EXPERIMENTS ON ACQUISITION AND PROCESSING OF VIDEO FOR FREE-VIEWPOINT TELEVISION

Marek Domański, Adrian Dziembowski, Agnieszka Kuehn, Maciej Kurc, Adam Łuczak, Dawid Mieloch, Jakub Siast, Olgierd Stankiewicz, Krzysztof Wegner

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

ABSTRACT

The paper describes an experimental multiview video production, processing and delivery chain developed at Poznań University of Technology for research on free-viewpoint television. The multiview-video acquisition system consists of HD camera units with wireless synchronization, wireless control, video storage and power supply units. Therefore no cabling is needed in the system, which is important for shooting real-world events. The system is mostly used for nearly circular setup of cameras but the locations of cameras are arbitrary, and the procedures for system calibration and multiview video correction are considered. The paper deals also with adoption for circular camera arrangement of the techniques implemented in Depth Estimation Reference Software and View Synthesis Reference Software.

Index Terms — FTV, multi-view, depth, rectification, view synthesis, circular setup, test video, frame grabber

1. FREE-VIEWPOINT TELEVISION

Free-Viewpoint Television (FTV) is an interactive video service that provides an ability for a viewer to choose freely an arbitrary virtual viewpoint and to move it smoothly [1,2]. Recent research resulted in significant progress in 3D scene representation [3], 3D video coding (e.g. [4]), multiview streaming, 3D displays (e.g. [5]), depth estimation (e.g.[6]) and view synthesis (e.g. [7]). Such achievements provide a solid basis for fast development of FTV technology and services. These achievements as well as interests of industry were already recognized by the ISO-based expert group MPEG that launched exploration activities towards FTV [8]. Now, the demand is growing to provide practical results on integration of individual techniques into experimental platforms. There are various attempts to achieve that. Some approaches relay on sophisticated 3D video acquisition systems [9,10], while a more straightforward approach is proposed in this paper. Here, we describe an FTV experimental processing chain with a simple and practical multi-camera video capture system.

2. MULTIVIEW VIDEO CAPTURE SYSTEM

An important requirement for the multi-camera system is its practicality including moderate cost, portability, easy operation etc. These features may be obtained by application standard portable television cameras usually used by reporters. The cameras should be equipped with the global shutter that provides exposure of all the pixels at the same time. This allows correct estimation of depth maps and proper spatio-temporal processing of the captured multiview video. In many cameras the rolling shutter is used. In such cameras a frame is captured not from the whole sensor, but sequentially from fragments of the sensor. In extreme cases, a frame is collected line by line, vertically or horizontally. Using such cameras, the whole image could not be recorded at the same time in all cameras, and the correct depth map estimation would be impossible.

For experiments, 10 HDTV global-shutter cameras (Canon XH G1) were used in nearly circular arrangement (Fig. 1). In order to avoid shooting the cameras themselves, their locations are limited to an angle of about 120 degrees. The viewer watches the scene from a virtual viewpoint located arbitrarily around the scene. For the sake of simplicity, in the current system, the virtual viewpoint is allowed to navigate on one horizontal plane. Nevertheless, the distance between a virtual camera and the scene may be set freely, thus no camera zooming is needed while shooting multiview video. Therefore, the complex and unreliable hardware zooming of multiple cameras is superfluous as it is replaced by software selection of a virtual camera position.

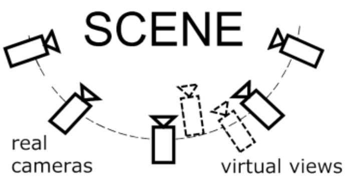

Fig. 1. Circular setup of the cameras

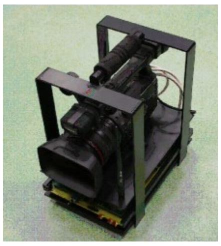

Fig. 2. Wireless camera module

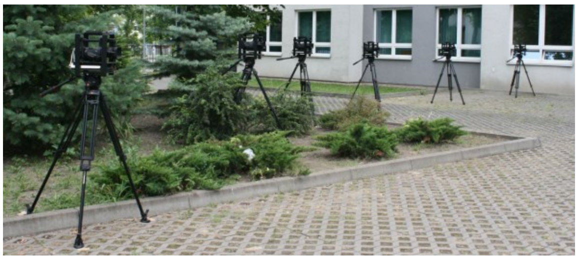

Fig. 3. Wireless multi-camera system for FTV test video capture

The multiple cameras need common control and exact synchronization. Together with signal and power supply these needs would create plethora of cables. In order to avoid problems with cable unplugging and cable layout around a real-world scene, a wireless camera module was developed by the authors (Fig. 2). Such a module can be easily mounted on a tripod thus allowing to place a camera in an arbitrary location (Fig. 3). Each module has local power supply and hard disk to store captured video. Therefore the system operates with no cable connecting the cameras. Such a system is used for acquisition of uncompressed multi-camera video, both indoor and outdoor.

The experimental system was built having in mind such future FTV applications like, e.g. sport broadcasts (like judo, wrestling, sumo, dance etc.), performances (theatre, circus), interactive courses (medical, cosmetics, dance etc.), manuals and school teaching materials.

3. WIRELESS CAMERA MODULE

The camera modules with all building blocks (Fig. 2) have been designed, developed and produced by the authors, and the authors do not know any other equivalent design elsewhere. Each camera module is designed for a single HD camera with an HD SDI output and a Genlock input/output. The modules provide wireless bi-directional communication with the management laptop. In that way a single human operator is able to control and manage capturing of multiview video. The essential digital circuitry of the module was developed using FPGA devices. The basic blocks of the camera module are described below (Fig. 4).

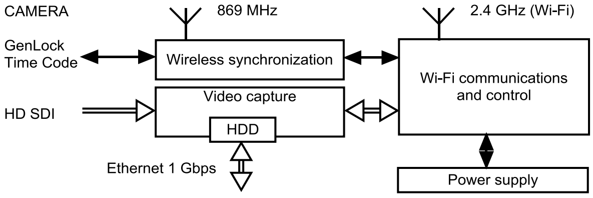

Fig. 4. The structure of the camera module

Wireless synchronization block includes integrated receiver and transmitter (FSK, Manchester codes) for Genlock signals and Time Codes transmitted in the license-free 869 MHz band. The maximum transmission rate of 30 kbps is sufficient for the Genlock-based frame synchronization by 25 or 30 fps. Measurements have shown that synchronization inaccuracy is within ±2 μs i.e., more than 2000 times less than the shutter speed that is 1/250 ÷ 1/25 second or 4 ÷ 40 ms. The Time Codes are stored for all frames in all camera modules for processing and display synchronization.

One camera module is programmed as a host that synchronizes generators of Genlock signals in the slave modules. The master module also distributes Time Codes to all slave modules.

Video capture block inputs the HD SDI signal from the camera and stores video data on a hard disk mounted in the module. The main part of this block is a frame grabber for 1920 × 1080, 25 fps video. The frame grabber was built by the authors on the basis of the XILINX Spartan-6 FPGA device. Its task is also to process video in the 10-bit 4:2:2 format produced by the camera at its HD SDI output. This video is converted on fly into the 8-bit 4:2:0 format, thus the bitrate is reduced enough to write the video data from one camera directly into a single SATA hard disk. After experimenting with a number of hard disks available on the market, the authors have chosen Seagate MOMENTUS ST95005620AS disks with capacity of 500 GB and the maximum write throughput of about 100 MBps needed for frame grabbing in real time.

Power supply block contains a 11Ah Li-ion battery that allows about 100 minutes of HD video recording. The power supply block maintains also the supply voltages during the switchover between battery and external supply for the battery charger.

Wi-Fi communication and control block controls the operation of the whole module and provides bi-directional communication with the laptop used by the human operator to manage the system. The Wi-Fi communication with the laptop is implemented as 802.11b/g link. After considering a number of products available on the market, the integrated circuit WIZ610wi (Wiznet) was selected. It is a low-power 802.11b/g wireless LAN client/AP/gateway with built-in IP protocol stack and TCP/UDP endpoint. This block comprises also a microcontroller (STM32F207) with MII/RMII Ethernet interface.

The tasks of this block are the following:

- management of the wireless network connection,
- passing commands from management laptop to other modules and passing status data from other modules to the management laptop through the wireless channel,
- periodical transmission the telemetry data (Time Code, current frame number, synchronization status, battery voltage, battery charger status etc.) to the management laptop,
- monitoring the status of each camera by sending to the laptop reading low-resolution frames (128 × 71, 8 fps),
- sending via Wi-Fi single full-resolution frames (before shooting the scene or even during shooting) from the camera to the laptop in order to calibrate the camera system.

4. CALIBRATION OF THE VIDEO CAPTURE SYSTEM

It is a very idealistic assumption that the cameras are located on a circle. Various obstacles, e.g. pillars in buildings, scene items, often cause that actual locations are quite far from a circle. Moreover, the cameras are mounted on individual tripods and their positions are difficult to control exactly. Therefore, the camera locations must be considered as arbitrary.

For calibration, it is very helpful to ensure that a calibration pattern, e.g. checkerboard (Fig. 5a), is visible by possibly many cameras simultaneously. In the course of the work, the authors experimented with various calibration patterns moving in the fields of view of all cameras. For example, good results have been obtained by the use of lighting diode (Fig. 5b). The authors also experiment with characteristic points in a scene used to calibrate the system.

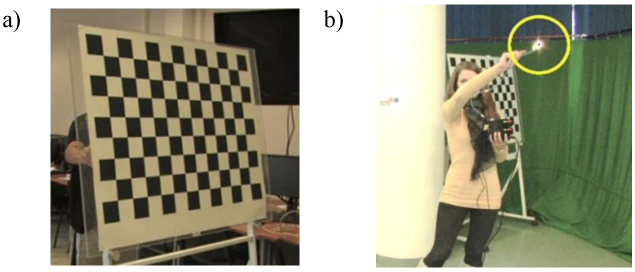

Fig. 5. a) A checkerboard used in the experiments,
b) lighting diode used for system calibration

The actual calibration is preceded by correction of barrel distortions of the lenses. In the current implementation, this operation is limited to the correction of radial distortions only, using the approach from [11,12].

The system calibration is aimed at precise estimation of intrinsic and extrinsic parameters of the multi-camera system [13]. Here, the authors adopted techniques [11,14] for circular camera arrangement. Moreover, the camera color characteristics are equalized during color correction of the captured views. For the sake of brevity, the details are omitted in this paper.

5. TEST SEQUENCE PRODUCTION FOR FTV

Preparation of the proper test material is an important and challenging task for future research. The authors use similar approach as it was used before for production of well-known standard test sequences for linear camera arrangement [15]. Using the new acquisition system (cf. Section 2), the new 10-view uncompressed sequences have been produced with cameras located circularly around scenes. Among the produced test sequences, there is the sequence "Poznan Blocks" [16] already provided for the FTV research community (Fig. 6). For research purposes, this sequence and the respective camera parameters

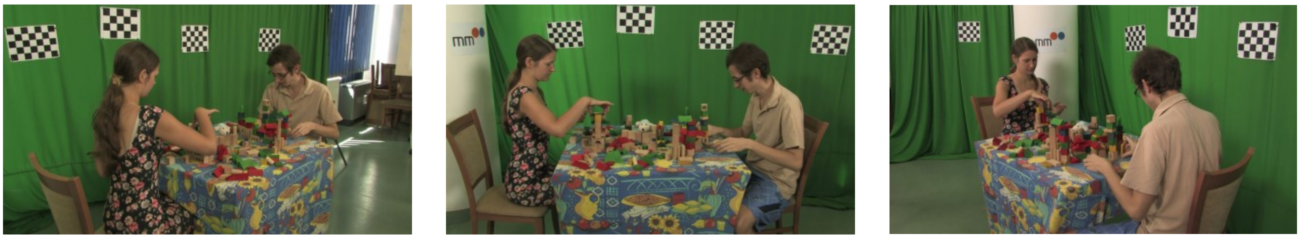

Fig. 6. A frame from the test sequence “Poznan Blocks”: the leftmost view (view 0), the 4th view, the rightmost view (view 9)

may be obtained from the authors (email: {ostank, kwegner} @multimedia.edu.pl). The features of this test sequence are the following: the angle between outermost cameras – about 100 degrees, resolution 1920 × 1080, 25 frames per second, progressive scan, 10 views, 1000 frames, i.e. 40 seconds.

6. VIDEO COMPRESSION

FTV application scenarios foresee transmission of multiview video to a server for further processing. For simulcast coding of the views, the state-of-the-art video compression technique HEVC (High Efficiency Video Coding) [17] can be easily used. Nevertheless, simulcast coding does not exploit redundancies between the individual views but there are already available very efficient 3D video coding techniques (e.g. [4]) that exploit extensively 3-D scene models but these techniques are yet not standardized. For experiments, the multiview video coding extensions of HEVC are used: MV-HEVC [18] and 3D-HEVC [19]. Both techniques exploit inter-view predictions, but MV-HEVC uses only standard HEVC low-level tools while 3D-HEVC uses some additional low-level coding tools.

Here, we report the very first HEVC results for multiview video with circularly distributed cameras. The configuration parameters for all three encoders (software version HTM 10.0) are the same: intra-period = 24 (each intra picture is IDR), GOP = 8 (each 8th inter-coded picture has no reference to future frames), 1 slice per picture, SAO and VSO switched on. For compression of the views 4, 5 and 6, the MV-HEVC and 3D-HEVC provide about 11 % and 13 % average bitrate reduction as compared to simulcast HEVC. This experiment demonstrates acceptable bitrates even for high quality (Fig. 7).

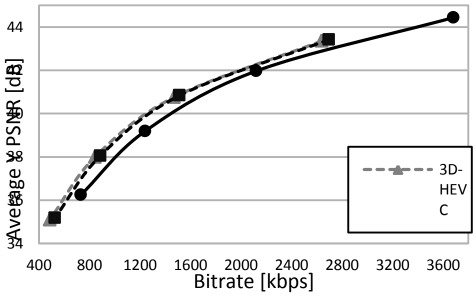

Fig. 7. Compression of the “Poznan Blocks” sequence (the first 50 frames): the average luma PSNR versus the total bitrate for 3 views

7. DEPTH ESTIMATION

Free navigation requires that a virtual view can be synthesized from an arbitrary position requested by a viewer. Such synthesis mostly exploits depth maps pre-calculated for individual views. Depth estimation is a computationally heavy task. For 3-D video processing, the state-of-the-art technique is the technique implemented in Depth Estimation Reference Software (DERS) used by the ISO/IEC MPEG in research related to standardization. Recently, the DERS technique has been extended for processing of multiview video with arbitrary camera positions [6].

In the original DERS correspondence search between the views is performed along horizontal lines. Such an approach is feasible for rectified views. For the circular camera arrangement, the rectification often cannot be performed either because it is technically not attainable (e.g. due to positions of the cameras) or because it would distort the views at unacceptable level. As the rectification is a transformation that aligns epipolar lines to horizontal rows in all views, the more skewed is the camera position, the more distorted is the rectified view. Therefore, an approach is employed that does not require the views to be rectified. Correspondence matching (e.g. pixel-, block- or soft-segment-based) is done along epipolar lines (Figs. 8, 9) calculated according to the camera parameters.

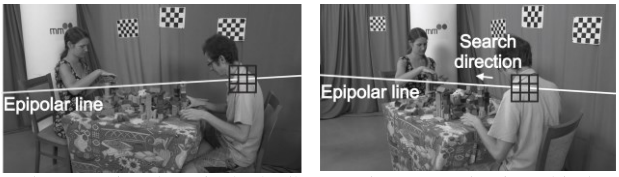

Fig. 8. Explanation of epipolar line search on example of 3×3 block matching in “Poznan Blocks” sequence.

For the circular camera arrangement, also the search range has to be defined in another way as compared to the standard DERS algorithm. Now, it is defined using the minimal and the maximal distances from the camera, denoted as z_{near} and z_{far}, respectively (Fig. 9). The search is done along the epipolar line, i.e. the depth z values are estimated. The output depth is represented as

$$\delta = \delta_{max} \cdot \left(\frac{1}{z} - \frac{1}{z_{far}}\right) / \left(\frac{1}{z_{near}} - \frac{1}{z_{far}}\right)$$

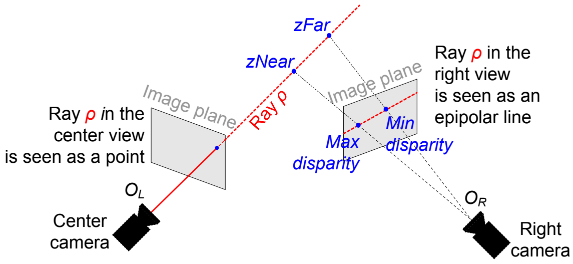
zFar
zNear
Ray ρ
Image plane
Image plane
Ray ρ in the right view is seen as an epipolar line
Ray ρ in the center view is seen as a point
Max disparity
Min disparity
O_L
O_R
Center camera
Right camera

Fig. 9. Search correspondence range for arbitrary camera setup.

Standard 8-bit depth format is not sufficient for the wide range of depth values that appear in multiview video captured with circularly distributed cameras. Therefore, the 16-bit depth format thus $\delta_{max} = 2^{16} - 1 = 65535$.

In the original DERS algorithm the dense linear camera arrangement was assumed. For circular camera arrangement, the content in subsequent views can be very different as the scene is viewed from different angles. The reduced overlap in the content results in more severe occlusion problems. In order to cope with them, the improved graph-cuts depth estimation algorithm is used, which iteratively estimates occlusions from several views [20]. In that way the high-quality depth maps are generated (Fig. 10) for the circularly-arranged multiview video.

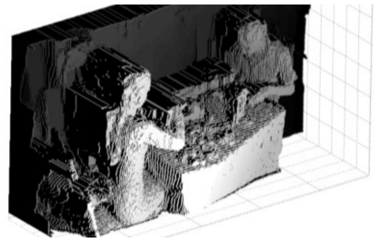

Fig. 10. Exemplary depth map for "Poznan Blocks" sequence, obtained using enhanced Depth Estimation Reference Software.

8. VIEW SYNTHESIS

In FTV, the viewer watches the synthesized views. In order to produce high-quality virtual views, the authors have modified the state-of-the-art algorithm implemented as View Synthesis Reference (VSRS) [7]. The modifications include the synthesis in 3D space, instead of row-wise warping, which was possible in linear camera arrangement. Moreover, several views are used for synthesis instead of the nearest views used in the classic approach. Such synthesis with 16-bit depth maps results in the virtual views that exhibit good perceptual quality (Fig. 11).

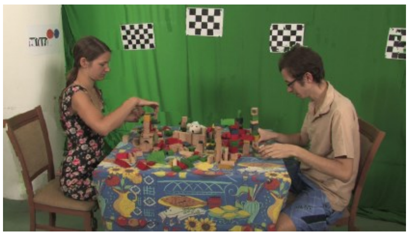

Fig. 11. A synthesized virtual view (position between cameras between the real cameras 4 and 5)

9. CONCLUSIONS

In the paper, described is a simple multiview video capture system and also the processing chain for experiments with free navigation around a dynamic scene. The proposed system and techniques are meant to be used in development of simple FTV systems where multi-camera video capturing will be operated by a single person. The described results suggest that development of such a practical simple and moderate-cost system is feasible in the very near future. This conclusion might be important for future development of FTV systems where prospective operators are scared about the expected high content production costs. The authors demonstrated that even their low-cost multi-camera systems can be efficiently used for preliminary educational and entertainment FTV broadcasts.

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

REFERENCES

[1] M. Tanimoto, M. Tehrani, T. Fujii, T. Yendo, "Free-viewpoint TV", *IEEE Signal Proc. Magazine*, vol. 28, pp. 67-76, Jan. 2011.

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

[3] M. Daribo, I. Cheung, G. Frossard, "Navigation domain representation for interactive multiview imaging", *IEEE Trans. Image Proc.,* vol. 22, pp. 3459 – 3472, Sept. 2013.

[4] M. Domański, 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 Trans. Image Proc*., vol. 22, pp. 3517 – 3527, Sept. 2013.

[5] H. Horimai, "360 degree viewable 3D Display Holo-Table", *ISO/IEC JTC 1/SC 29/WG 11*, Doc. M28301, Jan. 2013.

[6] O. Stankiewicz, K. Wegner, M. Tanimoto, M. Domański, "Enhanced Depth Estimation Reference Software (DERS) for Free-viewpoint Television", *ISO/IEC JTC 1/SC 29/WG 11*, Doc. MPEG M31518, Oct. 2013.

[7] O. Stankiewicz, K. Wegner, M. Tanimoto, M. Domański, "Enhanced View Synthesis Reference Software (VSRS) for Free-viewpoint Television", *ISO/IEC JTC 1/SC 29/WG 11*, Doc. M31520, Oct. 2013.

[8] M. Tanimoto, T. Senoh, S. Naito, S. Shimizu, H. Horimai, M. Domański, A. Vetro, M. Preda, K. Mueller, Proposal on a new activity for the third phase of FTV, *ISO/IEC JTC 1/SC 29/WG 11,* Doc. M30229/M30232, Vienna, Austria, July/Aug. 2013.

[9] M. Tanimoto, "Ray capture systems for FTV", *Signal Information Proc. Assoc. Annual Summit Conf. APSIPA ASC*, 2012, pp. 1 – 6.

[10] S. Prince, A. Cheok, F Farbiz, T. Williamson,N. Johnson, M. Billinghurst, H.Kato,"3D live: Real time captured content for mixed reality". *Int. Symp. Mixed and Augmented Reality ISMAR'02*, pp. 7–13, Sept. 2002.

[11] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations", *IEEE Int. Conf. Computer Vision*, vol. 1, pp. 666-673, 1999.

[12] A. Łuczak, S. Maćkowiak, D. Karwowski, T. Grajek, "A large barrel distortion in an acquisition system for multifocal images extraction", *Lecture Notes in Computer Science*, vol. 7594, ss. 164-171, Springer 2012.

[13] B. Cyganek, J. Siebert, *An introduction to 3D computer vision techniques and algorithms*, Wiley, 2009.

[14] T. Svoboda, D. Martinec, T. Pajdla, "A convenient multi-camera self-calibration for virtual environments", *Presence: Teleoperators and Virtual Environments*, vol. 14, pp. 407-422, Aug. 2005.

[15] M. Domański, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, K. Wegner, "Poznań multiview video test sequences and camera parameters," *ISO/IEC JTC1/SC29/WG11*, Doc. M17050, 2009, Xian.

[16] M. Domański, A. Dziembowski, A. Kuehn, M. Kurc, A. Łuczak, D. Mieloch, J. Siast, O. Stankiewicz, K. Wegner, "Poznan Blocks - a multiview video test sequence and camera parameters for Free Viewpoint Television", *ISO/IEC JTC1/SC29/WG11*, Doc. M32243, Jan. 2014, San Jose CA, USA.

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

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

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

[20] K. Wegner, O. Stankiewicz, M. Domański „Occlusion handling in depth estimation from multiview video", submitted to 3DTV Conference 2014.