7G-4

Multipoint Sound and Image Generation Using Large Microphone and Camera Array

Mehrdad Panahpour Tehrani[1], Yasushi Hirano[1], Toshiaki Fujii[2], Shoji Kajita[1], Kazuya Takeda[3], and Kenji Mase[1]

[1]Information Technology Center, Nagoya University
[2]Graduate School of Engineering, Nagoya University
[3]Graduate School of Information Science, Nagoya University

ABSTRACT

The advent of digital sensors, and the ability to create views and sounds that combine information from a number of sensed images and sounds, is changing the way we think about sensor network featuring 3D audio visual performance. In this paper, we describe an array of 100 cameras and Microphones that we have built, and we summarize our experiences to generate free listening-point and free viewpoint.

1. INTRODUCTION

This research is aim to represent 3D sound and Image without localization and propose to use ray-space representation of light rays for sound wave, which is independent of object's specifications, for arbitrary listening-point generation in 3D space, as it shown in Fig. 1.

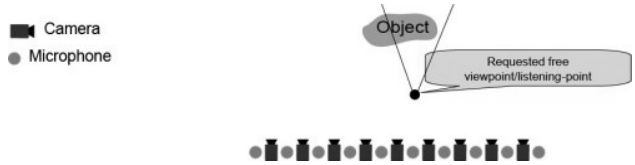

Fig 1: Arbitrary sound and image generation

Sound can be recorded, computed and replayed by directed speakers, using the well-known sound processing methods, efficiently. Several approaches tried to generate arbitrary listening-point generation of sound; however there are few effective model such as Head Related Transfer Function (HRTF) [1] and representation of the sound sources in 3D space to have an efficient processing. Meanwhile, images are rendered by computer graphics algorithms and have become more attractive and more efficient and image synthesis hardware has come to existence, such as Free viewpoint TV (FTV) [2]. The free viewpoint systems should have a correct correspondence of sound and images in an arbitrary viewpoint. Therefore, a representation method of sound sources in 3D space using computer graphics and image processing techniques is necessary. Many representation methods have already been proposed. These methods are categorized into image based rendering (IBR) [3], model based rendering (MBR) [4] methods and their combinations [5] Ray-space representation method is an IBR method and independent of object specifications. See Fig. 2.

Our goal is to integrate the 3D information of sound and images featuring multipoint generation capabilities that would be able to produce corresponded sound and images of any requested virtual point by user, where there is not any existing camera and microphone.

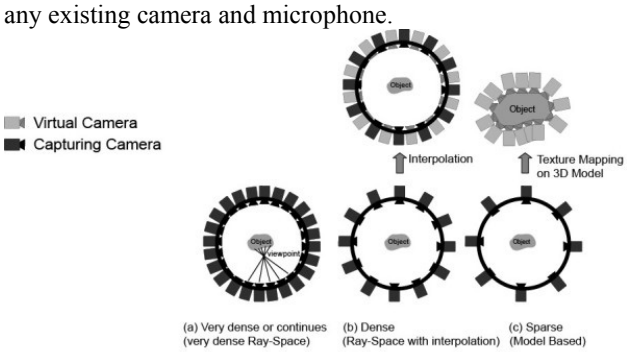

Fig. 2: Camera Interval and representation methods

2. HARDWARE

The system hardware is as following:

We used cameras (PULNiX TMC-1400CL, 1392x1040x1-Bayer Matrix, 29.411fps), and microphones (Sony ECM-77B, 16 Bits, 96~8 KS/sec). Each pair of camera and microphone is connected to a general purpose PC (Celeron 2GHz, 256 RAM, OS: Linux), and each 10 nodes are connected to a server (Xeon 3.60GHz Dual, OS: Windows). The nodes and server are connected with Gigabit Ethernet. The capturing and recording of video and audio are synchronized by triggering the nodes in a same time, with synchronizers which are receiving the GPS signal. The wired synchronizer clock is 1usec, were as the wireless is 1msec. Therefore, the developed system can contain nodes in different location and still senses the images and sound data being synchronized.

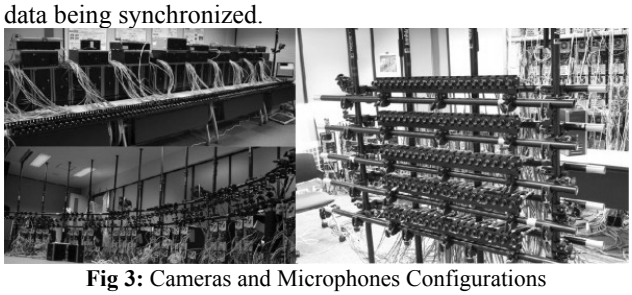

Fig 3: Cameras and Microphones Configurations

The cameras and microphones are set in three configurations, line, arc and 2D, as shown in Fig. 3.

3. MULTIPOINT SOUND AND IMAGE

Free view point generation in dense configuration of cameras can be done using ray-space method, and geometry compensation to generate a dense configuration, and consequently synthesizing any virtual viewpoint.

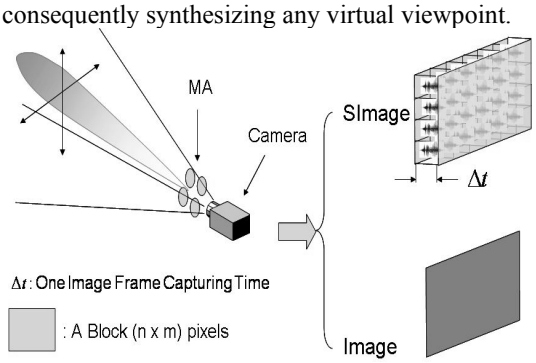

Fig. 4: Capturing Image and SImage with a camera and a MA

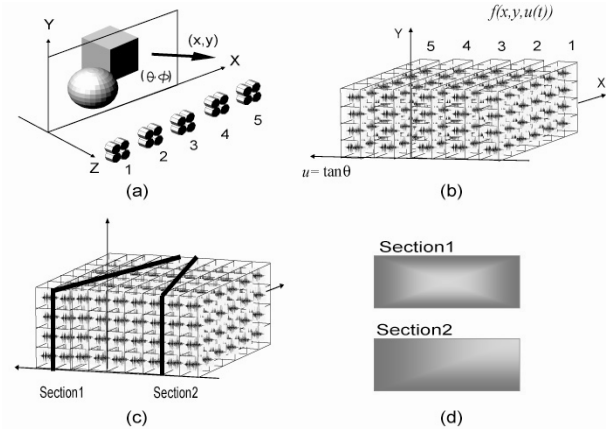

Fig. 5. Ray-space method of each frequency layer of an SImages (a) Ray space recording (b) Recorded ray-space (c) SImage generation (d) Generated SImage

In this research specifically we have proposed to represent the 3D sound wave in large microphone system using sub-band processing and ray-space method [6]. Fig. 4 and Fig. 5 show the sound signal capturing in image format using microphone array (MA), called sound image (SImage), and their representation in ray-space format, respectively. The sub-band processing is used to represent the SImages in multi-frequency layered, consequently their ray-space. Hence, we can integrate the 3D sound and image information, and generate any virtual view and listening points.

The geometry compensation is done using the corresponding images in the location of each MA or SImages or their combination, for each frequency layer. Sound of an SImage is generated by averaging the sound wave in each block. After combining the generated sound for certain duration of time in frequency domain and inverse transformation to time domain, the desired sound is generated. The sound of an SImage is generated by averaging all pixels or blocks of sound in the SImage.

Experimental results using SImages for arbitrary listening-point generation is done for three frequency layers. The experimental setup has 3 MAs on an arc (r=2.9m). Each array has 3 microphones with 183mm distance. Distance between each MA is 183mm. The sound source has 2.3m distance from the center of MAs line. Three SImages with the size of 6×1 are captured in f=5KHz, ranging 60-degree view, and the middle SImage is compared to the original one, and SNR of 22.86dB is obtained. The generated sound by virtual SImage has SNR equal to 10.36dB in comparison with the original sound in that location.

Note that installing a camera and a MA in same location is hardly possible. Nevertheless, close alignment of a camera and a MA will give quite good result due to static character of sound waves in space.

4. CONCLUSION

We assumed that processing large numbers of images and sound would eventually be easy using efficient integration method, and multi terminal signal processing and communication, since each node has computational power and communicate with other nodes.

The applications we have explored include next generation of 3D-TV or Free viewpoint TV (FTV), audio/visual sensor networks featuring 3D signal processing and integration, used in security, education, and hospitable systems.

This work has been supported by the SCOPE Fund project, Ministry of Internal Affairs and Communication, Japan (ref. No.: 041306003).

5. REFERENCES

[1] F.L Wightman, D.J. Kistler, “A Model of HRTFs Based on Principal Component Analysis and Minimum-phase Reconstruction”, Journal of the Acoustical Society of America, Vol. 91, No. 3, pp. 1637- 1647, 1992.

[2] P. Na Bangchang, T. Fujii, M. Tanimoto, “Experimental System of free viewpoint television”, Proc. SPIE, Vol. 5006-66, Santa Clara, CA, USA, pp. 554-563, 2003.

[3] T. Fujii, T. Kimoto, M. Tanimoto, “A new flexible acquisition system of ray-space data for arbitrary objects”, IEEE Transaction On Circuit and Systems for Video Technology, Vol. 10, No. 2, pp. 218-224, 2000.

[4] T. Matsuyama, X. Wu, T. Takai, T. Wada, “Real-time dynamic 3-D object shape reconstruction and high-fidelity texture mapping for 3-D video”, IEEE Trans. on Circuits and Systems for Video Technology, Vol. 14, No. 3, 2004.

[5] W.C. Chen, J.Y Bouguet, M.H. Chu, R. Grzeszczuk, “Light Field Mapping: Efficient Representation and Hardware Rendering of Surface Light Fields” ACM Trans. on Graphics, Vol. 21, No. 3, pp. 447-456, 2002.

[6] M.P. Tehrani, Y. Hirano, T. Fujii, S. Kajita, K. Takeda, K. Mase, “The Sub-band Sound Wave Ray-Space Representation”, IWAIT 2006, pp. 291-296, Japan, Jan 2006.