MULTI-VIEW VIDEO AND MULTI-CHANNEL AUDIO BROADCASTING SYSTEM

Kwan-Jung Oh[1], *Manbae Kim*[2], *Jae Sam Yoon*[1], *Jongryool Kim*[1], *Ilkwon Park*[3], *Seungwon Lee*[3], *Cheon Lee*[1], *Jin Heo*[1], *Sang-Beom Lee*[1], *Pil-Kyu Park*[1], *Sang-Tae Na*[1], *Myung-Han Hyun*[1], *JongWon Kim*[1], *Hyeran Byun*[3], *Hong Kook Kim*[1], *and Yo-Sung Ho*[1]

[1]Gwangju Institute of Science and Technology (GIST)
[2]Kangwon National University, [3]Yonsei University
Republic of Korea

ABSTRACT

In recent years, various multimedia services have become available and the demand for realistic multimedia systems is growing rapidly. The multi-view video and multi-channel audio are expected to satisfy the user demand for realistic multimedia services. In this paper, we present a new broadcasting system incorporating multi-view video and multi-channel audio over IPTV and MPEG-21 DIA. The proposed system includes data acquisition, camera calibration, data encoding and decoding, transmission, intermediate view reconstruction, and multi-view display and multi-channel audio play. In this paper, we discuss the main features of multi-view video and multi-channel audio.

Index Terms—broadcasting system, multi-view video, multi-channel audio, IPTV

1. INTRODUCTION

With the rapid development of video and audio processing technologies, multi-view video and multi-channel audio are recently attracting extensive interests. Following the current technology trend, various 3D stereoscopic and multi-view video processing systems have been proposed.

In 1994, the RACE DISTIMA project [1] carried out a real-time transmission experiment of stereoscopic video via ATM at 10 Mbps to link the laboratories of KPN Research in Leidschendam, the Netherlands, and the laboratories of Deutsche Telekom in Berlin, Germany. In the ATTEST project, they also developed a 2D-compatible multi-view 3DTV system for broadcasting applications [2]. In order to deliver natural 3D viewing experiences, the ATTEST project proposed an efficient approach to generate 3D contents using a depth camera.

In addition, Kimata et al. [3] proposed a free-viewpoint video communication system using a multi-view video compression. A free-viewpoint viewer generates a natural view from any arbitrary viewing positions and directions. Hur et al. [4] developed an experimental testbed for the 3DTV broadcasting system that are compatible with the HDTV broadcasting infrastructure such as terrestrial and satellite DS-3 networks. Vetro *et al.* [5] developed a multi-view transmission system where each encoded view sequence is transmitted independently and displayed in the stereoscopic monitor. Yang et al. [6] proposed a multi-view system focusing on system components. Lou *et al.* [7] presented a system architecture for real-time capturing, processing, and interactive delivery of multi-view video. They employed 33 cameras to capture multi-view video.

Recently, a number of 3D multi-view systems have been proposed [2-7]. However, in most cases, the integration of the proposed multi-view systems with the audio signal has not been considered. Advances in digital audio technology have made high-quality multi-channel audio. While the CD format specifies only two channels, multi-channel audio provides end users with much more involving experiences. Since multi-channel audio is essential for the realistic broadcasting system, we can include the widely adopted 5.1 channel configuration by placing three loudspeakers in the front of a listener and two in the rear side [8].

In this paper, we present a multi-view video and multi-channel audio broadcasting system which is currently under development in our research center. The main features of our system includes: (1) unlike the previous multi-view systems, multi-channel audio is integrated, and it is expected to support video with increased realism, (2) our system is designed for the IPTV network; therefore, interaction between server and client is easily implemented, (3) MPEG-21 DIA (digital item adaptation) plays an important role in server-client communication so that our system can fit into international standard. Finally, the key part of the multi-view system is the 3D display device. Our system can easily be adjusted to different types of 3D monitors. With the proposed system, users can enjoy the added realism supported by multi-view video and multi-channel audio though our system still has some problems in real-time processing and synchronization.

2. SYSTEM ARCHITECTURE

The structure of the proposed multi-view video and multi-channel audio broadcasting system is shown in Fig. 1. As shown, the proposed broadcasting system mainly consists of ten video cameras, five microphones, ten acquisition PCs, a transport server, clients, a DIA server, video display devices, and audio play environment. These components can be classified into four modules: acquisition/encoder part, transport part, client part, and display part.

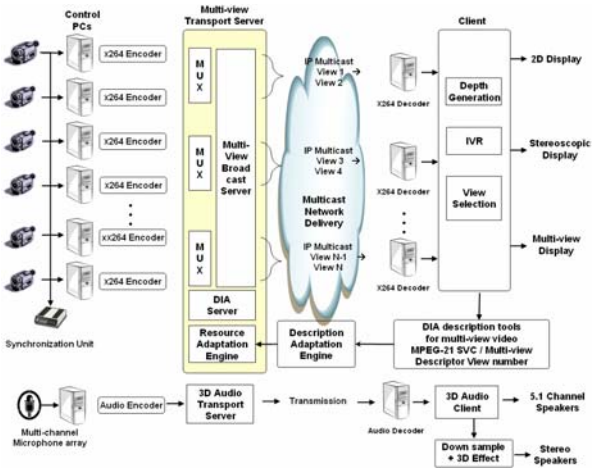

Fig. 1. Overall structure of the proposed system

The acquisition/encoder part is composed of ten cameras, a pan-tilt frame, five microphones, ten control PCs, and a synchronization unit. The cameras are put on a pan-tilt frame that can not only control the height and the distance between the cameras, but also support 1D parallel and arc camera arrangements. The data acquired from cameras and microphones are encoded by H.264/AVC and AAC (advanced audio coding), respectively.

The transport part transports the compressed multi-view video stream and multi-channel audio streams to clients. Additionally, ten multi-view video streams are multiplexed into five streams. In addition, MPEG-21 DIA server deals with the user preference information such as view selection, 3D depth from the back channel. Video and audio servers act independently.

The client part receives the compressed multi-view video and multi-channel audio streams from the transport server. The compressed streams are decoded and played with the 3D monitor and multi-channel speakers. A client can send the user preference data to the DIA server if necessary.

The display part renders the multi-view video and multi-channel audio. The former can be displayed by 2D, stereoscopic monitors or 3D multi-view monitors. The latter is played by 5.1 channel speakers or stereo speakers. In addition, the multi-channel audio can be down-sampled and added with 3D effects for stereo speakers.

3. DESCRIPTION OF MODULES

This section describes some important modules of the broadcasting system in details. They are data acquisition, data encoding and decoding, data transmission, camera calibration, depth map generation, intermediate view reconstruction, and 3D display.

3.1. Data Acquisition

In the proposed system, the multi-view video and multi-channel audio are captured independently. The multi-view videos are captured by ten cameras. Each control PC captures one view and the synchronization unit controls the synchronization of multi-view cameras. Our system supports two types of camera arrangements such as 1D parallel and arc. The camera model is FLEA-HICOL-CS. The image resolution is 1024x768 in pixels and the frame rate is 30 fps. The focal length of a camera lens is 6～20 mm. Figure 2 shows the multi-view camera configuration that is used in our system. The cameras are set in 1-D parallel and the baseline between two neighboring cameras is 20 cm.

Fig. 2. Multi-view camera configuration (GIST, Korea)

The multi-channel audio is acquired by a multi-channel microphone array [9]. The audio recording system uses five microphones so that a total sound field can be acquired for the 5.1 channel loudspeakers. The LFE (low frequency effects) signal is extracted from the low frequency band of five microphone signals.

3.2. Data Encoding and Decoding

Each of the captured multi-view videos is compressed by the H.264/AVC encoder in a simulcast approach. Although real-time hardware encoders are available, the proposed system employs an optimized x264 codec that can encode and decode a video in real-time. Each control PC employs a x264 encoder that compresses captured video and produces

NAL units. Similarly, the x264 decoder decodes the encoded bitstream in real-time. Based upon the processing performance of the system, we assign ten PCs for encoding and five PCs for decoding. In other words, bitstreams of two video sequences are processed at the same PC.

The multi-channel audio acquired by a microphone array is compressed with AAC that is one of the audio compression formats defined by the MPEG-2 standard. The bitrate is 320 kbps.

3.3. Data Transmission

The transport server handles the multiplexing of the synchronized video streams. To provide multi-view video services to clients, we multiplex the video streams together and transport the video streams over a separate IP multicast channel. The multicasting efficiently transports the packets from one or more servers to a group of clients. Then, the multiplexed video streams are simultaneously transported over different multicast channels. At each client, we can select a multicast channel according to a desired view of the end user by joining the multicast channel selectively.

In order to minimize the jitter effect of the transmission network, we multiplex streams in the MPEG-2 TS (transport stream) level. The MPEG-2 TS consists of the sequence of fixed sized TS packets. Each packet has 188 bytes, including 184 bytes for payload and a 4-byte header. One of the items in this header is 13-bit PID (packet identifier) which plays a key role in the operation of the MPEG-2 TS.

Streams are separately read through ten different channels, and then ten video streams are multiplexed into five streams by assigning the different PID of MPEG-2 TS. Then, the server adds the assigned PID of MPEG-2 TS packet and its description into the PMT (program map table) of the MPEG-2 TS. In order to make one time line program, we need to synchronize video streams. Since we obtain synchronous state among video streams by using the synchronization unit, one program is developed by multiplexing two MPEG-2 TSs into one. Finally, the transport server transports five multiplexed streams over different multicast channels.

The multi-channel audio bitstream is transmitted by RTP (real time protocol). The RTP provides end-to-end network transport functions suitable for real-time data such as audio and video, over unicast or multicast network services.

3.4. Camera Calibration

The camera calibration is necessary for the intermediate view reconstruction. The intrinsic and extrinsic camera calibration parameters are estimated from the pattern images for multi-view cameras. Each camera has its intrinsic and extrinsic parameters. Rotation and translation matrices are external parameters that define the position and orientation of the camera.

These camera parameters are used to generate a depth map and subsequently intermediate frames. If we know the relationship between the positions of world coordinate and the 2D image coordinate, we can obtain elements of the matrix. A 2D point is denoted by $m=[u, v]^T$. A 3D point is denoted by $M=[X, Y, Z]^T$. By adding 1 to previous equations, we can obtain augmented vectors like $m'=[u, v, 1]^T$ and $M'=[X, Y, Z, 1]^T$. The relationship between M' and its image projection m' is given by $sm'=K[R|t]M'$, where s is an arbitrary scale factor, $[R|\ t]$ is the extrinsic parameter, and K is the intrinsic parameter.

We employ an open Matlab toolbox provided by [10]. For ten cameras, we use ten pattern images respectively. We choose the fourth camera as the reference camera. The size of the rectangle on the pattern board is 250 mm. The results of the calibration are quite reliable, although they have small estimation errors.

3.5. Depth Map Generation

BP (belief propagation) has been applied successfully to stereo matching algorithms. BP is an iterative inference algorithm that propagates messages in the network [11]. Despite of the good approximation for energy minimization problems, BP approach still requires several minutes for solving stereo problems on the present desktop computer [12]. Since our system is targeting for the real-time processing, we apply the modified BP algorithm that provides less computational cost in generating depth data of multi-view video. The visual cues such as segmentation and edges can be incorporated into the intensity constraint to improve stereo matching. Segmented regions are generated by mean shift segmentation. It is one of the fast and reliable segmentation algorithms.

To solve a depth boundary and occlusion regions, we assign each segmented region to each depth region by the minimization of the cost function considering a relation with neighboring segments. Therefore, we can generate a more reliable depth image from the modified BP algorithm that is incorporated with the segmentation algorithm.

3.6. Intermediate View Reconstruction

In general, there are two major problems in multi-view camera configuration. The first problem is the reliability of disparity. The disparity originated from multiple cameras is often too large. If there is an excessive disparity, it will exhaust viewer's eyes. The other problem happens when view is changed. When users change their views, if the base line of cameras is large, the flickering will then occur on 3D display like a sudden scene change. It also causes a visual discomfort to viewer's eyes.

To solve these problems, we generate intermediate frames. An intermediate frame is an image captured from a virtual

camera between multi-view cameras. Thus, we can provide a high quality image that reduces visual discomfort of the viewer by adjusting the disparity.

A depth image is required to generate intermediate images at given viewpoints. Due to the computational complexity, the depth acquisition is processed off-line. However, intermediate view images are reconstructed in real-time from acquired depth images.

3.7. 3D Display

The multi-view video can be displayed at several types of display as illustrated in Fig. 1. In 2D display, users can select arbitrary viewpoints. Stereoscopic display device can display arbitrary stereo views based on a selected viewpoint. For multi-view (e.g., 9-view) 3D monitor, 9 views can be chosen and displayed to different viewers. Here, some format conversion is needed. Figure 3 shows stereo images displayed on a 9-view 3D monitor.

Fig. 3. Stereo images viewed at different viewpoints

The audio play environment is a 5.1 channel loudspeaker system. In the environment, decoded six audio signals are played by six loudspeakers. In addition, down-mixing is implemented for mono/stereo compatibility. When the multi-channel signal is down-mixed for a stereo signal, HRTF (head related transfer function) is utilized to provide a 3D audio effect by the support of spatial cues of the multi-channel sound. In case of down-mixing into a mono signal, six signals are simply summed without any spatial cues.

4. CONCLUSION

In this paper, we presented a multi-view video and multi-channel audio broadcasting system which can generate more realistic multimedia. The proposed system targets for the system which processes from acquisition to display of multi-view video and playback of multi-channel audio in real time. System components of the proposed system are designed to operate over IPTV network and MPEG-21 multimedia framework. In addition, MPEG-21 DIA plays an important role in server-client communication so that we expect the proposed system can fit into international standards. Based on the developed multi-view video and multi-channel audio system, we plan to integrate the MVC (multi-view video coding) codec and implement the synchronization between the video and audio.

ACKNOWLEDGMENT

This work was supported by the MIC, Korea, under the ITRC support program supervised by the IITA and in part by MOE through the BK21 project.

REFERENCES

[1] A. Smolic and P. Kauff, “Interactive 3D video representation and coding technologies,” *Proceedings of the IEEE, Spatial Issue on Advances in Video Coding and Delivery*, Vol. 93, pp. 99-110, 2005.

[2] A. Redert, M. Beeck, C. Fehn, W. IJsselsteijn, M. Pollefeys, L. Gool, E. Ofek, I. Sexton and P. Surman, “Advanced three-dimensional television system technologies,” Proc. 3D Data Processing Visualization and Transmission, 2002.

[3] H. Kimata, M. Kitahara, K. Kamikura and Y. Yashima, “Free-viewpoint video communication using multi-view video coding,” NTT Technical Review, 2004.

[4] A. Vetro, W. Matusik, H. Pfister and J. Xin, “Coding approaches for end-to-end 3D TV systems,” Picture Coding Symposium, 2005.

[5] N. Hur, G. Lee, W. Yoo, J. Lee and C. Ahn, “An HDTV-Compatible 3DTV Broadcasting System,” *ETRI Journal*, Vol. 26, No. 2, April 2004.

[6] Z. Yang, B. Yu, K. Nahrstdet and R. Bajscy, “A multi-stream adaptation framework for bandwidth management in 3D tele-immersion,” Network and Operating System Support for Digital Audio and Video (NOSSDAV), Newport, Rhode Island, May 2006.

[7] J. G. Lou, H. Cai and K. Li, “A real-time interactive multi-view video system,” Multimedia '05: Proceedings of the 13th annual ACM international conference on multimedia, pp. 161-170, ACM Press, 2005.

[8] M. Bosi, “High Quality Multichannel Audio Coding: Trends and Challenges,” *Proc. 16th AES International Conference*, pp. 393-400, March 1999.

[9] M. Williams and G. Le Dti, “The Reference Model Architecture for MPEG Spatial Audio Coding,” *Proc. 108th AES convention*, Paris, France, Preprint 5157, Feb. 2000.

[10] Camera Calibration Toolbox for Matlab at: http://www.vision.caltech.edu/bouguetj/calib_doc/

[11] S. Jian, Z. Nan-Ning and S. Heung-Yeung, “Stereo matching using belief propagation,” *IEEE Trans. On Pattern Analysis and Machine Intelligence*, Vol. 25. No. 7, July 2003.

[12] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient belief propagation for early vision,” *IEEE Proceedings of Computer Vision and Pattern Recognition*, Vol. 1, pp. 261-268, July 2004.