METHODS OF HIGH EFFICIENCY COMPRESSION FOR TRANSMISSION OF SPATIAL REPRESENTATION OF MOTION SCENES

Marek Domański, Adrian Dziembowski, Tomasz Grajek, Adam Grzelka, Łukasz Kowalski, Maciej Kurc, Adam Łuczak, Dawid Mieloch, Robert Ratajczak, Jarosław Samelak, Olgierd Stankiewicz, Jakub Stankowski, Krzysztof Wegner

Poznań University of Technology, Chair of Multimedia Telecommunications and Microelectronics, Poznań, Poland
domanski@et.put.poznan.pl, ostank@multimedia.edu.pl, jstankowski@multimedia.edu.pl

ABSTRACT[1]

The paper reports the hitherto obtained results of the research project that deals with scene representations that are suitable for free navigation around the scene. These scene representations are acquired from a limited number of cameras located around the scene. The project is aimed at the development of an efficient compression technology and the experimental environment for testing such technology. The paper reports some new multiview test sequences with the circular camera arrangement, including the hybrid ones that have been acquired using also the depth sensors. Finally, the paper also briefly reports the recent results of the new compression technology being under development. This technology outperforms the state-of-the-art 3D-HEVC.

Index Terms— Multiview video, free-viewpoint video, free navigation, 3D video compression, interview prediction

1. INTRODUCTION

The paper reports the project *Methods of High Efficiency Compression for Transmission of Spatial Representation of Motion Scenes* funded by National Science Centre of Poland. The project aims at gaining new knowledge about the compression methods suitable for multiview videos of natural scenes, captured by cameras. The emphasis is put on the videos acquired with the use of cameras placed around a scene nearly on an arc. This is the practical scenario foreseen for capturing the video for the free-viewpoint applications, including the free-viewpoint television (FTV) that is an interactive video service that provides the ability for a viewer to navigate freely around a scene [1,2]. Firstly, the project considers the reduction of the number of cameras that yields the reduction of the input data volume.

The project is not focused on video and depth acquisition. At the start of the project it was assumed that the respective multiview test data would be publicly available during the project, but it was not the case. Therefore, a substantial effort has been made in order to acquire the multiview test video sequences together with the corresponding depth data. This test material is needed for the experimental assessment of the compression techniques.

The implementation of this compression project requires the depth data for the scene representations. Therefore, an additional effort was needed to improve the system calibration techniques, the video correction techniques and the depth estimation techniques that needed to be adopted to the circular camera arrangements. As the quality of the decoded video is often measured by the quality of the synthesized virtual views, also the view synthesis software was adopted to the needs of the free navigation around a scene [3].

For the development of the new compression techniques, the 3D-HEVC software (HTM version 13.0) [4] was the starting point. The new and modified compression tools have been implemented on the top of this state-of-the-art 3D/multiview video compression software.

2. VIDEO COMPRESSION IN FREE-VIEWPOINT TELEVISON SYSTEMS

In the course of the project, the considerations were focused on practical free-viewpoint television systems that provide the feature of free navigation. In principle, the system consists of 4 basic units [5]:

- The content acquisition system (cameras, microphones, depth cameras, potentially light-field cameras);
- The representation server where the system calibration, the video and audio preprocessing and the 3D scene representation estimation (including the depth estimation) are implemented;
- The rendering server where the virtual views and the corresponding audio are synthesized according to the requests of the viewers;

[1] Research project was supported by National Science Centre, Poland, according to the decision DEC-2012/05/B/ST7/01279.

- The user terminal where requested views are presented together with the corresponding audio. The terminal is capable of bidirectional communications thus allowing the view requests to be transmitted in the uplink.

The block diagram of an FTV system is depicted in Fig. 1.

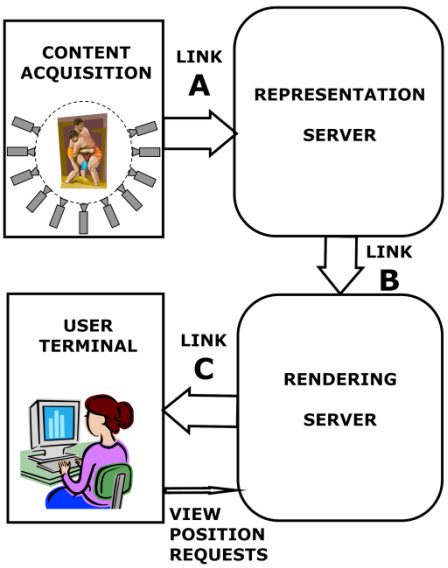

Fig. 1. The generic block diagram of an FTV system.

In general, the basic functional blocks of an FTV system are linked by three communication links named as A, B and C. Some communication links may vanish in particular configurations of the system. For example, the representation may be calculated directly on the site of video and audio acquisition. In such a case, the link A does not exist. The rendering server may be incorporated into the user terminal. Then the communication link C is superfluous. Also, the representation server may be co-located with the rendering server. In such a case, the link B would vanish. The characteristics of the individual communication links are as follows.

Link A is needed to transmit the multiview video together with the audio and the calibration data, and possibly also the data from the depth sensors or the light-field cameras. The link belongs to the contribution environment, and very high video quality is required. The depth is either not estimated yet or only raw depth data from the depth sensors are available, probably not useful for the improvement of video compression efficiency. The challenge is the high-fidelity multiview compression for the sparse real-camera locations distributed around a scene roughly on an arc, or even distributed more randomly.

Link B is between the representation server and the rendering server. The video quality must be high enough to synthesize the virtual views with the broadcast quality. Here, the likely data model is "multiview + depth" and an efficient compression technology is needed for views taken by the cameras sparsely distributed around a scene.

Link C is used to send the requested virtual view (or views) to the user terminal. The transmitted video may be monoscopic or stereoscopic, or even multiview with linearly and densely distributed views (for an autostereoscopic display). For all these cases of Link C, very efficient compression technology is available – either HEVC (High Efficiency Video Coding) [6] or MV-HEVC [7], or 3D-HEVC [8]. Therefore, Link C is not a challenge for compression technology.

A quite different situation appears for Links A and B if the source cameras are sparsely distributed in arbitrary locations around a scene. The state-of-the-art techniques MV-HEVC and 3D-HEVC are optimized for densely distributed cameras located on a line. For the circular camera arrangement and for the sparse camera distribution, their performance is not much better than just HEVC simulcast. For the inter-camera angle of about 10 degrees, the respective compression gain is around 3 – 13 % as compared to the simulcast [3].

For a practical system, at least one of Link A and Link B exists. These communication links need a new multiview video compression technology that will demonstrate the high compression efficiency for the circular camera arrangement and for the sparse camera distribution.

The respective international standardization project is likely to start soon, as MPEG has already organized an exploration expert group on FTV. This project should be based on the existing HEVC technology but it should be aware of the coming new successor of HEVC. For this paper, let us call this prospective compression technology SHEVC (Super High Efficiency Video Coding), see Fig. 2.

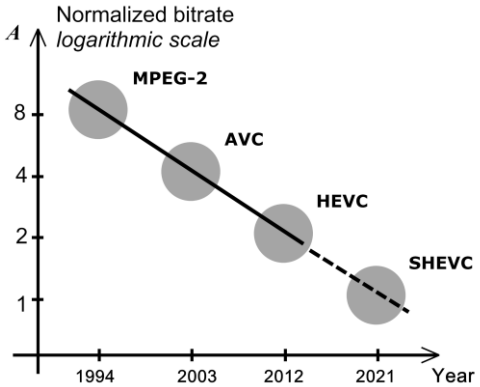

Fig. 2. The normalized bitrate (exactly the technology factor A from Eq. 1) for the consecutive compression generations.

For television services, for demanding monoscopic content, the average bitrate B may be very roughly estimated by the formula (a result of the research of an author)

$$B \approx A \cdot V \cdot 2^{19} \ \left[\text{bps}\right] , \qquad (1)$$

where A is technology dependent factor, A=1 for SHEVC, A=2 for HEVC, A=4 for AVC, A=8 for MPEG-2, and V is video format dependent factor V=1 for SDTV (720×576, 25i), V=4 for HDTV (1920×1080, 25i) and V=16 for UHDTV (3840×2160, 50p).

About every 9 years we have a new generation of video compression technology that provides halved bitrates. For the year 2021, one may expect the new compression technology (SHEVC) capable to reach 2 Mbps for broadcast-quality HD content, i.e. 20 Mbps would be needed for simulcasting 10-view video. For Link A the contribution quality would be expected, thus increasing the bitrate to the numbers exceeding 100 Mbps. Therefore, there is a need for research towards new technology that would efficiently exploit the inter-view redundancies for arbitrary camera locations. The technology should be built on the top of HEVC technology, and later it should be adopted to the prospective SHEVC in a similar way as MVC was adopted to HEVC resulting in MV HEVC.

3. TEST VIDEO SEQUENCES FROM CAMERAS LOCATED ON AN ARC

For the sparse circular camera arrangements and for the natural (not synthetic) scenes, the set of the publicly available multiview test video sequences is quite small hitherto. The best known are Microsoft 8-view 1024×768 sequences "Breakdancers" and "Ballet" [10]. Therefore, the authors have produced new HD (1920×1080, 25 fps) multiview test sequences that depict natural dynamic scenes. The sequences are available together with the respective camera parameters.

There are two groups of these test sequences produced by the authors at Poznań University of Technology:

1) The first group comprises the sequences acquired using 10 video cameras placed roughly on an arc with about 11 degrees spacing;
2) The second group comprises the test sequences acquired using 9 video cameras and 5 depth sensors. The video cameras are also placed on an arc but with about 13 degrees spacing. Each second video camera is co-located with a depth sensor (Fig. 3). Moreover, an additional camera is used to capture the view that corresponds to the intermediate view, in order to be able to estimate the fidelity of the virtual-view synthesis.

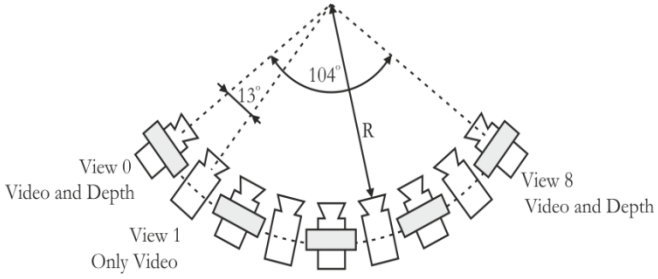

Fig. 3. The circular video camera setup with additional depth sensors.

In the first group we have the test sequences "Poznan Blocks" (indoor, the camera setup radius R ≈ 3 meters, see Fig. 4) and "Poznan Team" (outdoor, the camera setup radius R ≈ 15 meters, see Fig. 4). In the second group we have, for example, the test sequence "Poznan Service" (indoor, the camera setup radius R ≈ 3 meters, see Fig. 5). The is one of the very first test sequences produced for bi-modal depth estimation, i.e. both the outputs from the video cameras and from the depth sensors are available. Therefore, the depth may be estimated by merging the information from the two sources.

Fig. 4. Views 0, 5 and 9 of indoor and outdoor test sequences: "Poznan Blocks" (top) and "Poznan Team" (bottom).

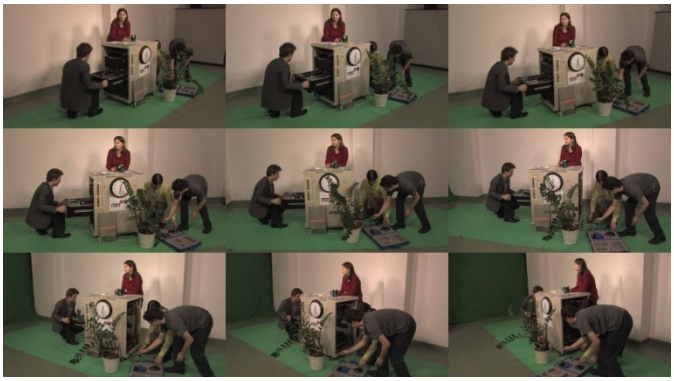

Fig. 5. Views 0 to 8 from the test sequence "Poznan Service."

The test sequences have been produced for the testing of the new compression techniques, but the authors believe that the sequences will be useful also for the other research on the multiview video processing. The abovementioned test sequences together with the respective camera parameters are now available to the FTV research community and can be obtained from the authors: {ostank, kwegner}@multimedia.edu.pl.

For circular camera arrangement, the authors have adopted also the techniques and the respective software for the camera system calibration, depth estimation and view synthesis [3].

4. 3D-HEVC ADOPTED TO CIRCULAR CAMERA ARRANGEMENT

For the sequences with the sparse circular camera arrangement, the results of the authors' experiments show, that the state-of-the-art 3D-HEVC codec provides only a small improvement over HEVC simulcast, i.e. bitrate is reduced only by 13.1% for "Poznan Blocks" test sequence,

and by 3.6%, for "Poznan Team" test sequence [3]. The average bitrate reductions are calculated using the Bjøntegaard formula [11] for the luma PSNR values in the range of about 33÷43 dB. The configuration parameters for all encoders were the same: intra-period = 24, GOP = 8, 1 slice per picture, SAO and VSO switched on.

For the sparse circular camera arrangements, the relatively poor compression efficiency of 3D-HEVC is related to the very simple inter-view prediction model. In 3D-HEVC, these predictions (of samples, motion vectors etc.) are made as the purely horizontal shifts defined by the disparity values. Such a prediction is unable to effectively remove the inter-view redundancy from the views with optical axes in the significantly different directions.

In our proposal, we replace the horizontal block shifts by the true mapping in the 3D space. The parameters of such mapping are completely defined by the intrinsic and the extrinsic camera parameters that should be acquired in the process of the calibration of the multicamera system.

The respective codec software was developed by the authors as a modification of the HTM 13.0 software of 3D-HEVC [4]. For the luma PSNR of about 35÷42 dB, the average bitrate reductions versus the standard 3D-HEVC are 5.7%, 6.9% and 6.2% for the test sequences "Ballet", "Breakdancers", and "Poznan Blocks", respectively.

5. ENCODING OF DISOCCLUDED REGIONS IN 3D-HEVC

The tool called Disoccluded Region Coding (DRC) was already proposed by the authors in the course of preparation of 3D-HEVC standard [12]. Now, this tool has been adopted to the arbitrary locations of the cameras and added on the top of the 3D-HEVC codec modified as described in the previous section of this paper. In DRC, the view synthesis is used as a primary inter-view prediction mechanism. With reference to the already encoded views and the respective depth maps, a virtual view is synthesized in the position of the view currently being coded. The locations and borders of the disoccluded regions can be estimated as a by-product of the view synthesis. This may be done in the same way both in the encoder and in the decoder. Therefore, there is no need to transmit any side information regarding the locations and borders of the disoccluded regions.

The preliminary experimental results demonstrate a significant encoder complexity reduction sometimes exceeding even 50%. For a given bitrate, the subjective quality is mostly similar or slightly better than for the modified 3D-HEVC. The numbers for the luma PSNR are the numbers for virtual views, thus they are not very meaningful. Nevertheless, these numbers remain comparable for DRC switched off and on.

6. CONCLUSIONS

The paper reports new and original results that comprise new multiview test sequences with the circular camera arrangement, the general considerations on compression in FTV systems and the original extensions of 3D-HEVC technology for multiview video obtained from cameras with arbitrary and sparse locations. The sequences are provided for the research usage. The sequence "Poznan Service" is probably the first publicly available multiview test video sequence that provides data for bi-modal depth estimation, i.e. from the video data and from the depth sensors. The proposed extensions of 3D-HEVC constitute the promising techniques that provide an increased compression performance for the content acquired from the cameras placed in arbitrary locations that are sparsely distributed around a scene.

REFERENCES

[1] M. Tanimoto, M. Tehrani, T. Fujii, T. Yendo, "FTV for 3-D spatial communication," *Proc. IEEE*, Vol. 100, pp. 905-917, April 2012.

[2] M. Domański, A. Dziembowski, A. Kuehn, M. Kurc, A. Łuczak, D. Mieloch, J. Siast, O. Stankiewicz, K. Wegner, "Experiments on acquisition and processing of video for free-viewpoint television," *3DTV-CON*, Budapest 2014.

[3] M. Domański, A. Dziembowski, D. Mieloch, A. Łuczak, O. Stankiewicz, K. Wegner, "A practical approach to acquisition and processing of free viewpoint video," *Picture Coding Symposium, PCS*, Cairns, June 2015.

[4] 3D HEVC reference codec available online https://hevc.hhi.fraunhofer.de/svn/svn_3DVCSoftware/tags/HTM-13.0.

[5] M. Domański, A. Dziembowski, K. Klimaszewski, D. Mieloch, A. Łuczak, O. Stankiewicz, K. Wegner, "Comments on further standardization for free-viewpoint television," *ISO/IEC JTC1/SC29/WG11* Doc. MPEG2015 /M35842, Febr. 2015.

[6] "High Efficency Video Coding", *ISO/IEC* Int. Standard 23008-2, *ITU-T* Rec. H.265, 2013.

[7] G. Tech, K. Wegner, Y. Chen, M. Hannuksela, J. Boyce, "MV-HEVC Draft Text 8", *JCT-3V of ITU-T, ISO/IEC* Doc. JTC3V-H1004, 2014.

[8] G. Tech, K. Wegner, Y. Chen, S. Yea, "3D-HEVC Draft Text 4", *JCT-3V of ITU-T and ISO/IEC* Doc. JTC3V-H1001, 2014.

[9] M. Domański *et al*, "Coding of multiple video+depth using HEVC technology and reduced representations of side views and depth maps," *29th Picture Coding Symposium*, *PCS*, Kraków, May 2012.

[10] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. *Microsoft Research 3D Video Download*. http://research.microsoft.com/en-us/um /people /sbkang/3dvideodownload, last visited: April 2015.

[11] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," *ITU-T Study Group 16 Question 6 (VCEG)*, Doc. VCEG-M33, Austin TX, March 2001.

[12] M. Domanski, O. Stankiewicz, K. Wegner, M. Kurc, J. Konieczny, J. Siast, J. Stankowski, R. Ratajczak, T. Grajek, *High efficiency 3D video coding using new tools based on view synthesis*, IEEE Transactions on Image Processing, vol. 22, pp. 3517 – 3527, 2013.