MULTI-VIEW VIDEO PLUS DEPTH REPRESENTATION AND CODING

Philipp Merkle, Aljoscha Smolic, Karsten Müller, and Thomas Wiegand

Image Communication Group, Image Processing Department
Fraunhofer Institute for Telecommunications - Heinrich-Hertz-Institut
Einsteinufer 37, 10587 Berlin, Germany
{merkle/smolic/kmueller/wiegand}@hhi.de

ABSTRACT

A study on the video plus depth representation for multi-view video sequences is presented. Such a 3D representation enables functionalities like 3D television and free viewpoint video. Compression is based on algorithms for multi-view video coding, which exploit statistical dependencies from both temporal and inter-view reference pictures for prediction of both color and depth data. Coding efficiency of prediction structures with and without inter-view reference pictures is analyzed for multi-view video plus depth data, reporting gains in luma PSNR of up to 0.5 dB for depth and 0.3 dB for color. The main benefit from using a multi-view video plus depth representation is that intermediate views can be easily rendered. Therefore the impact on image quality of rendered arbitrary intermediate views is investigated and analyzed in a second part, comparing compressed multi-view video plus depth data at different bit rates with the uncompressed original.

Index Terms— Multi-view video coding, video plus depth format, free viewpoint video, 3DTV, point-based rendering.

1. INTRODUCTION

Convergence of technologies from computer graphics, computer vision, multimedia and related fields together with rising interest in 3D television (3DTV) and free viewpoint video (FVV) lead to the development of these types of new media [1][2]. While 3DTV offers a 3D depth impression of the observed scenery (also known as stereo), FVV allows for interactive selection of viewpoint and direction within a certain operating range as known from computer graphics. Both technologies do not exclude each other, but rather can be very well combined within a single system. This progress is enabled by research and development, regarding the complete processing chain, from capturing, representation, compression, transmission to interactive presentation.

One common characteristic of many 3DTV and FVV systems is that they use multiple camera views of the same scene, often referred to as multi-view video (MVV). Since this approach causes a vast amount of data to be stored or transmitted to the user, efficient compression techniques are essential for realizing such applications. MVV contains a large amount of inter-view statistical dependencies, since all cameras capture the same scene from different viewpoints. These can be exploited for combined temporal/inter-view prediction, where images are not only predicted from temporally neighboring images but also from corresponding images in adjacent views [3]-[5], referred to as multi-view video coding (MVC).

A popular format for 3DTV uses a conventional color video and an associated per sample depth map [6]. MPEG specified a standard for efficient compression and transmission of such data. This format can be combined with multi-view video to the *multi-view video plus depth* format [7][8]. In this paper we investigate how multi-view video plus depth data can be compressed with MVC and how this compression influences the quality of rendered intermediate views in a FVV scenario.

The paper is organized as follows: Section 2 introduces the multi-view video plus depth representation. Section 3 describes and evaluates coding of multi-view video plus depth data and finally Section 4 analyses and evaluates the impact of this compression on the quality of rendered intermediate views for 3DTV or FVV applications.

2. MULTI-VIEW VIDEO PLUS DEPTH REPRESENTATION

The video plus depth representation used for 3DTV and FVV is an extension of the conventional color video format and suitable for rendering and compression [6]. For example stereo image pairs can be generated by view interpolation from one video and depth data associated to each sample. From the coding efficiency point of view the video plus depth format is also very valuable, as the per sample depth data can be regarded as a monochromatic, luminance-only video signal.

Fig. 1 illustrates the video plus depth format with an image and it's associated per sample depth map. The depth range is restricted to the minimum z_{near} and maximum z_{far} distance from the camera for the corresponding 3D points. With that, the depth map (Fig. 1 middle) is specified, resulting in a grey scale image. A sequence of such depth images can be converted into a YUV 4:0:0 format video signal and compressed by any state-of-the-art video codec.
The video plus depth format provides a very limited FVV functionality. If the head position of the user is tracked, the rendered stereo pair can be adjusted to the actual position. With

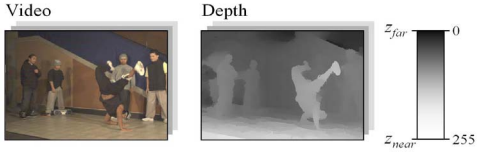

Fig. 1. Video plus depth data representation format consisting of regular 2D color video and accompanying 8-bit depth-images.

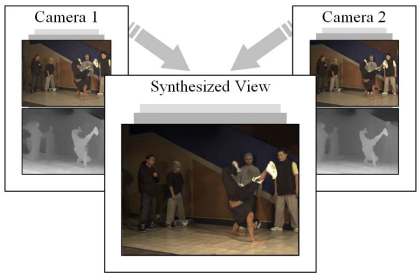

Fig. 2. Synthesis of arbitrary intermediate views from video plus depth of adjacent camera views.

this head motion parallax viewing becomes possible in a very limited navigation range. The extension to multi-view video plus depth extends the navigation range significantly [7][8]. Virtual intermediate views can be rendered for any position in between the cameras, thus providing advanced FVV functionality. Fig. 2 illustrates the functionality of synthesizing arbitrary intermediate views with this representation. Multi-view video plus depth data together with camera geometry provides the capacity of generating colored 3D point clouds [9]. This representation enables to render any intermediate view and free navigation in between the original cameras also enhancing 3DTV.

3. MVC FOR MULTI-VIEW VIDEO PLUS DEPTH

The results in [6] show that depth data can be very efficiently compressed using standard video coding algorithms. Therefore in this section compression of multi-view video plus depth data with compression algorithms for multi-view color video data is investigated.

3.1 Multi-view Video Coding

Encoding and decoding each view of a multi-view test data set separately, referred to as simulcast coding, can be done with any video codec including H.264/AVC [10]. This would be a simple, but inefficient way to compress multi-view video sequences, due to not exploiting the inter-view statistical dependencies. Regarding single view video coding, where prediction is limited to reference pictures in temporal dimension, the concept of *hierarchical B pictures* (see [11] for a detailed description) has proven to provide high compression efficiency. The coding schemes described below are well known from MVC and can identically be applied to multi-view depth data.

Fig. 3 illustrates how the concept of hierarchical B pictures is used with multi-view video plus depth data for a sequence with five cameras views and a GOP length of 8. Simulcast coding will be used as a reference to compare highly efficient temporal prediction structures with prediction structures that additionally use inter-view prediction. For multi-view video as well as for depth data the use of inter-view reference pictures offers the potential to improve coding efficiency. In order to exploit all statistical dependencies within a multi-view data set, inter-view prediction has to be combined with temporal prediction. Fig. 4 illustrates how the advantages of hierarchical B pictures are combined with inter-view prediction, without any changes regarding the temporal prediction structure. For view *Cam 1* the prediction structure is identical to simulcast coding and for the remaining views inter-view reference pictures are additionally used for prediction.

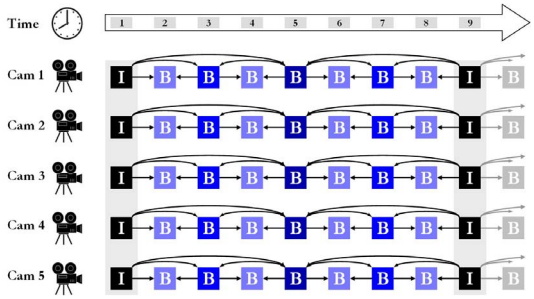

Fig. 3. Simulcast coding structure with hierarchical B pictures for temporal prediction (black arrows).

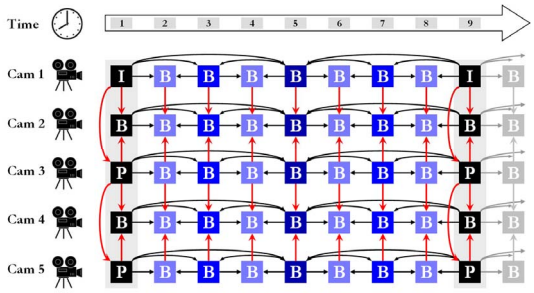

Fig. 4. Multi-view coding structure with hierarchical B pictures for both temporal and inter-view prediction (red arrows).

3.2 Coding Results

The MVC prediction structure presented in section 3.1 are realized using an H.264/AVC encoder with extended memory capabilities. For that the multi-view video sequences are combined into one single uncompressed video stream using a specific scan. This uncompressed video stream, containing either color or depth information, is used as the input of standard encoder software. The prediction structure itself is controlled by appropriate settings of the encoder's parameters for reference picture selection and memory management. This kind of encoder configuration is well-established for hierarchical B pictures with temporal prediction.

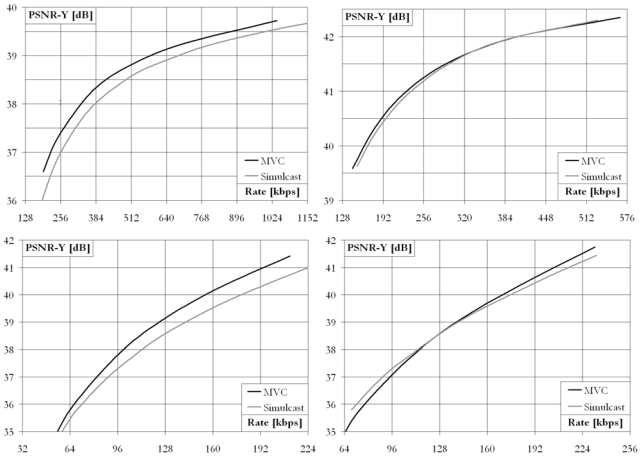

Fig. 5. Results of coding experiments for multi-view data (left: *Breakdancers*, right: *Ballet*, top: video, bottom: depth).

The results for simulcast and multi-view coding are shown in Fig. 5. The PSNR-Y values are plotted over bit rate averaged over all views of the two investigated multi-view video plus depth data sets. Their depth maps feature logarithmic quantization with a fixed depth range for the complete sequence. Although simulcast and multi-view coding perform almost identical for the *Ballet* sequence, both methods provide very efficient compression, especially for the color video component. Simulcast coding is outperformed by the MVC coding method for the *Breakdancers* sequence. Here compression performance is improved significantly by exploiting inter-view statistical dependencies for both components.

4. INTERMEDIATE VIEW RENDERING USING COMPRESSED MULTI-VIEW VIDEO PLUS DEPTH

The previous chapter investigated efficient compression techniques for multi-view video and depth data separately by comparing compressed pictures with their original in terms of PSNR. However, for multi-view video plus depth representations it is more important to evaluate the impact of coding on the visual quality of rendered arbitrary intermediate views. Especially the distortion of compressed depth data has to be studied, as these depth values represent geometrical 3D positions of scene points. By rendering the 3D point clouds of two adjacent cameras from arbitrary intermediate viewpoints, as already depicted in Fig. 2, the subjective and objective quality can be evaluated. For this purpose the results rendered from compressed multi-view video plus depth data at different bit rates are compared with the uncompressed originals in terms of PSNR and visual quality.

The multi-view video plus depth data sets used for this evaluation consist of linearly arranged camera views. For each camera view at a certain time instance, a 3D point cloud with per point color is generated. The point clouds of two adjacent cameras are then rendered from an intermediate viewpoint, resulting in the associated 2D image. Consequently the PSNR-Y calculation uses two such synthesized images with identical virtual viewpoint. The reference is synthesized from uncompressed video plus depth data, and the other one is synthesized from compressed data. For creating the intermediate view, the associated projection matrix has to be interpolated from the matrices of the two original cameras involved. This is achieved by using spherical linear interpolation [12], which originates from the context of quaternion curves. With this method it is possible to calculate the projection matrix of any intermediate view from two original camera's projection matrices. This means that a geometrically correct view from any position on a linear path between the two original cameras can be rendered.

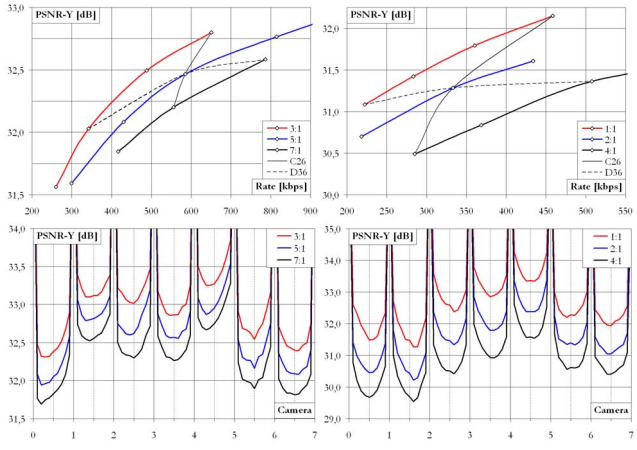

Fig. 6. Results of objective evaluation with compressed multi-view video plus depth data (left: *Breakdancers*, right: *Ballet*, top: equidistant intermediate views, bottom: virtual camera path).

The results of this evaluation are presented in Fig. 6, where PSNR-Y values are plotted over bit rate in the top diagrams. Each point in these charts represents the average value over the complete sequence. As the two investigated multi-view data sets consist of eight camera views, seven intermediate views are rendered and analyzed for each time-point. Each of the seven virtual camera views is equidistant from the two adjacent original camera views. The plotted rates represent the combined bit rate of color video and depth data as a per camera average value, resulting from the MVC coding experiments in section 3. The PSNR values of the bold curves are obtained by varying this bit rate, but keeping a constant ratio between the rates for color and depth. For example a ratio of *3:1* distributes 75% of the total rate to color and 25% to depth. Additionally *C26* connects rate points with identical quality of compressed color video with varying depth quality, while *D36* keeps constant depth and varies color quality accordingly. In the bottom diagrams of Fig. 6 PSNR-Y values are plotted over the rendered camera position for the three rate points of the *C26* curve. Here not only one but several intermediate camera views are rendered and analyzed. Accordingly the influence of depth and color compression is evaluated for a virtual camera path along the original cameras. Again each curve represents the average values over the complete sequence.

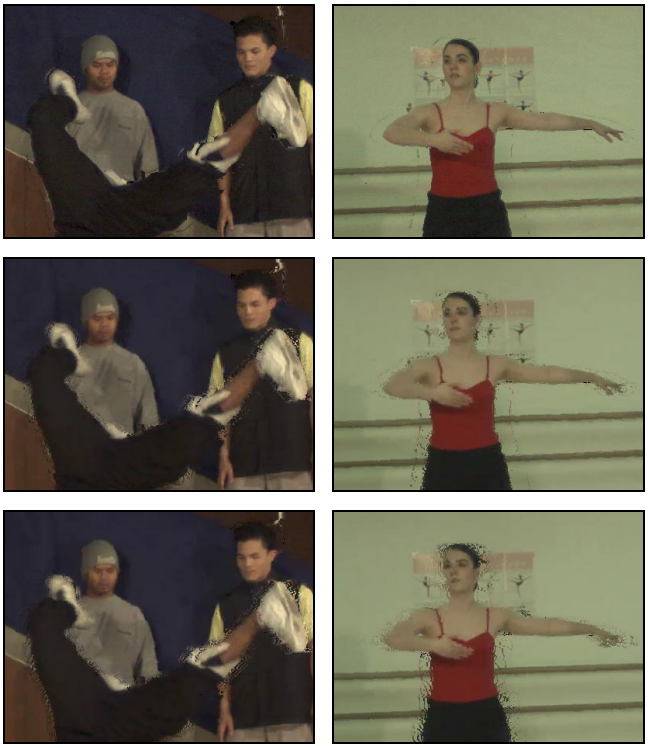

Fig. 7. Results of subjective evaluation with original and compressed multi-view video plus depth data (left: *Breakdancers*, right: *Ballet*, top: original, middle: *C26_red*, bottom: *C26_black*).

The results of both evaluations clearly indicate the strong influence of depth data coding quality on the quality of rendered intermediate views. With lower bit rate (and higher distortion) coding of the depth maps the distortion of rendered intermediate views increases, as coding leads to distorted areas especially around depth discontinuities at the borders of objects with different scene depth. On the other hand the quality of compressed color video influences the quality of rendered intermediate views much less, basically showing well-known effects like a loss of sharpness. The second evaluation additionally indicates that the quality of rendered intermediate views strongly depends on the distance between the virtual camera and an original camera. Due to coding artifacts in the depth maps the distortion of rendered intermediate views increases with increasing distance. Therefore the first evaluation analyses the worst case scenario by using the intermediate view equidistant from two original cameras. At original camera positions the PSNR-Y value does not at all depend on the depth information and is therefore equal to the results for color coding in section 3. These characteristics of the objective quality can be confirmed subjectively, as shown by the sample pictures in Fig. 7. The top pictures display details from original multi-view video plus depth data, middle and bottom from compressed data with constant color quality *C26*, but varying depth quality for highest (red bold line) and lowest (black bold line) render quality respectively. Again the strong correlation between compression level of depth data and the quality of rendered intermediate views becomes visible.

5. CONCLUSIONS

This paper presented a study on the multi-view video plus depth representation, where depth data is associated with color video data for each camera view of a multi-view video sequence. Starting from an evaluation of multi-view coding approaches for video plus depth, where the depth data associated with each view is treated as a monochromatic video sequence. Multi-view coding was applied similar to both color video and depth data. In a second step the influence of compression on the quality of synthesized intermediate views was investigated. The obtained results clearly show that coding artifacts on depth data strongly influence the reconstruction quality of rendered arbitrary views in a FVV scenario. For depth data coded at low bit rates, the reconstruction results show scattering artifacts, mainly around depth discontinuities, which occur at object boundaries. Consequently, reasonable results can only be obtained at relative high bit rates for depth data compression.

Possible extensions could include post-processing of the decoded depth maps, especially around the depth discontinuities of object boundaries, to avoid sample scattering. On the other hand new coding approaches might be appropriate, exploiting the specific statistics of depth data with its sharp edged low-frequency areas for objects. Such coding approaches need to preserve these boundaries as best as possible.

6. ACKNOWLEDGMENT

We would like to thank the Interactive Visual Media Group of Microsoft Research for providing the *Ballet* and *Breakdancers* data sets. This work is supported by European Commission Sixth Framework Program with grant No. 511568 (3DTV Network of Excellence Project).

7. REFERENCES

[1] A. Smolic, K. Müller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and Thomas Wiegand, "3D Video and Free Viewpoint Video – Technologies, Applications and MPEG Standards", *ICME 2006, IEEE International Conference on Multimedia and Expo*, Toronto, Ontario, Canada, July 2006.

[2] A. Smolic, and P. Kauff, "Interactive 3D Video Representation and Coding Technologies", *Proc. of the IEEE, Special Issue on Advances in Video Coding and Delivery*, vol. 93, no. 1, Jan. 2005.

[3] P. Merkle, K. Müller, A. Smolic, and T. Wiegand, "Efficient Compression of Multi-view Video Exploiting Inter-view Dependencies Based on H.264/MPEG4-AVC", *ICME 2006, IEEE International Conference on Multimedia and Exposition, Toronto*, Ontario, Canada, July 2006.

[4] K.-J. Oh, and Y.-S. Ho, "Multi-view Video Coding based on the Lattice-like Pyramid GOP Structure", *Proc. PCS 2006, Picture Coding Symposium*, Beijing, China, April 2006.

[5] Y. Yang, G. Jiang, M. Yu, F. Li, and Y. Kim, "Hyper-Space Based Multiview Video Coding Scheme for Free Viewpoint Television", *Proc. PCS 2006, Picture Coding Symposium*, Beijing, China, April 2006.

[6] C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W. Ijsselsteijn, M. Pollefeys, L. Vangool, E. Ofek, and I. Sexton, "An Evolutionary and Optimised Approach on 3D-TV", *IBC 2002, Int. Broadcast Convention*, Amsterdam, Netherlands, Sept. 2002.

[7] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger, "Depth Map Creation and Image Based Rendering for Advanced 3DTV Services Providing Interoperability and Scalability", *Signal Processing: Image Communication. Special Issue on 3DTV*, February 2007.

[8] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-Quality Video View Interpolation Using a Layered Representation", *ACM SIGGRAPH and ACM Trans. on Graphics*, Los Angeles, CA, USA, August 2004.

[9] S. Würmlin, E. Lamboray, and M. Gross, "3D video fragments: dynamic point samples for real-time free-viewpoint video", *Computers and Graphics 28 (1), Special Issue on Coding, Compression and Streaming Techniques for 3D and Multimedia Data*, pp. 3-14, Elsevier Ltd, 2004.

[10] H. Schwarz, D. Marpe, and T. Wiegand, "Analysis of hierarchical B pictures and MCTF", *ICME 2006, IEEE International Conference on Multimedia and Expo*, Toronto, Ontario, Canada, July 2006.

[11] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard", *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 13, No. 7, p. 560, July 2003.

[12] K. Shoemake, "Animating Rotation with Quaternion Curves", *ACM SIGGRAPH*, San Francisco, USA, July, 1985.