^m in view m. This is a typical example of inverse problem, which needs suitable regularization to be solved. In the following, for the sake of simplicity, we will consider only the case $m = n - 1$.

At the basis of the estimation methods, there is the hypothesis that the image intensity is roughly constant once one has compensated for the disparity. As a consequence, a common method to estimate d is to minimize a cost function such as the sum of squared differences between the current image and the one compensated by disparity.

$$d^*(\cdot) = \underset{d \in \Omega}{\operatorname{argmin}} \sum_{(x,y) \in \mathcal{P}} [I_t^n(x, y) - I_t^{n-1}(x + d(x, y), y)]^2 \tag{1}$$

where $\mathcal{P}$ is the picture support and Ω is the range of candi-

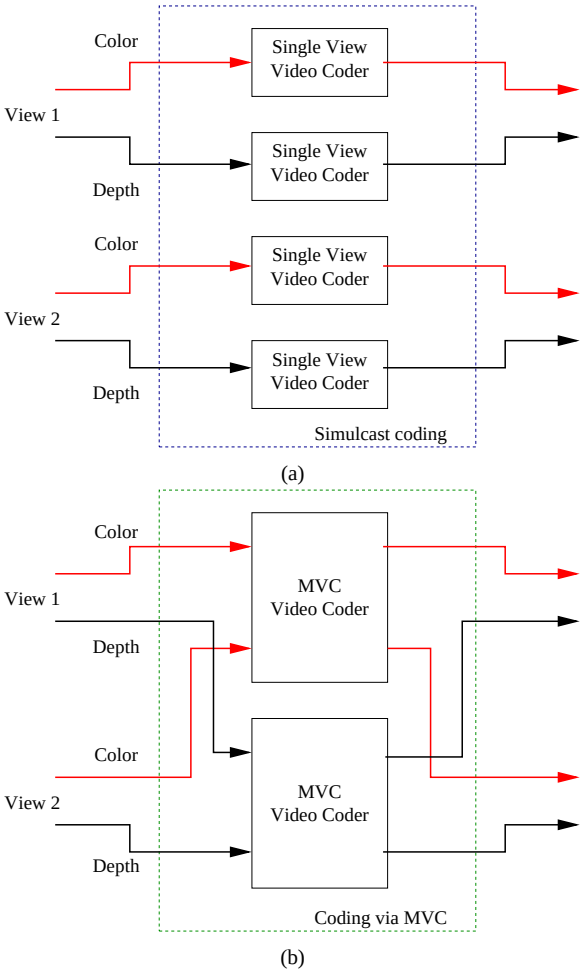

Fig. 2. Reference MVD coding methods: (a) Simulcast; (b) MVC of texture and depths.

date disparity fields. However this criterion is hardly if ever useful, since any disparity field linking equally luminous pixels would make it equal to zero. In order to find a significant solution, we have to inject into the criterion other constraints, accounting for known characteristics of the solution (*a priori* information). This is the regularization needed to solve the problem.

However, before introducing regularization, we want to simplify the criterion. If we assume that an initial coarse estimate $\bar{d}$ of d is available (*e.g.* thanks to block-matching method), and that the difference between $\bar{d}$ and d is small, the warped image can be approximated as:

$$I_t^{n-1}(x+d,y) \simeq I_t^{n-1}(x+\bar{d},y) + (d-\bar{d})\frac{\partial}{\partial x} I_t^{n-1}(x+\bar{d},y) \quad (2)$$

This linearization allows to rewrite the criterion $J[d(\cdot]$ as a quadratic convex functional:

$$J[d(\cdot)] = \sum_{\mathbf{p}\in\mathcal{P}} [r(\mathbf{p}) - L(\mathbf{p})\, d(\mathbf{p})]^2 \quad (3)$$

where

$$\begin{aligned} L(\mathbf{p}) &= \tfrac{\partial}{\partial x} I_t^{n-1}(x+\bar{d}(\mathbf{p}),y) \\ r(\mathbf{p}) &= I_t^n(\mathbf{p}) - I_t^{n-1}(x+\bar{d}(\mathbf{p}),y) + \bar{d}(\mathbf{p})\, L(\mathbf{p}) \end{aligned}$$

As pointed out before, the minimization of J is an ill-posed problem, demanding for additional constraints, which reflect the *a priori* knowledge about the disparity.

This problem can be solved in the context of the set theory. We introduce M constraints. The m-th of them is represented by a closed convex set S_m in a Hilbert space $\mathcal{H}$. We call S the intersection of all the M sets S_m. Then, S is the set of candidate solutions [9], *i.e.* the set where we have to look for the field minimizing J:

$$d^*(\cdot) = \underset{d\in\bigcap_{m=1}^M S_m}{\operatorname{argmin}} \; J(d) \quad (4)$$

This formulation is useful, since the constraints can be described as level sets of suitable continuous convex real functions $\{f_m\}_{m\in\{1,\dots,M\}}$:

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

where $(\delta_m)_{1\le m\le M}$ are real-valued parameters such that $S = \bigcap_{m=1}^M S_m \neq \emptyset$.

Now we shall define the constraints. We consider two simple but effective constraints. The first one specifies the range of values of the disparity field $[d_{\min}, d_{\max}]$, and can be expressed by the constraint set S_1:

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

The second imposes the regularity of the disparity field, limiting the amount of variability of d. This can be achieved by limiting the total variation of the disparity field. The total variation $\mathrm{tv}(d)$ is defined as the sum over $\mathcal{P}$ of the norm of the (discrete) spatial gradient of d [10]. As a conclusion, the total-variation based regularization constraint amounts to impose an upper bound τ on tv:

$$S_2 = \{d\in\mathcal{H} \mid \mathrm{tv}(d) \le \tau\} \quad (7)$$

The total variation limit τ depends on the characteristics of the scene and of the camera configuration: therefore finding an optimal value for it can be an hard task [11]. One of the contribution of this work is to explore the relationship between this parameter and the quantization parameters of the compressed. MVD sequence.

We introduce a last regularization term, which penalizes solutions too much different from the initial one. This is accounted for by a weight α. In conclusion, the criterion to minimize becomes:

$$J(d) = \sum_{\mathbf{p}\in\mathcal{P}} [r(\mathbf{p}) - L(\mathbf{p})\, d(\mathbf{p})]^2 + \alpha||d-\bar{d}||^2 \quad (8)$$

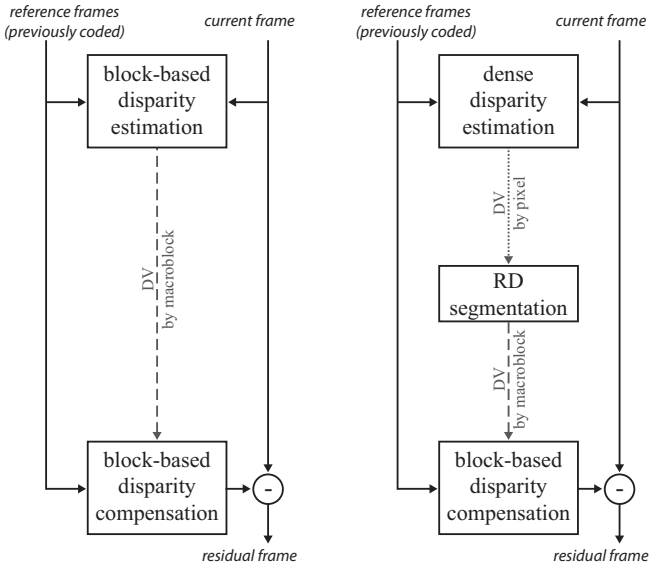

Fig. 3. Disparity prediction: (left) block-based estimation, (right) enhanced by a dense estimation.

In our implementation, described in [8], we used the efficient constrained quadratic minimization technique developed in [9, 12] which is adapted to problems with quadratic convex objective functions.

3. PROPOSED METHOD

The proposed coding scheme for MVD is based on our previous work [8], on which we build in order to take into account the depth maps.

3.1. Starting scheme: RD segmentation of dense disparity fields

This encoder, described in [8], allows to use a dense disparity field for efficiently encoding a multiple view video (without depth) with a standard encoder. For the sake of simplicity, the description and the figures will refer only to the stereo case (*i.e.*, two views). However the encoding schemes are promptly extended to the case of more than two views.

The reference encoder workflow is the following. First, a dense disparity field (DDF) is computed for the color sequence. This DDF is then segmented into 16×16 blocks, corresponding to the H.264 macroblocks (MBs). Then, for each MB we start from the 256 candidate vectors of the dense field, and we have to chose one, in order to represent the current block as it was an ordinary *INTER* block: we will encode the chosen vector and the corresponding motion-compensated residual. The representative vector is chosen with an RD criterion, from a set made up by the average vector of the 256 candidates, the median (in the sense of the norm) vector, and the 4 closest to the median vector. The difference between

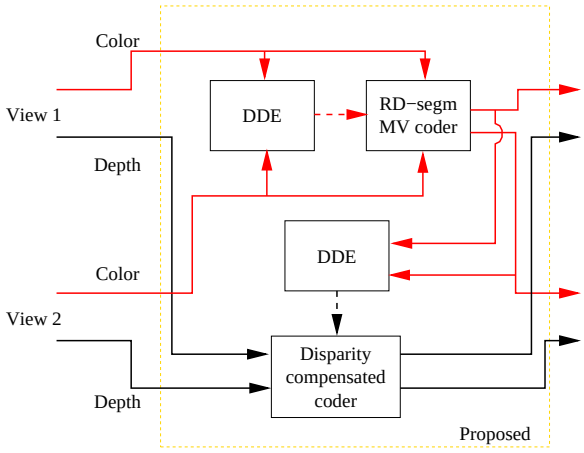

Fig. 4. Proposed MVD coding methods.

classical, block-based disparity compensation, and the proposed one is depicted in Fig. 3.

This process is repeated for all the possible partition of an H.264 macroblock. This means that for smooth disparity regions the RD choice will tend to favor large partitions, while for "active" regions (*i.e.* those where the disparity varies significantly, like across object contours) small partitions will be more likely. Therefore we end up with a RD-driven segmentation of the disparity map, that allows to efficiently encode the stereo pair.

3.2. Proposed scheme for MVD coding

The proposed scheme takes advantage of the dense disparity estimation algorithm described in Section 2 and of the RD-driven segmentation-based multiview (RD-MV) coder described in Section 3.1. This coder is firstly used to encode the color sequences. In this case we set the values for the parameters (α, τ, $d_{\min}$ and $d_{\max}$ using the results of our previous work [8]. Then we use the same RD-MV coder to represent the depth maps. However, in order to save bitrate taking advantage from the correlation between texture and depth, the depths disparity is computed using color images. In this way, we do not need to send the dense disparity map to the decoder, which instead can compute it from the compressed color sequence, and on the other hand, we take benefit from a dense disparity map, that is used to perform a disparity-compensated coding of the depth. Then, the compressed left depth map and the compressed residual of the disparity-compensated right depth map are sent to the output.

One of the key steps of the implementation of this algorithm is the choice of the DDE parameters for depths, in particular the amount of total variation τ. It is intuitive that they depend on the quality of the compressed color sequence. In particular, for high-quality color images, we expect that all small details are represented, and therefore the corresponding disparity field can have a higher total variation. On the con-

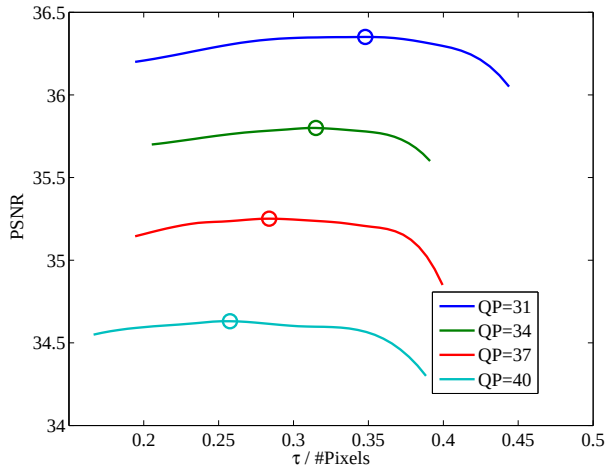

Fig. 5. Effect of τ (normalized over the number of image pixels) on the disparity quality, for several values of the quantization step

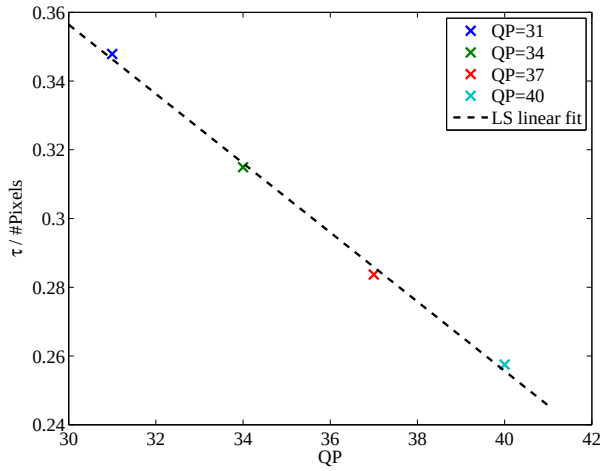

Fig. 6. Best normalized τ values in function of QP. The best fitting first order polynomial is shown as well.

trary, heavily quantized images are much smoother, and we should allow a smaller variation of the disparity field. This intuition is confirmed by the experiments, as shown in the next section.

4. EXPERIMENTAL RESULTS

4.1. Tuning the total variation parameters

In a first set of experiments, we looked for the optimal value of the total variation parameter τ. In particular, the intuition suggests that the best value for τ should depend on the quantization step of the color images used for the estimation. Therefore we ran the following experiment. We considered the multiview video sequence *break dancer*and *ballet*, compressed with QPs in the range $\mathcal{Q} = \{31, 34, 37, 40\}$. For each QP value, we ran the DDE algorithm using the compressed images from several couples of views, using several values of τ. Then we used the resulting dense field to compute a disparity-compensated prediction of the depth map, and finally we compute the PSNR of this prediction with respect to the actual depth map. The results are shown in Fig. 5, where we report, for each QP, the PSNR as a function of the normalized value of τ (the normalization is computed with respect to the number of pixels). In this graph, we also point out the best value of normalized τ, *i.e.*, the one maximizing the PSNR. We refer to this values as τ^*. We remark that these values are strongly correlated to the QP. We compute the sample correlation coefficient obtaining:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{QP_i - \overline{QP}}{\sigma_{QP}}\right)\left(\frac{\tau_i^* - \overline{\tau^*}}{\sigma_{\tau^*}}\right) = -0.9987$$

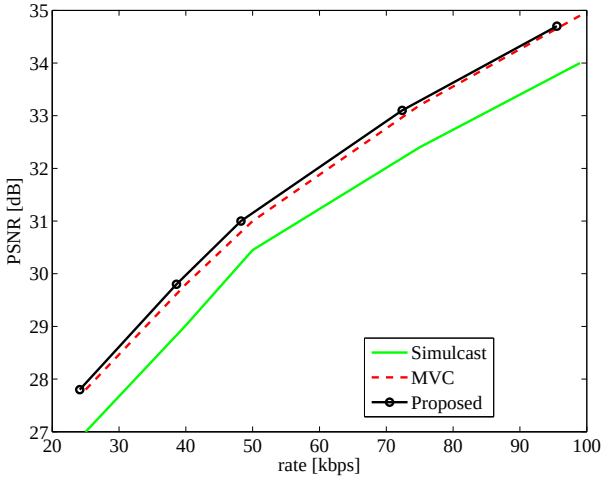

Fig. 7. RD compression performance of the proposed scheme for depth maps. The references are the schemes in Fig. 1

where, as usual, the bar represents the (sample) mean and σ represents the (sample) standard deviation. Finally we computed the least square linear fitting of QP and τ^*. We found the following regression equation:

$$\tau^* = -0.0101QP + 0.6587 \quad (9)$$

In Fig. 6 we show at the same time the experimental points and the least square linear fit. As expected, we obtained a very good match, and so Eq. (9) can be used in the proposed encoder in order to quickly find a good value for τ, at least in the considered range of QP values.

4.2. Compression performance

In a second set of experiments, we evaluated the performance of the disparity-compensated depth map coder, using the op-

timized values of total variation for the disparity field estimation. We encoded the first depth map as an ordinary video sequence, while for the second one, we considered the disparity-compensated residual. The disparity field rate was not take into account since this field is available at the decoder side as well. The resulting RD performances are compare to those of the reference schemes (Simulcast and MVC, see Fig. 1) and shown in Fig. 7. We remark a non-negligible rate reduction (computed using the Bjontegaard metric [13]) with respect to the reference, estimated to 3% less than MVC and 17% less than Simulcast over the test sequences.

Global compression performance were misurated as well. Cumulating the gains obtained by the RD-segmentation driven encoder on the texture and those of the presented encoder for the depth maps, we register an average rate reduction of 11% with respect to an MVC-based scheme as the one shown in Fig. 1(b).

5. CONCLUSIONS AND FUTURE WORK

Multiview video plus depth is a format for 3D and free view point television which is gathering more and more attention, due to its flexibility in representing arbitrary views of a given scene. However, the MVD format is extremely demanding in terms of storage space and transmission bandwidth, and so compression is mandatory. In this paper we propose an algorithm for MVD compression, based on dense disparity estimation, and on the exploitation of the redundancy between color and depth information. The experimental results reveal us how to tune the dense disparity estimation algorithm in order to extract the disparity from compressed color images. Therefore we use disparity to perform efficient coding of depth maps. Compression results are encouraging as well, showing a 3% average rate reduction for the depth maps and a global 11% rate reduction with respect to popular reference methods.

Future works are intended to better exploit the depth-color correlation, by further exploiting the color information into the depth map encoder.

6. REFERENCES

[1] M. Tanimoto, M.P. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint TV," *Signal Processing Magazine, IEEE*, vol. 28, no. 1, pp. 67 –76, 2011.

[2] C. Fehn, P. Kauff, M. Op De Beeck, F. Ernst, W. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, and I. Sexton, "An evolutionary and optimised approach on 3d-tv," in *In Proceedings of International Broadcast Conference*, 2002, pp. 357–365.

[3] Philipp Merkle, Aljoscha Smolic, Karsten Müller, and Thomas Wiegand, "Multi-view video plus depth representation and coding," in *IEEE International Conference on Image Processing*, Oct. 2007, vol. 1, pp. 201–204.

[4] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski, "High-quality video view interpolation using a layered representation," in *ACM SIGGRAPH 2004 Papers*, New York, NY, USA, 2004, SIGGRAPH '04, pp. 600–608, ACM.

[5] "Draft call for proposals on 3D video coding technology," Tech. Rep., ISO/IEC JTC1/SC29/WG11, Daegu, Korea, Jan. 2011, Doc. N11830.

[6] Y. Chen, Y. K. Wang, K. Ugur, M. 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.

[7] Valentina Davidoiu, Thomas Maugey, Béatrice Pesquet-Popescu, and Pascal Frossard, "Rate distortion analysis in a disparity compensated scheme," in *IEEE International Conference on Audio, Speech and Signal Processing*, Prague, Czech Republic, May 2011, ICASSP '11, To appear.

[8] I. Daribo, M. Kaaniche, W. Miled, M. Cagnazzo, and B. Pesquet-Popescu, "Dense disparity estimation in multiview video coding," in *IEEE Workshop on Multimedia Signal Processing*, Rio de Janeiro, Brazil, 2009.

[9] W. Miled and J.C. Pesquet, "Disparity map estimation using a total variation bound," in *3rd Canadian Conference on Computer and Robot Vision*, 2006, pp. 48–48.

[10] Leonid I. Rudin, Stanley Osher, and Emad Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D*, vol. 60, no. 1-4, pp. 259–268, 1992.

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

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

[13] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," in *VCEG Meeting*, Austin, USA, Apr. 2001.