DEMONSTRATION OF A SIMPLE FREE VIEWPOINT TELEVISION SYSTEM

Marek Domański, Adrian Dziembowski, Tomasz Grajek, Adam Grzelka,
Krzysztof Klimaszewski, Dawid Mieloch, Robert Ratajczak, Olgierd Stankiewicz,
Jakub Siast, Jakub Stankowski, Krzysztof Wegner

Poznań University of Technology, Chair of Multimedia Telecommunications and Microelectronics, Poznań, Poland

ABSTRACT

The demonstration is aimed at presenting a simple free-viewpoint television system that is under development at Poznań University of Technology, Poznań, Poland. Video acquisition using sparsely distributed pairs of video cameras, the depth estimation from video, the virtual video rendering on a server feature the perspective FTV system. The original depth estimation and view synthesis algorithms ensure good quality of the virtual views generated during virtual walks around a scene. The system is developed for broadcasting sport (e.g. judo, karate, volleyball) and cultural (e.g. amateur or professional theater performances) events, as well as for interactive courses and manuals (medical, cosmetics, dancing, technical etc.).

Index Terms— virtual navigation, free-viewpoint television, multiview video.

1. INTRODUCTION

The virtual navigation is a functionality of future interactive video services where a user is able to navigate freely around a scene, also stepping towards and outwards the scene. The systems that provide such a functionality are often called free-viewpoint television (FTV) [1]. Here, we focus on the prospective FTV that will be an interactive internet-based system operating like a video on demand service that is shared by many viewers who all navigate independently from each other.

The system is developed for broadcasts of sport (e.g. judo, karate, volleyball) and cultural (e.g. amateur or professional theater performances) events or for interactive courses and manuals (medical, cosmetics, dancing, technical etc.

In an FTV system, a viewer watches a scene from virtual viewpoints on an arbitrary navigation trajectory. At each virtual viewpoint, the corresponding view has to be synthesized and made available at the receiver. Here, we use the centralized model of view synthesis where the views requested by all viewers are synthesized in the servers of the service provider [2], [3]. The centralized model may suffer from delays in the bidirectional server-to-terminal communications [3], [4], similarly to networked gaming. Therefore, the server-to-terminal distance has to be limited in order to limit the network delay.

The demonstration is aimed at presenting new results in the design and practical implementations of the FTV systems that use sparse setups of moderate numbers of cameras. The system is developed using some extended and improved solutions already described in [5],[6].

The crucial block of the system is the one related to preparation of the 3D scene representation. The acquisition system comprises cameras with microphones. The authors have experimented with depth sensors but finally the depth estimation is developed as software using video data from several cameras. The depth sensors are not used because they face severe problems related to limited resolutions of the acquired depth maps, mutual interferences between several sensors simultaneously used in the same system, limited measured distance ranges, problems related to synchronization of the video and depth cameras, and sensitivity to the environmental factors including solar illumination. Moreover, we focus on the multiview recording of real events where additional infrared illumination might be unacceptable.

The demonstration presents the results related to the consecutive blocks of the system.

2. MULTIVIEW VIDEO ACQUISITION

It has already been discovered that for the scenes with substantial amount of occlusions, the use of camera pairs is more efficient than the uniform distribution of the camera locations around a scene [6],[7]. Therefore, the video acquisition system consists of camera pairs sparsely distributed around a scene. For practical reasons, the locations usually exhibit some irregularities due to the limitations of a real scene, e.g. the requirements to leave free communication routes, windows, commercial banners and displays etc. The distances between camera pairs depend on the dimensions of a scene, on the distances between the scene and cameras, and on camera lenses and sensor parameters. The video cameras need to be precisely synchronized, i.e. the synchronization jitter has to be much

smaller than the shutter opening time. For the sake of simplicity, the system developed employs video cameras with the lenses of fixed focal length.

3. SCENE REPRESENTATION SERVER

The scene representation server is the critical part of the system. It outputs the corrected multiview video and the respective depth maps.

The software of the scene representation server includes the calibration software that calculates the intrinsic and the extrinsic camera parameters.

The other video processing pipeline comprises the following blocks:

- Colour correction and illumination differences compensation;
- Depth estimation: The depth is estimated always from at least two camera pairs: a pair with a small and with a large camera base (the respective algorithms and software were developed by the authors).

Among the abovementioned tasks, the depth estimation seems to be the most challenging tasks. The fidelity of the estimated depth maps defines the quality of the virtual views. Therefore the authors have decided to develop special software for the multiview plus video compression.

The scene representation server produces also the respective 3D audio signals.

4. EFFICIENT MULTIVIEW PLUS DEPTH VIDEO COMPRESSION

The visual scene representation needs to be compressed prior to being transmitted to remote edge servers (rendering servers) that render the views requested by the users. In the scene representation, only one video channel and one audio channel is included from each camera pair.

Multiview video with depth may be compressed using the state-of-the-art 3D-HEVC encoder. Nevertheless, for the sparse 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.

Therefore, the authors have adopted the 3D-HEVC video compression scheme to the cameras allocated around a scene. 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 calibrating the multicamera system [8]. The proposed compression technique provides some further bitrate reduction as compared to 3D-HEVC.

The audio signals may be transmitted using the state-of-the-art 3D audio codecs but also simulcast coding of all audio channels might be acceptable.

5. RENDERING SERVER

The rendering server responds to the requests from a user and streams video for the requested viewpoint. This requires the video frames to be synthesized according to the current viewpoint defined by a user. For cameras located on an arc, the synthesis is more complex than for the linear camera arrangement but still doable in real time. Thus, we demonstrate the real-time software for virtual view synthesis that is capable of processing the input views from the cameras located around the scene, further is capable to synthesize the arbitrary view from any virtual viewpoint around the scene (also with close-ups).

From the rendering server to a user, the video is streamed as a single bitstream. Each such stream has to be generated individually for each user terminal, as the frames correspond to various viewpoints, or more generally viewports, currently chosen by an individual user. In order to avoid substantial coding delays, the video is encoded in the all-intra mode.

The demonstration includes also the presentation of the service where a viewer uses his or her smartphone or tablet to watch the virtual video and to change the virtual observation point.

6. RESULTS

In the demonstration we present an entire structure of an original simple low-cost FTV system in contradiction to the sophisticated systems proposed in other contributions. Also straightforward schemes for the rendering server, as well as for video streaming are proposed. The novelty of the demonstration is related also to the proposal of building the acquisition system using the two-camera modules. The advantages of such a system were already demonstrated empirically using some video sequences acquired from such camera modules [6], [7].

From the perspective of a user, the FTV service is available through a webpage, thus no special application needs to be installed on the user's device. From the business perspective, it is a kind of the video-on-demand service.

All these results together with other results cited in the paper encourage us to believe that the development of usable FTV systems will be possible within the next very few years.

ACKNOWLEDGEMENT

The research project was supported by The National Centre for Research and Development, Poland. Project no. TANGO1/266710/NCBR/2015.

REFERENCES

[1] M. Tanimoto, et al., "FTV for 3-D spatial communication", *Proc. IEEE*, vol. 100, pp. 905-917, 2012.

[2] M. Domański, "Practicing free-viewpoint television: multiview video capture and processing," in: M. Tanimoto, T. Senoh, "FTV seminar report," *ISO/IEC JTC 1/SC 29/WG 11*, Doc. MPEG M34564, July 2014.

[3] J. Kim, J. Jang, D. Ho Kim, "Design of platform and packet structure for the free-viewpoint television", *18th IEEE International Symposium on Consumer Electronics*, Jeju Island, 2014.

[4] J. Osada, N. Fukushima, Y. Ishibashi, "Influence of network delay on viewpoint change in free-viewpoint video transmission", *18th Asia-Pacific Conference on Communications*, Jeju Island, pp. 110 –115, 2012.

[5] 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 2015*, Cairns, 2015, pp. 10-14, also in IEEExplore.

[6] M. Domański, M. Bartkowiak, A. Dziembowski, T. Grajek, A. Grzelka, A. Łuczak, D. Mieloch, J. Samelak, O. Stankiewicz, J. Stankowski, Krzysztof Wegner, "New results in free-viewpoint television systems for horizontal virtual navigation," in *2016 IEEE International Conference on Multimedia and Expo (ICME),* Seattle, WA, 2016, pp. 1-6.

[7] M. Domański, A. Dziembowski, A. Grzelka and D. Mieloch, "Optimization of camera positions for free-navigation applications," in *2016 International Conference on Signals and Electronic Systems* (ICSES), Kraków, 2016, pp. 118-123, also in IEEExplore.

[8] J. Stankowski, et al., „3D-HEVC extension for circular camera arrangements," *3DTV-CON*, Lisbon 2015.