FLEXIBLE MOTION MODEL WITH VARIABLE SIZE BLOCKS FOR DEPTH FRAMES CODING IN COLOUR-DEPTH BASED 3D VIDEO CODING

B. Kamolrat, W.A.C. Fernando, M. Mrak, and A. Kondoz

Centre for Communication System Research
University of Surrey, Guildford, United Kingdom
{B.Kamolrat, W.Fernando, M.Mrak , A.Kondoz}@surrey.ac.uk

ABSTRACT

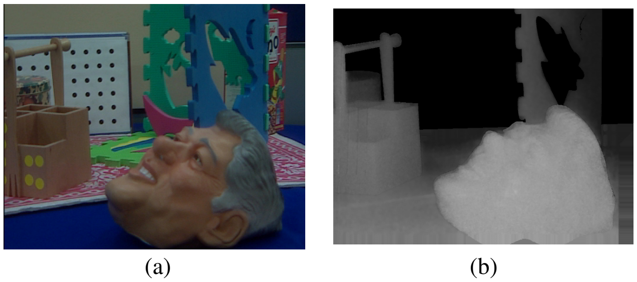
New techniques for representing 3D video are often realised using monoscopic video and associated per-pixel depth information. While the compression of both monoscopic and depth video channels can be achieved using traditional video coding techniques, in this paper, specific properties of the depth channel are exploited to further compress the depth information. More precisely, enhanced compression of the depth channel is achieved by using a highly flexible motion model, Binary Partition Tree (BPT) which enables adaptive partitioning of the depth frames leading to better rate-distortion optimisation for inter-frame prediction. Comparing to a conventional approach, Limited Variable Size Block (LVSB) which enables only limited frame partitioning options, the presented approach brings a gain of up to 70 % in bitrate reduction for the depth information.

Index Terms – *3D video, motion compensation, variable size blocks*

1. INTRODUCTION

The research on stereoscopic video has received high interest over the past decade in order to provide viewers with more realistic vision than traditional 2D video. Stereoscopic images and display systems are designed to emulate human stereo perception by capturing a 3D scene with two cameras located in slightly shifted positions. The captured pictures are then projected to left and right eyes, called left and right views. Subsequently, the human visual system will perceive left and right views as one image with the sense of depth.

Instead of using left and right views to represent 3D video, new techniques, like depth image-based rendering (DIBR) [1] and screen parallax [2], represent 3D video based on a monoscopic video and associated per-pixel depth information (simply called depth map). The example pictures of monoscopic video and the associated depth map of popular 3D test video sequence are shown in Fig. 1. While the monoscopic video consists of three components - Y, U and V as in the traditional video applications, the depth map video has one component. The advantage of such a scheme is, it can capture the stereoscopic sequences more easily compared to the traditional left-right view technique. However, like in the conventional stereoscopic coding, problems of compression are still there with this DIBR technique. Several basic techniques have been used to encode DIBR sequences, but nothing has been able to reduce the data rate significantly. In this paper, we propose an application of a flexible motion model with variable block size to encode the depth sequence in DIBR sequences.



(a) (b)

Fig. 1. An example of 3D video - frame from the "Orbi" test sequence; (a) monoscopic and (b) depth map frame

The rest of the paper is organized as follows Section 2 presents related works in this area. The proposed technique which enables the reduction of stereoscopic video bitrate is discussed in section 3. Section 4 provides results of the experiments and discussion. Section 5 concludes the paper.

2. RELATED WORK

When 3D video is represented by DIBR technique, it is found that each region of depth map has different importance. Precisely, sharp discontinuities in depth and intensity require more accurate depth map. Therefore, the concept of region-of-interest (ROI) integrated with JPEG2000 was applied in [3] to gain quality improvement for image coding. Moreover, due to the relationship between the monoscopic video and depth map, the idea of sharing motion vectors between two sequences was introduced in [4]. The performance of sharing motion vector approach was

978-1-4244-2571-6/08/$25.00 ©2008 IEEE

found to be superior to encoding motion vectors separately at low bitrate when the coding cost for motion vectors outweighs the one for coding the residual. However, these techniques have not shown considerable bitrate reductions. Therefore, further development of techniques to improve 3D video coding is needed.

In video coding, the motion compensation is used to remove temporal redundancy between successive frames of a video signal. Motion compensation is performed in the direction of motion vectors (MV) which are obtained during the motion estimation for each frame region. Popular frame region shapes are blocks which are commonly used in standard video coding techniques since handling of such shapes is easy to implement and provides reasonable results across a wide range of compression ranges.

The block models can be categorized into fixed size block (FSB) models and variable size block (VSB) models. For FSB, a frame is divided into evenly distributed fixed size blocks. However, in general, block boundaries do not coincide with motion boundaries of objects in video, which means that a block may contain regions corresponding to more than one type of motions. Therefore, it is desirable to not only vary the block size depending on motion in specific regions but also to enable adaptable frame partitioning into motion blocks.

3. APPLICATION OF FLEXIBLE MOTION MODELS TO 3D VIDEO

The VSB approach used in the model is based on the block structure of H.264 / AVC and the BPT [6]. The block-size of the model used in H.264 can be varied from 16x16 down to 4x4 pixels. For BPT, the decision on a frame partitioning is related to the accuracy of predicted motion compensation at certain motion estimation levels. The block size is not predefined but is fully adaptable to the designed rate-distortion. This can be achieved by two steps: growing the tree which describes a frame partitioning (top-down approach) and pruning the tree which finds the optimal partitioning (bottom-up approach). At the tree growing step, the entire picture is repeatedly split up to a target number of blocks N which is defined as. $N = \alpha \cdot N_R \cdot N_C / 256$, where N_R and N_C are the numbers of rows and columns respectively, and α has a value of 2 for P-frames and 1 for B-frames. To split the blocks, all blocks are considered and the block which provides largest minimum motion compensation error is split by a vertical or horizontal line at the position that minimizes the block motion compensation error. The vertical line is used if the block width is bigger than block height and horizontal line is used if the block height is bigger than block width. These two blocks are added to the tree as children nodes of the partitioned block. The example of tree growing algorithm is illustrated in Fig. 2. to generate 4 blocks of a frame.

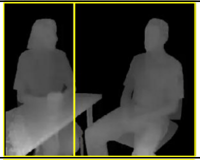
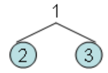
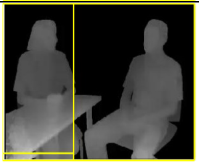
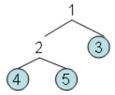
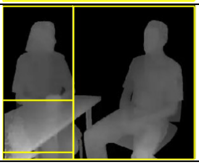
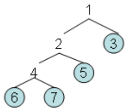
Tree-growing iterations	Original frame and its partitioning	BPT and the energy of motion compensated frame
0		 E = 53.39
1		 E = 51.77
2		 E = 42.37
3		 E = 38.02

Fig. 2. Growing the tree to generate 4 blocks at the 4th frame of "Interview" test sequence.

At the bottom up step, a pair of child blocks, which provides the smallest gain in motion compensation, is re-merged. The process is repeated until number of block n within the frame minimises a Lagrangian cost function defined as $J(n) = D(n) + \lambda \cdot R(n)$, where D, R and λ denote the distortion, bitrate and Lagrangian parameter, respectively.

To perform the VSB approach, there are three main types of data have to be coded: the information required to represent the block structure for each inter frame, the motion information for each block and the residual after motion compensation.

Since depth map consists of large smooth areas as well as sharp edges, our evaluation focuses on application of suitable motion compensation methods which can exploit the specific structure of depth images. In order to gain better temporal redundancy reduction in the depth sequence, we use the above algorithm to encode depth images. Flexible motion models are applied to 3D video sequences: monoscopic video and depth map, in order to reduce the overall bitrate. The flexibility of the applied model is capable to capture the most significant characteristics of depth map - large smooth areas and sharp edges. Since the depth map indicates the depth of objects within the scene, motion of objects related to the background only depends on the motion of objects themselves or camera, not on the changing of environment such as light or colour. Therefore, it is desirable to have a motion model functionality which will allow for assignment of as large as possible block to background in order to reduce the amount of motion information and small block on object edges to obtain the accurate motion prediction.

4. SIMULATION RESULTS AND DISCUSSION

The experiments are carried on the "Orbi" and "Interview" sequences of CIF (352 x 288) resolution. Two tested motion model settings use the VSB structures that follow the H.264 block model for which the block size is varied between 16x16 and 4x4 pixels and the BPT approach which the block size is varied according to the motion of video content. These two approaches are simply called LVSB and BPT for the rest of paper. Video sequences are encoded as IBBPBBP… and a group of picture (GOP) equals to 15. At B-frames, the past and future reference frames are used (bi-directional prediction) while only one reference frame is used for P-frames. I frames are in both cases encoded using the same settings.

Lagrangian rate-distortion optimization is employed to ensure that each component of the encoder operates using the same rate-distortion trade-off relationship. The Lagrangian parameter λ is used to maintain the picture quality to be almost the same within the GOP.

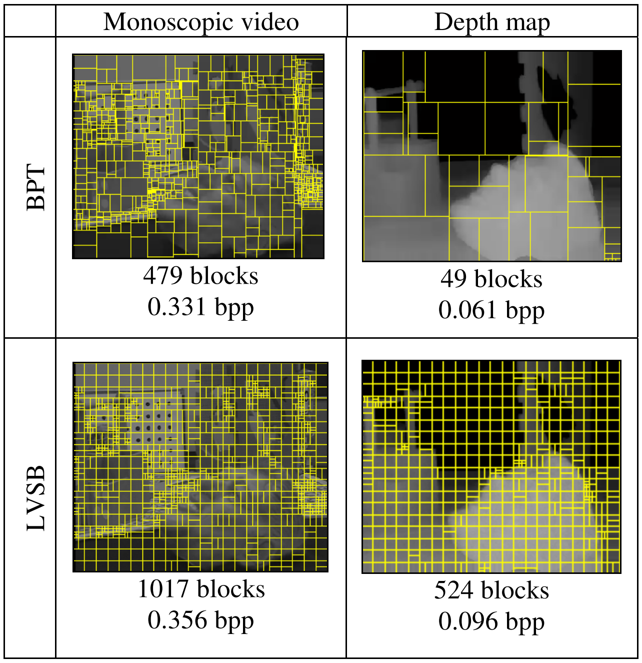


Fig. 3. The 3rd frame of "Orbi" -Monoscopic image and Depth map- segmented by BPT and LVSB

		"Orbi"		"Interview"	
		Mono-scopic	Depth map	Mono-scopic	Depth map
BPT	Average PSNR [dB]	38.35	42.25	36.00	41.64
	Average number of blocks/frame	230	60	157	130
	Average rate [bpp]	0.125	0.036	0.073	0.052
LVSB	Average PSNR [dB]	38.35	42.25	36.00	41.64
	Average number of blocks/frame	811	510	743	552
	Average rate [bpp]	0.141	0.063	0.082	0.064

Table 1. The average number of blocks per frame and proportion of the bitrate used for compressed P and B frames.

The examples of monoscopic video and depth map of Orbi (P-frame) which are divided into variable size blocks by the BPT and LVSB approaches are presented in Fig. 3. In case of the monoscopic image, to obtain the same PSNR at about 36 dB the image is segmented into 479 blocks by BPT and 1017 blocks by LVSB. In addition, for the depth image, it is segmented into 49 blocks by BPT and 524 blocks by LVSB in order to obtain image quality of about 39 dB. Moreover, the bitrate used in BPT in the monoscopic video is about 7.5 % lower than in LVSB and about 57 % lower in depth map. It means that the BPT can significantly reduce bitrate (compared to LVSB) in the depth map by using fewer blocks to provide good quality of motion compensation. As a result, for the same bitrate the application of BPT will provide higher decoding quality compared to the compression which uses LVSB.

The average numbers of blocks used in all P and B frames of the test sequences are presented in Table 1. for almost the same video quality. Since a typical depth map frame is smooth and its background occupies large portion of frame, assignment of larger blocks to background area reduces the amount of motion information. On the other hand, assignment of smaller blocks on object edges is needed to obtain the accurate motion prediction for those areas that change faster. As a result, a smaller number of blocks can be used in BPT to provide better rate-distortion optimisation. Especially, at the B-frame where bi-directional prediction is applied, even smaller number of blocks is used. Consequently, fewer bits are required to encode the motion component. On the other hand, a block size of LVSB, which has the same block structure used in H.264 / AVC, is varied between 4x4 to 16x16 pixels.

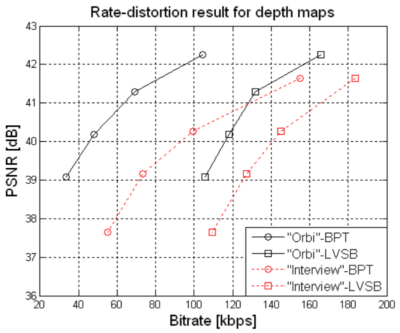


Fig. 4. Rate-Distortion results on Orbi and Interview (depth map)

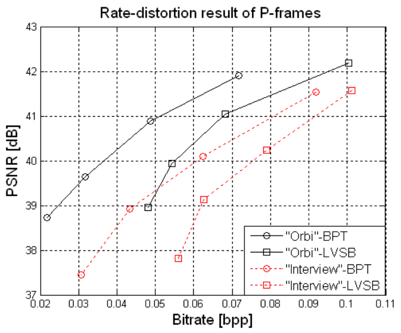


Fig. 5. Rate-Distortion results on Orbi and Interview (depth map) at P-frames

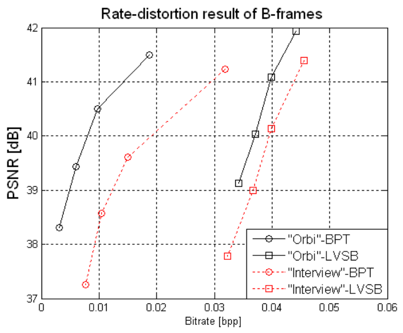


Fig. 6. Rate-Distortion results on Orbi and Interview (depth map) at B-frames

Therefore, the biggest block is limited to 16x16 pixels, leading to high number of bits required for the motion component. This reflects to the overall bitrate which is larger when LVSB is used. Most importantly, it can be seen that the bit savings achieved by application of BPT are much higher for the depth map compared to the monoscopic video.

The rate-distortion results for test sequences (depth map) are presented in Fig. 4. for wide range of bit-rates. The presented results suggest that the performance of BPT is much superior to the performance of LVSB. This is quite obvious at the low bitrate region since the rate-distortion optimization algorithm used in BPT is capable to find the optimum trade-off between bitrate for motion information and residual to provide the possible best video quality. Moreover, the rate distortion results in P and B frames are shown in Fig. 5. and Fig. 6. The experiment results suggest that the performance of BPT is better than LVSB in both frame types, especially in case of the B-frame where bi-directional prediction is implemented. The bi-directional prediction allows for more accurate motion estimation. As a result, even fewer blocks can be used leading to reduction in motion information.

5. CONCLUSIONS

In this paper, a flexible motion model is proposed to code the depth map in DIBR. We have used the features of the depth map and based on that we used BPT to design a flexible motion model to encode the frame. Simulation results clearly suggest that the flexible motion model can improve the RD performance quite significantly compared to the LVSB which is used in standard video codecs. The proposed technique will add some computational complexity to the system to optimise the partitioning of the frames. In our future work we will address this issue.

ACKNOWLEDGEMENT

This work was partly supported by the European Commission IST FP6 program, under VISNET II project, a European Network of Excellence.

5. REFERENCES

[1] W. A. IJsselsteijn, P. J. H. Seutiens, and L. M. J. Meesters, "State-of-the-Art in Human Factors and Quality Isses of Stereoscopic Broadcast Television," Technical Report D1, IST-2001-34396 ATTEST 2002.

[2] A. Bourge and C. Fehn, "ISO/IEC CD 23002-3 Auxiliary Video Data Representation," *ISO/IEC JTC 1/SC 29/WG 11/N8038,* April 2006.

[3] R. Krishnamurthy, B.-B. Chai, H. Tao, and S. Sethuraman, "Compression and Transmission of Depth Maps for Image-Based Rendering," *Int. Conf. Image Processing,* pp. 828-831, 2001.

[4] S. Grewatsch and E. Miiler, "Sharing of Motion Vectors in 3D Video Coding," *Int. Conf. Image Processing,* pp. 3271-3274, 2004.

[5] "Draft ITU-T recommendation and final draf international standard of joint video specification (ITU-T Rec. H.264/ISO/IEC 14 496-10 AVC," *in Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVTG050,* 2003.

[6] M. P. Servais, T. Vlachos, and T. Davies, "Motion Compensation using Variable-Size-Block-Matching with Binary Partition Trees," *Int. Conf. Image Processing,* pp. 157-60, 2005.