Joint Intra Coding of Video and Depth Maps

Krzysztof Klimaszewski, Krzysztof Wegner, Marek Domański
Chair of Multimedia Telecommunications and Microelectronics,
Poznań University of Technology,
Polanka 3, Poznań, Poland,
e-mail: {kklima, domanski}@et.put.poznan.pl, kwegner@multimedia.edu.pl

***Abstract*— The paper describes a technique for compression of a stereoscopic video pair together with the corresponding two depth maps. The proposed encoder produces a joint bitstream for two viewpoint video sequences and two depth maps. Thus the headers with the control information is shared for all 4 video items. Moreover mutual correlation between these items is exploited by use of the multiview compression tools from MPEG-4 AVC/H.264 coding technology. The performance of the proposed technique is evaluated and compared to the performance of independent coding of video and depth maps.**

I. INTRODUCTION

The recent developments of three dimensional (3D) video technology allow to envisage a great popularity of the 3D-based services. These services include stereoscopic television systems with various types of displays and free viewpoint television. In most of the cases, information about 3D structure of a scene is required, provided, for example, in a form of a depth map.

To provide the new 3D services, all necessary data have to be transmitted to the users, including the view video sequences and the corresponding depth data. There is, therefore, a need to compress these data in an efficient way.

The aim of this paper is to present a new technique for efficient encoding of video and accompanying depth maps. A scenario with two views and depth maps is considered, although an extension to any multiview scenario is mostly straightforward.

II. VIDEO AND DEPTH MAP CODING

Coding of multiview video can be conducted in several ways, like independent coding of two separate views (simulcast) and coding using specialized multiview codec, such as H.264/AVC codec [1] with its Annex H - multiview coding extension (MVC). Gains of approximately 20% of bitrate can be attained when using MVC compared to simulcast. In addition to video data, receivers need the depth data in order to be able to synthesize virtual views. This depth data is represented by a disparity map, a grayscale image of the same resolution as the luminance image of the video data. The most obvious approach would be to encode required disparity maps in the same way as the video data. This scenario is depicted on the left of Figure 1. An example use of the method is presented in [2] and [3]. This method, however, neglects all the special properties of the depth data. There are other methods, such as platelet coding [4][5], that are more suited to depth compression. However, they use a completely different approach to coding and therefore are more difficult to merge into existing video coding standards. Special prediction techniques that use knowledge of depth data, like [6] are proposed. Their drawback is that a separate bitstream is produced. There are also papers describing joint coding of depth maps and video, as depicted on the right of Figure 1. This approach allows to reduce an overhead of control data and inherently ensures synchronization of video and depth data. For example, in [7] authors propose to use scalability mechanism of H.264/AVC to encode video and depth maps. They focus their attention on the similarity of motion field in video and depth maps.

In our approach, we propose to use MVC with its multiview compression capabilities and extend them to the joint video plus depth case. In the first stage we propose joint intra coding technique.

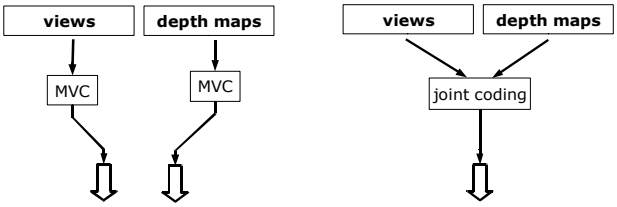

Fig. 1. Two possible coding scenarios for video and depth. Independent coding with two output streams on the left and proposed joint coding with one output stream on the right.

III. DEPTH MAP FEATURES

There are several differences between video and depth data, the most important ones can be summarized in a table shown below.

TABLE I
COMPARISON BETWEEN VIDEO AND DEPTH DATA PROPERTIES

Depth Maps	Views
only one color component	three color components
limited texture	abundance of texture
prominent contours	less prominent contours
accurate depth value important	accurate value not necessary

From the coding point of view, the most important difference is the lack of textures in depth map. This property results in different statistics of macroblock data.

The other difference is the existence of prominent contours of scene objects, that need to be preserved through compression and decompression process.

The importance of exact pixel value is a concern mainly for large quantization parameter indices.

Apart from differences, several similarities between view and depth data exist. The object boundaries are exactly in the same places, both for disparity map and video frame. This correlation can be used to increase coding efficiency of joint coding of video and depth data. In this work we focus on intra coding only. On the Figure 2, intra prediction modes selected by the encoder are presented (different shades of grey correspond to different prediction directions of intra coding) for a selected area of a single frame of coded sequence. For reference, there are also luminance and disparity images of the same region of the frame.

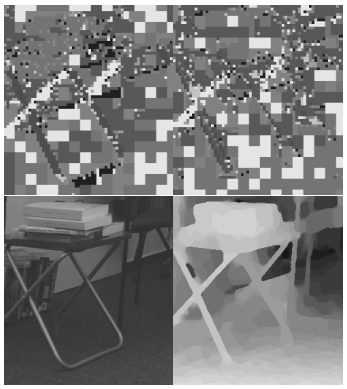

Fig. 2. Prediction modes (top) and input images (bottom) for view and depth coding. View on the left, disparity map on the right.

In both pictures representing prediction direction the chair is easily discernable, with its parts being coded using the same prediction directions. Block sizes are highly correlated between depth and view coding - on the borders of objects there are 4x4 blocks, and in less complex areas 16x16 blocks are the more frequent choice. Also intra prediction modes are similar for video and depth, especially along object boundaries.

IV. Proposed coding technique

To exploit the correlation between depth and video data, a joint compression of depth and video data has to be developed. It is possible to treat the depth data as a fourth component, that can be added to a well known YUV representation of images. In the proposed method, the depth data is coded as an additional component with the same resolution as the luminance component. Joint intra coding for this representation is presented.

In the approach presented in this paper it is assumed, following the works described in [3], that video quality is more important for high quality view synthesis than quality of depth map. It is therefore assumed that, in proposed coder, reconstructed video has to be identical to the video reconstructed in independent video coding.

From the above assumption it follows that the proposed encoder has to select the same modes as for independent coding of luminance for a given macroblock. There is an additional, separate procedure of choosing best macroblock and prediction modes for depth component coding. It is possible to choose another macroblock mode and other prediction mode, exclusively for depth, depending on the result of the rate - distortion optimization.

For every macroblock, depth map data in the stream is preceded by a field that carries information about which parameters are sent in the bitstream and which should be inherited from luminance block. The field length is variable, and lengths for all possible cases are described in the Table II.

TABLE II

Bitstream field for coding depth map information

Condition		Code length [bits]
depth MB mode	**depth prediction mode**	
Same as Luminance	Same as Luminance	1
Same as Luminance	Other than Luminance	2
Other than Luminance	Same as Luminance	3
Other than Luminance	Other than Luminance	3

In case when a given mode is the same as for luminance, it is not included in the bitstream. This contributes to the proposed method's gain.

For 16x16 luminance blocks, all the information, like prediction mode and presence of AC transform coefficients, are included in the macroblock mode field. Therefore it is necessary to include additional flag signaling presence of AC coefficients for depth data.

For 4x4 luminance blocks, an additional field is sent in the same purpose, namely Coded Block Pattern (CBP). It is necessary to send such information for luminance and depth. It was found that the best solution is to send CBP fields separately, with slight modification for depth.

The AVC standard defines codewords for coding all possible values of CBP field using VLC codes. Reassignment of available codes and removing redundant information increases coding performance. This is because for depth coding, the most probable situation is the one where none of the blocks in the macroblock contain non-zero AC coefficients.

The overall quality of reconstructed data is directly connected to quantization parameter, that is selected by a quantization parameter index. In the proposed encoder, there are two separate quantization parameter indices, one for luminance (QP) and one for depth (QD). The greater the index value, the smaller the bitstream, at cost of decreased quality of reconstructed video.

The base indices, QP and QD, have to be sent in the bitstream. In the proposed encoder, the base QD index is sent in the same structure (Picture Parameter Set) as the base QP index for standard AVC bitstream is.

The probability of choosing the same parameters for depth and view macroblocks for two different test sequences, Newspaper [8]and Book Arrival [9], is shown on Figure 3. The parameters of the plots are QP and QD indexes, while the result is a number in range 0 to 1, corresponding to the probability of choosing the same parameter for coding luminance and depth components for a given macroblock.

The probability of choosing the same type of macroblock is high (above 0.5 for most QP-QD combinations), while the situation where both, the macroblock mode and the prediction mode are the same, is less frequent, especially for low values of QP and QD indexes. This probability increases to well above 0.5 for QP and QD larger than 30 for both tested

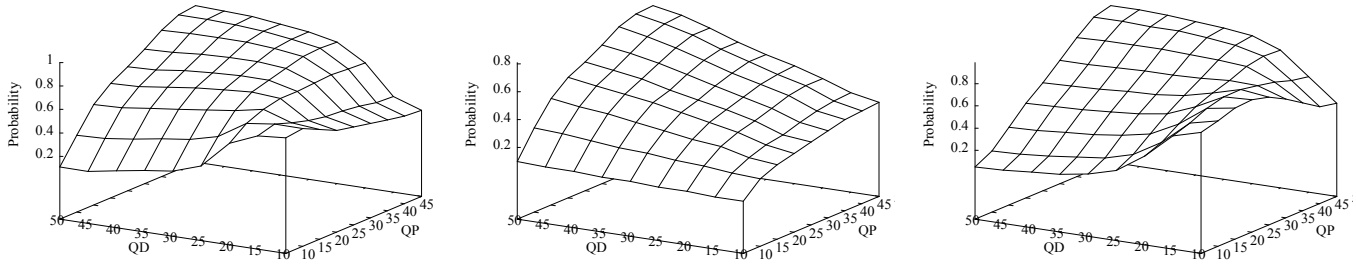

Fig. 3. Probabilities of choosing the same macroblock parameters for view and depth encoding. The same macroblock mode (left) for Book arrival and the same macroblock and prediction mode (middle) for Book arrival. The same macroblock mode for Newspaper on the right for comparison.

sequences. In this range the reduction of bitrate for the proposed codec is the most noticeable.

V. Quality evaluation

For video the most obvious quality measures are PSNR or subjective quality testing. Depth maps, however, are the auxiliary data, and are not meant to be presented to the viewers. They serve only the purpose of providing necessary data to synthesis algorithm. Therefore, it is better to evaluate the quality of a view synthesized from compressed data. The quality of the virtual view can be assessed using PSNR measure. The reference data can be either a view from a real camera positioned in the same location as the virtual camera or a view synthesized using uncompressed image and depth data. This way of evaluation of depth compression performance is adopted by MPEG in its works on 3D video [11].

It would be prohibitively time consuming to perform subjective visual tests for such a huge amount of cases as presented in this paper, however, previous works show that subjective quality measure tends to follow the trend of a PSNR measure when the synthesized view is compared to a real camera view [3].

VI. Test setup

In order to evaluate performance of the proposed codec, a series of tests were performed. In the tests, three test sequences were used: Book arrival [9], Newspaper [8] and Poznan street [10]. For each of them, two selected views were coded using proposed encoder and using standard MVC codec with view and depth data being coded separately. For both cases, reconstructed video and depth data were used as an input to a view synthesis procedure, performed using MPEG view synthesis reference software [12]. Resulting virtual view was compared to a real reference view from a camera (for sequences Book arrival and Newspaper) and, in the second case, to a view synthesized using uncompressed video and depth data (for all 3 sequences). For this case also Bjontegaard metrics [13] are presented for several QP values. For all cases, only Intra macroblocks were enabled during coding. The tests were performed using VLC coding.

VII. Coding results

Figures 4, 5 and 6 present the results obtained using proposed encoder. The results obtained using independent view and depth coding are also shown for reference.

TABLE III

Bjontegaard metrics for proposed technique

Sequence name	QP	ΔPSNR [dB]	Δbitrate [%]
Book Arrival	22	0.26	-4.0
	28	0.21	-8.3
	34	0.11	-10.9
Newspaper	22	0.05	-1.3
	28	0.21	-3.6
	34	0.12	-7.1
Poznan Street	22	0.14	-1.0
	28	0.10	-4.2
	34	0.05	-9.2

Bjontegaard metrics show a noticable gain of several percent in terms of bitrate gain, higher for the lower bitrates. In terms of PSNR, the gain is in the order of 0.1 to 0.2 dB.

On graphs, each line represents quality for the combinations where QP index is kept constant and only QD index is changed. The indexes' values range is 10 to 46 with step 3.

The greatest gains can be observed for the cases where QP and QD indexes have the similar value. This is especially visi-ble for QP values in the range of 20 to 30. It can also be seen that for the smallest QP values, below 20, there is a loss of efficiency of the proposed codec for QD values above 34. These cases are, however of little practical use.

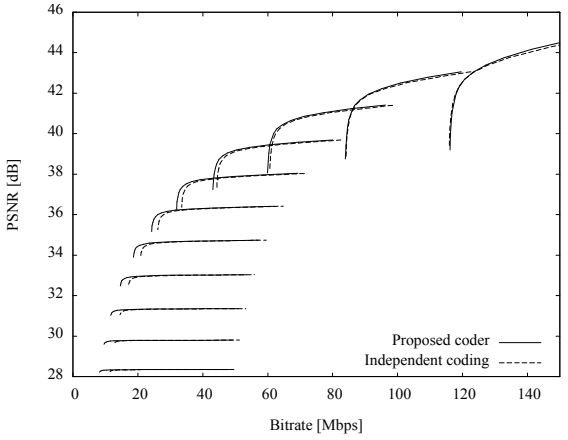

Fig. 4. View synthesis quality for Poznan Street sequence, referenced to virtual camera synthesized using uncompressed data.

VIII. Conclusions

The presented method of joint coding of views and depth data is superior to independent coding of video and depth maps. Compared to MVC in scenario presented on the left of Figure 1, it provides a coding gain at no cost. There is no

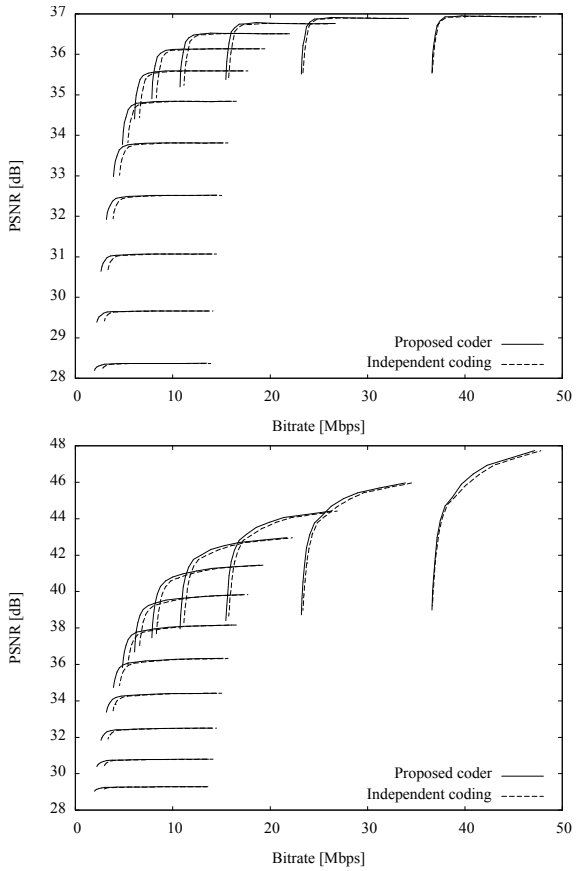

Fig. 5. View synthesis quality for Book Arrival sequence, referenced to real camera view (top) and virtual camera synthesized using uncompressed data (bottom).

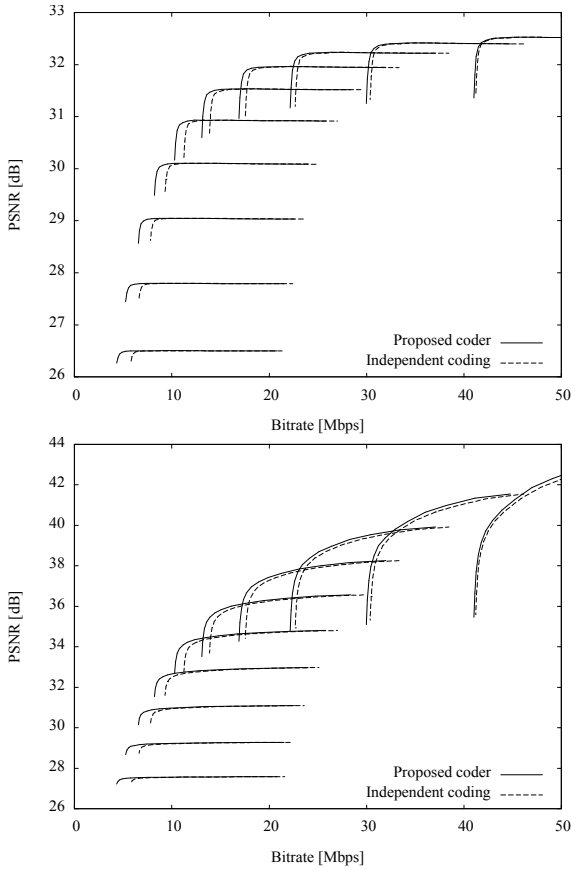

Fig. 6. View synthesis quality for Newspaper sequence, referenced to real camera view (top) and virtual camera synthesized using uncompressed data (bottom).

increase in the coding complexity, and the use of the proposed method requires only one pass of coding process. The gain is obtained thanks to omitting redundant control data and most of the mode selection information for depth. The proposed method produces a single stream, so there is no need to multiplex or synchronize two separate streams containing video and depth data. The proposed method is using the well known MVC codec as a base, contrary to other described methods [4], where a completely new technique is proposed, that might be more difficult to merge with the currently used compression tools.

ACKNOWLEDGMNENTS

The method described here was developed during the works that were supported by the public funds as a research project.

REFERENCES

[1] International Standard ISO/IEC 14496-10:2009, Information technology - Coding of Audio-Visual Objects, Part 10, Advanced Video Coding, 5th Ed. Annex H (2009).

[2] P. Merkle, A. Smolic, K. Müller, and T. Wiegand, “Multi-view Video Plus Depth Representation and Coding”, ICIP 2007, IEEE International Conference on Image Processing, San Antonio, TX, USA, September 2007.

[3] K. Klimaszewski, K. Wegner, M. Domański, “Influence of Distortions Introduced by Compression on Quality of View Synthesis in Multiview systems”, 3DTV-Conference 2009 The True Vision Capture, Transmission and Display of 3D Video, Potsdam (2009).

[4] Y. Morvan, P.H.N. de Witha, D. Farina, “Platelet-Based Coding of Depth Maps for the Transmission of Multiview Images”, Proceedings of SPIE, Stereoscopic Displays and Applications, vol. 6055 p. 93-100, January 2006, San Jose (CA), USA.

[5] P. Merkle, Y. Morvan, A. Smolic, D. Farin, K. Muller, P.H.N. de With, T. Wiegand, “The Effect of Depth Compression on Multiview Rendering Quality”, Proceedings of 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, pp. 245-248, May 2008.

[6] Sang-Tae Na, Kwan-Jung Oh, and Yo-Sung Ho, “Joint Coding of Multi-view Video and Depth Map”, ICIP 2008, IEEE International Conference on Image Processing, San Diego, California, USA, October 2008.

[7] Siping Tao, Chen Ying; M.M. Hannuksela, Ye-Kui Wang, M. Gabbouj, Li Houqiang, “Joint Texture and Depth Map Video Coding Based on the Scalable Extension of H.264/AVC”, ISCAS 2009, IEEE International Symposium on Circuits and Systems, Taipei, Taiwan, May 2009.

[8] Yo-Sung Ho, Eun-Kyung Lee, Cheon Lee, “Multiview Video Test Sequence and Camera Parameters”, ISO/IEC JTC1/SC29/WG11 MPEG2008/M15419, Archamps, France, April 2008.

[9] I. Feldmann, M. Mueller, F. Zilly, R. Tanger, K. Mueller, A. Smolic, P. Kauff, T. Wiegand, “HHI Test Material for 3D Video”, ISO/IEC JTC1/SC29/WG11 MPEG 2008/M15413, Archamps, France, April 2008.

[10] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, K. Wegner, “Poznan Multiview Video Test Sequences and Camera Parameters”, ISO/IEC JTC1/SC29/WG11 MPEG/M17050, Xian, China, October 2009.

[11] MPEG document “Description of Exploration Experiments in 3D Video Coding”, ISO/IEC JTC1/SC29/WG11 MPEG2010/N11274, Dresden, Germany, April 2010.

[12] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, Y. Mori, “Reference Softwares for Depth Estimation and View Synthesis”, ISO/IEC JTC1/SC29/WG11 (MPEG) Doc. M15377, Archamps (2008).

[13] G. Bjontegaard, “Calculation of average PSNR differences between RD-curves”, VCEG Contribution VCEG-M33, Austin, TX, USA, April 2001.