BIT-RATE ALLOCATION BETWEEN TEXTURE AND DEPTH: INFLUENCE OF DATA SEQUENCE CHARACTERISTICS

Emilie Bosc, Riou Paul, Muriel Pressigout, Luce Morin

BIT-RATE ALLOCATION BETWEEN TEXTURE AND DEPTH: INFLUENCE OF DATA SEQUENCE CHARACTERISTICS

Emilie Bosc, Paul Riou, Muriel Pressigout, Luce Morin

Université Européenne de Bretagne, France,
INSA de Rennes, IETR, UMR 6164, F-35708, Rennes

ABSTRACT

This paper questions the existence of factors influencing the quality of the synthesized views, in the context of multi-view video plus depth coding (MVC). The issue of bit-rate allocation between texture and depth data in MVC is still open, despite the many efforts already raised for the development of optimization techniques. The originality of this study lies in the investigation of direct relationships between the best bit-rate allocation, in terms of objective quality of synthesized views, and the sequence characteristics (entropy of depth maps, depth complexity and camera baseline distance, background/foreground contrast areas). The results confirm our assumptions regarding the impact of the sequence features on the bit-rate allocation. The results and the limitations of the study are also discussed.

Index Terms — 3DTV, MVD, 3D video, MVC, quality assessment

1. INTRODUCTION

Multi-view-plus-depth video sequences compression is a huge challenge for researchers. Not only the amount of data to process is significant, but quantization degradations can lead to downsides in terms of quality whose range has never been experienced before with 2D media. Indeed, synthesized views, e.g. novel viewpoints, can be rendered from texture and depth information, through Depth-Image-Based-Rendering methods [1]. However, previous studies ([2]) showed that the impact of coding artifacts on depth data can dramatically influence the visual quality of synthesized views. Yet, both main applications of 3D Video - 3D television (3DTV) that provides a depth feeling, and Free Viewpoint Video (FVV), that allows navigation inside the scene - require the synthesis of novel viewpoints.

Most of the state-of-the-art used codecs for depth maps are inspired from 2D video codecs that are optimized for human visual perception of color images. Thus, they can induce non-perceptible artifacts in the depth map, but unexpected distortions in the synthesized views. While researchers rated their compression methods' performances through the evaluation of reconstructed texture and depth information separately, for a long time, they now often focus on the quality of the view synthesized from the decompressed texture and depth map [3, 4]. Indeed, the success of 3D Video widely depends on the ability to provide high quality synthesized views. Yet, there is no standardized quality assessment framework dedicated to synthesized views, up to now.

Numerous studies addressing the problem of MVD compression proposed bit-rate allocation strategies aiming at minimizing the synthesized views' distortions. In [3], the authors proposed an efficient joint texture/depth rate allocation method based on a view synthesis model distortion, for the compression of MVD data. According the the bandwidth constraints, the method delivers the best quantization parameters combination for depth/texture sets that maximizes the rendering quality of a synthesized view in terms of Mean-Squared-Error (MSE). Similarly, in [4], the authors proposed a joint depth/texture bit allocation algorithm for the compression based on the MSE of the synthesized view. MSE and Peak-Signal-to-Noise (PSNR) that is based on MSE, are generally used as quality criterion by bit-rate allocation optimization methods because of their computational simplicity. However, recent studies [5] suggested that this metric was not sufficient for assessing the quality of synthesized views.

Our goal is to answer the following question: regarding the constraints of complexity of coding methods and considering PSNR as a simple indicator of presence of inaccurate warped areas in the synthesized views, can we rely on this measure to determine cues for a priori setting the bit ratio between texture and depth data without the need for on line Rate/Distortion optimization?

The preliminary step for this study is presented in this paper. It first consists in investigating the relationships between texture and depth data in the context of MVD compression. We believe that our investigations are useful for the improvement of coding strategies. A previous study [6] already addressed the question of the optimal ratio between texture and depth data in this context. This ratio was constant for a fixed sequence whatever the total bit-rate. As this optimal bit-rate would vary according to the sequence content, it suggested that a further analysis of the sequences' features may lead to an other source of information. This analysis is the object of this paper.

The rest of this paper is organized in five sections. Section 2 presents the main conclusion of the previous results and the experimental conditions of the extended study. Section 3 presents the investigated features. Section 4 concludes the paper.

2. SEQUENCE DEPENDENT RATIO

This section sets the framework of our study by reminding the main conclusions of the previous work and by presenting the experimental conditions of the extended work. In particular, this section shows that the extended work confirms the main conclusions of the previous work regarding the sequence dependent ratio between texture and depth data.

2.1. Previous work

In [6], we aimed at evaluating the required ratio between depth and texture data when relying on the quality of a reconstructed view in terms of PSNR. H.264/MVC reference software, JMVM 8.0 (Joint Multiview Video Model) was used to encode three views, as a realistic simulation of a 3D-TV use in a first case study. To vary

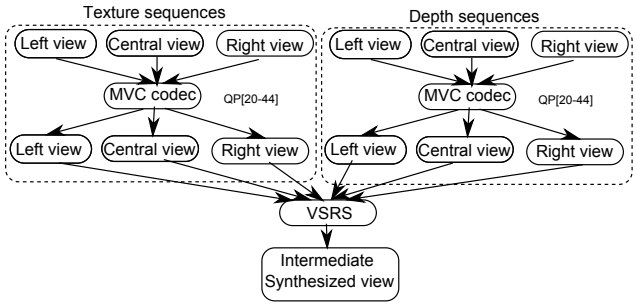

Figure 1. Experimental protocol.

the bit-rate ratio and the total bit-rate, the quantization parameter QP varies from 20 to 44 for both depth and texture coding. The central view predicts the two other views. Then, from the decompressed views, we computed the intermediate view between the central view and the right one, by using the reference software: VSRS [7], version 3.5, provided by MPEG.

Figure 1 illustrates the described protocol. In this figure, "MVC codec" thus refer to H.264/MVC. We used two different types of sequences to answer our question: *Ballet*, and *Book Arrival*, (1024 $\times$ 768). The considered views were 2, 4 and 6 for *Ballet*, and 6, 8 and 10 for *Book Arrival*. For each couple $(QP_{texture}, QP_{depth})$, the average PSNR score of the synthesized sequence was evaluated, compared to the original acquired view. The obtained results showed that the most faithful reconstruction by using VSRS may require to affect between 40% and 60% of the total bit-rate to depth data, depending on the available MVD data (in these experiments, i.e. with *Ballet* and *Book Arrival*). This is far from the recommended ratio in the ATTEST project [1] (below 10-20% of the basic color video bitrate would be allocated to depth data, in the context of videdo-plus-depth data), based on the fact that "*the per-pixel depth information doesn't contain a lot of high frequency components*"[1]. In average, the percentage of bit-rate allocated to depth data leading to the maximum PSNR is 60.5% for *Ballet* and 36.1% for *Book Arrival*. Those observations are related to H.264/MVC encoding. We assume that using a different encoding framework may lead to different figures: if the assumption that being a monochromatic signal, depth data requires less than 20% of the total bit-rate is appropriate ([1], [8] who consider using only 5-10% of the total bit-rate for encoding depth data with H.264/AVC, based on the observation that it "*produced acceptable view synthesis performance*"), then the used codec (H.264/MVC in these tests) is not appropriate for depth maps coding. We also assume that the optimal ratio was dependent on the sequence.
We ran similar experiments with a larger base of MVD sequences to confirm these first results. They are discussed in the next subsection.

2.2. Experimental protocol

In total, 11 sequences were included in the extended tests. For each sequence, "key" frames were encoded ("key" frames were randomly chosen). Then, from the decoded "key" frames, we generated the virtual viewpoints with various baseline distances between the reference viewpoints. Table 1 gives the summary of the used material. Encoding followed the same protocol as described in [6]: considered views (textures and depth maps) were encoded through H.264/MVC reference software (JMVM 8.0). Two test sequences were acquired in toed-in camera configuration (*Ballet* and *Breakdancers*). The other sequences were acquired in parallel configuration, and camera spacing varies between 3.5 cm and 6.5 cm. Test sequences were real natural scenes except for *Mobile* that is a computer generated sequence with ground truth depth data.

Sequence Name	Frame no.	Views
Ballet	1	0-**1**-2
	100	0-**1**-2
Balloons	1	1-**2**-3
		1-**3**-5
		3-**4**-5
	10	1-**3**-5
	50	1-**3**-5
	300	1-**2**-3
		1-**3**-5
		3-**4**-5
Book Arrival	1	8-**9**-10
	99	8-**9**-10
Breakdancers	1	0-**1**-2
		0-**1**-4
		0-**2**-4
		0-**3**-4
		0-**1**-6
		0-**3**-6
		0-**5**-6
		0-**1**-7
		0-**4**-7
		1-**3**-4
		2-**4**-6
		4-**5**-6
		4-**6**-7
	100	0-**1**-2
Cafe	1	2-**3**-4
	300	2-**3**-4

Sequence Name	Frame no.	Views
Champagne	1	37-**39**-41
	300	37-**39**-41
Kendo	1	1-**2**-3
		1-**2**-5
		1-**3**-5
		1-**4**-5
		3-**4**-5
	300	3-**4**-5
Lovebird	1	4-**6**-8
Mobile	1	3-**4**-5
		3-**5**-7
		3-**6**-7
	100	3-**4**-5
		3-**5**-7
		3-**6**-7
	200	3-**4**-5
		3-**5**-7
		3-**6**-7
Newspaper	1	2-**3**-4
		2-**3**-6
		2-**4**-6
		2-**5**-6
		4-**5**-6
	2	2-**4**-6
	10	
	50	
	300	
Pantomime	1	37-**39**-41
	500	37-**39**-41

Table 1. Test material.

Sequence Name	Ratio Depth/Texture in %
Ballet	51.6
Balloons	28.21
Book Arrival	31.97
Breakdancers	46.27
Cafe	38.38
Champagne	52.11
Kendo	27.6
Lovebirds	23.58
Mobile	16.57
Newspaper	30.97
Pantomime	19.48

Table 2. Ratio between texture and depth information allowing the minimal distortion in terms of PSNR.

2.3. Results and discussion

The optimal ratio between depth and texture are calculated through PSNR of the intermediate synthesized view compared to the original acquired view, as an indicator of inaccurate warped areas. Table 2 summarizes these results. This table confirms the assumptions raised by the previous experiments since the ratios vary from 16% to 52%, depending on the sequence: there is a relationship between the content and the required ratio. This is studied in the next section.

3. FEATURES ASSUMED TO INFLUENCE THE SYNTHESIZED VIEW QUALITY

Based on the previous results, we aim at investigating the relationships between texture and depth data. The aspects to study are the features which differ from one content to another. We tested various features but only the most relevant are presented in this paper.

Relative depth maps entropy, baseline distance between the reference cameras, features of discovered areas are discussed in the following.

3.1. Relative depth maps entropy

The ratio that rules the optimal synthesized views in terms of PSNR, is related to the amount of information contained in the original data. In other words, the entropy of depth against the entropy of texture is expected to influence the optimal allocation between depth and texture. Let e_d be the average entropy of the encoded depth maps, for a given content. Let e_t be the average entropy of the encoded texture frames, for the same content. Thus, for a given sequence and a couple {depth, texture} data, we consider the following relative depth map entropy:

$$R_e = \frac{e_d}{e_d + e_t} \tag{1}$$

The consideration of this relative depth map entropy is an alternative idea to that of the fixed ratio of 20% of the total information owned by depth data.

Fig. 2 plots the mean R_e (from Eq. 1) per sequence against the "optimal" percentage of bit-rate allocated to depth data according to our previous experimental protocol. There is a linear relationship between R_e and the "optimal" percentage of bit-rate allocated to depth data. The correlation coefficient between R_e and the "optimal" percentage of bit-rate allocated to depth data reached 76.95%. These results are understandable because a high entropy value for the depth implies a highly detailed depth structure. If the level of details of depth is higher than that of the texture, the synthesis quality mostly relies on the accuracy of the depth map. These results suggest that an preliminary analysis texture and depth entropies can be used as an indicator for automatic bit-rate allocation between these two types of data.

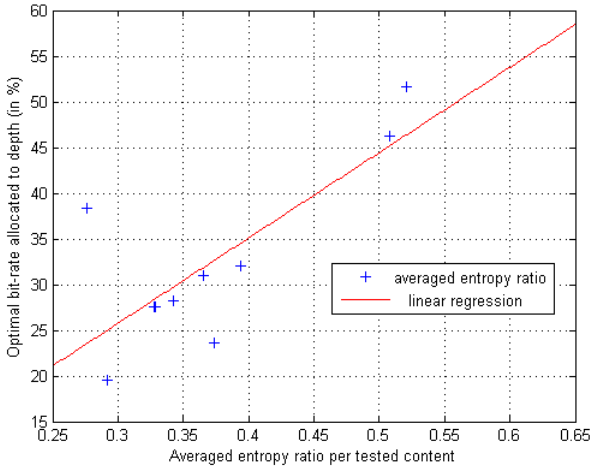

Figure 2. Ratio of entropy between texture and depth data against optimal percentage of bit-rate allocated to depth data according to our previous experimental protocol, in terms of PSNR, per sequence.

3.2. Baseline distance between cameras and discovered areas

We assume that there is a relationship between the structure of the scene depth and the "optimal" percentage of bit-rate allocated to depth data in MVC. According to the depth structure complexity and the baseline distance between the reference cameras, discovered areas in the novel virtual viewpoints are relatively large and difficult to fill-in by through the synthesis process. Since, the discovered areas are filled in with in-painting methods whose texture estimation quality differs according the used strategy, these areas are prone to perceptible synthesis errors. We aim at evaluating the influence of the discovered areas on the "optimal" percentage of bit-rate allocated to depth data. Let T_r and T_l be the original texture right and left view, respectively and D_r and D_l be the original right and left depth maps respectively. Let $T_{r\to v}$ the projection of T_r into the target virtual viewpoint, and $T_{l\to v}$ the projection of T_l into the target virtual viewpoint. $T_{r\to v}$ and $T_{l\to v}$ contain undetermined areas that correspond to the discovered areas. $T_{r\to v}$ and $T_{l\to v}$ are used to create logical masks $M_{r\to v}$ and $M_{l\to v}$ defined as:

$$M_{r\to v}(x,y) = \begin{cases} 0 & \text{, if } T_{r\to v}(x,y) \text{ is determined} \\ 1 & \text{, if } T_{r\to v}(x,y) \text{ is not determined} \end{cases} \tag{2}$$

$$M_{l\to v}(x,y) = \begin{cases} 0 & \text{, if } T_{l\to v}(x,y) \text{ is determined} \\ 1 & \text{, if } T_{l\to v}(x,y) \text{ is not determined} \end{cases} \tag{3}$$

Then we consider the importance of the discovered areas, denoted as I, according to its depth by applying the masks on the respective depth maps as follows:

$$I = \frac{1}{2\times M\times N}\sum_{x=1}^{N}\sum_{y=1}^{M} \begin{array}{l} \big(D_r(x,y)\times M_{r\to v}(x,y) \\ +D_l(x,y)\times M_{l\to v}(x,y)\big) \end{array} \tag{4}$$

where N and M are the width and height of the original image. Score I, from Eq. 4, is computed for each piece of the tested material. The results are plotted in Fig. 3. As expected, this figure shows a relation between the computed importance score I and the "optimal" percentage of bit-rate allocated to depth data. Then, this indicator might be useful for a priori setting the bit ratio between texture and depth data for a given sequences. However, the limit of this cue lies in its dependency to the target synthesized viewpoint and the baseline distance between the reference viewpoints.

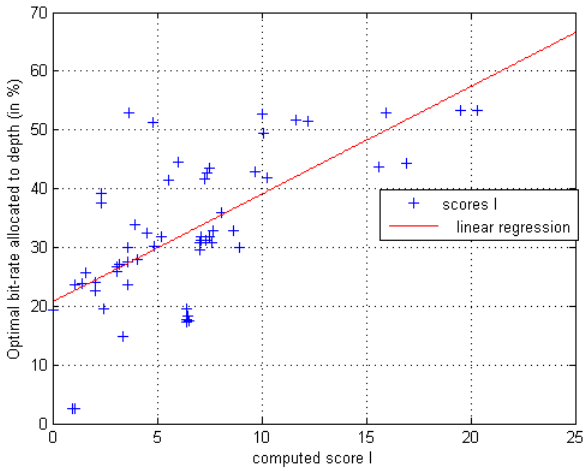

Figure 3. Importance of discovered area against optimal percentage of bit-rate allocated to depth data according to our previous experimental protocol, in terms of PSNR.

3.3. High contrast background/foreground areas

We assume that errors occurring after the synthesis process are not only more noticeable when the contrast between background

objects and foreground objects is high, but also more penalized by signal-based objective metrics. To investigate this assumption, we consider the strong depth discontinuities (highlighted by an edge detection algorithm) and evaluate the standard deviation of the texture image around these discontinuities. Let ΔD be the gradient image of depth map D. Any pixel located at coordinates (x, y) is noted p and we consider the set of pixels Γ such as:

$$\Gamma = \{p = (x, y) | \Delta D(x, y) > 0\} \quad (5)$$

The investigated feature, denoted C, that expresses the contrast between foreground and background areas around the strong depth discontinuities, is computed as:

$$C = \frac{1}{|\Gamma|} \sum_{p=1}^{|\Gamma|} \sigma(T(p)), p \in \Gamma \quad (6)$$

where $|\Gamma|$ is the cardinality of Γ and $\sigma(T(p))$ is the standard deviation of 5×5 window centered on pixel p in texture view T. For each piece of the tested material, we compute C for left and right views and we consider the mean of these two coefficients. Fig. 4 plots the obtained scores. Unexpectedly, the results show that the higher the contrast, the less bit-rate allocated to depth: two main point clouds are distinguishable. The point cloud corresponding to 40-55% of bit-rate allocated to depth belongs to the two toed-in camera configuration sequences. The second cloud corresponds to the parallel camera configuration sequences. So, our assumption is that despite the high contrast around objects contours, the camera configuration (and thus the distance to the virtual view) might reduce the impact of the synthesis distortions.

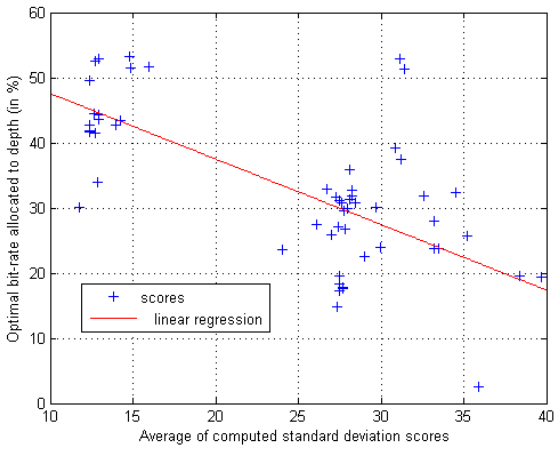

Figure 4. Influence of high contrast background/foreground areas: Average of computed standard deviation scores around gradient pixels of depth maps against optimal percentage of bit-rate allocated to depth data according to our previous experimental protocol, in terms of PSNR.

4. CONCLUSION

The study presented in this paper aimed at highlighting cues for a priori bit-rate allocation strategies through the analysis of MVD sequences features.The experiments consisted in encoding both texture and depth data by the same compression scheme, varying the ratio between texture and depth information and analyzing the quality of the rendered virtual view. The experiments relied on the use of H.264/MVC. The attributed depth ratio varied from 2% to nearly 95% and the synthesis of an intermediate view was performed. The analysis of the MVD data features and related parameters such as video contents and camera settings revealed the existence of their impact on the best trade-off for bit-rate allocation between texture and depth data. The assumed factors influencing the best trade-off for bit-rate allocation between texture and depth data were the depth map entropy, its complexity coupled with the camera baseline distance, and the contrast of neighboring background/foreground pixel areas.

Despite its limitations concerning the choice of the distortion indicator (PSNR) that is not perceptually oriented but that addresses the computational contrainsts in video coding, and the assessment targeting monoscopic viewing, this study highlighted the existence of cues for the conception of a priori bit-rate allocation strategies. The new bit-rate allocation strategies might consider a weighted combination of indicators presented in our study, depending of the used codec. A similar work relying on a different quality criterion could be an interesting source of information for comparative study.

The next step of our research implies the integration of a weighted combination of indicators presented in our study for the conception of a bit-rate allocation method and the evaluation of the results by objective and subjective quality assessment methods.

5. ACKNOWLEDGMENTS

This work is supported by the French National Research Agency as part of PERSEE project (ANR-09-BLAN-0170). We would like to acknowledge Microsoft Research, Fraunoffer HHI, Nagoya University, GIST, Nokia, ETRI/MPEG Korea Forum and MPEG for providing the tested sequences and VSRS.

6. REFERENCES

[1] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," in *Proceedings of SPIE Stereoscopic Displays and Virtual Reality Systems XI*, 2004, vol. 5291, p. 93104.

[2] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proceedings of ICIP*, 2007, p. 201204.

[3] Y. Liu, Q. Huang, S. Ma, D. Zhao, and W. Gao, "Joint video/depth rate allocation for 3D video coding based on view synthesis distortion model," *Signal Processing: Image Communication*, vol. 24, no. 8, pp. 666–681, Sept. 2009.

[4] Y. Morvan, D. Farin, and P.H.N. de With, "Joint depth/texture bit-allocation for multi-view video compression," in *Proceedings of Picture Coding Symposium (PCS 2007)*, Lisboa, Portugal, Nov. 2007, vol. 10, p. 4349.

[5] E. Bosc, R. Pepion, P. Le Callet, M. Koppel, P. Ndjiki-Nya, M. Pressigout, and L. Morin, "Towards a new quality metric for 3-D synthesized view assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 7, pp. 1332–1343, Nov. 2011.

[6] E. Bosc, V. Jantet, M. Pressigout, L. Morin, and C. Guillemot, "Bit-rate allocation for multi-view video plus depth," in *Proc. of 3DTV Conference 2011*, Turkey, 2011.

[7] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference softwares for depth estimation and view synthesis," Apr. 2008.

[8] E. Martinian, A. Behrens, J. Xin, A. Vetro, and H. Sun, "Extensions of h. 264/AVC for multiview video compression," in *Image Processing, 2006 IEEE International Conference on*, 2006, p. 2981–2984.