ARBITRARY LISTENING-POINT GENERATION USING SUB-BAND REPRESENTATION OF SOUND WAVE RAY-SPACE

Mehrdad Panahpour Tehrani[(1)], Yasushi Hirano[(1)], Toshiaki Fujii[(2)], Shoji Kajita[(3)], Kazuya Takeda[(4)], and Kenji Mase[(3)]

[(1,3)]Information Technology Center, Nagoya University
[(2)]Graduate School of Engineering, Nagoya University
[(4)]Graduate School of Information Science, Nagoya University
[(1)]{mehrdad, hirano}@itc.nagoya-u.ac.jp
[(2)]fujii@nuee.nagoya-u.ac.jp
[(3,4)]{kajita, kazuya.takeda, mase}@nagoya-u.jp

ABSTRACT

This paper proposes arbitrary listening-point generation of sound without source localization, and a theory based on the ray-space representation of light rays using sub-band signal processing. An array of beam-formed dynamic microphone-arrays (MAs), are set and each MA generates a multi frequency layer sound-image (SImage) by scanning the viewing range of a camera. Each layer is captured by active microphones in MA for a given frequency range which makes the correct interval to capture that frequency layer, and a sound wave ray-space. Arbitrary listening-point generation of each layer is done by geometry compensation of corresponding images in the location of each MA or SImages or their combination. Each layer sound of an SIamge is generated by averaging the sound wave in each pixel or group of pixel. The listening-point is generated after combining each layer sound in frequency domain and inverse transformation from frequency to time domain.

1. INTRODUCTION

So far, several approaches tried to generate arbitrary listening-point 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]. However the corresponding sound in the virtual location of camera cannot be generated without sound source localization. Many representation methods have already been proposed for 3D images. These methods are categorized into image based rendering (IBR) [3],[4], model based rendering (MBR) [5] methods and their combinations [6]. Ray-space representation method is an IBR method and independent of object specifications. Based on the ray-space representation of light rays, we propose a theory to represent the 3D sound wave. In [7], we have proposed a theory to represent 3D sound wave based on ray-space method for a single frequency sound source. However, sound signal are in a frequency range. To expand the theory in [7] to a practical case, in this paper we propose to use the sub-band signal processing of the sound wave [8]. In other words, we represent each SImage in different frequency layers.

Captured multi-layered SImages generate multi layered sound wave ray-space and will be dense by geometry compensation [9],[10] of corresponding camera viewing images or SImages or their combination, for each frequency layer. Hence, any virtual SImage, which corresponds to an arbitrary listening-point can be generated, and it perfectly corresponds to its virtual viewpoint image. The corresponded sound in each frequency layer is generated by averaging the sound wave in each pixel or group of pixel. The generated sound of all layers for certain duration of time are transferred to frequency domain and combined, then the inverse transform is done, and the desired listening-point can be generated.

2. MULTI FREQUENCY LAYERED SIMAGE

Sound is 1D signal in time domain. In order to process sound signal like image, a 2D×1D signal of sound wave is generated in space and time domains. We capture a sound wave with duration of a frame for each pixel or group of pixel in different frequency ranges. After capturing the sound wave for all location of an image, we can generate the multi frequency layered SImage.

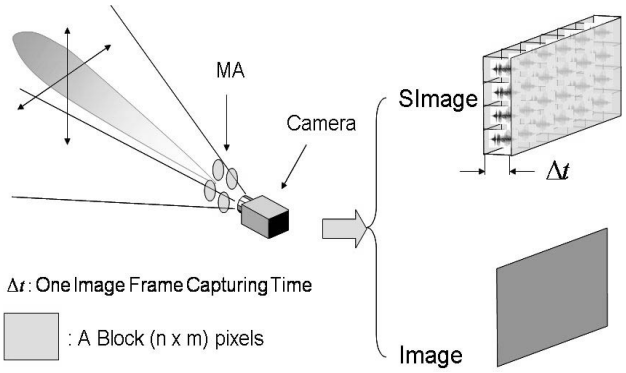

Fig. 1. Capturing an Image and an SImage with a camera and a MA

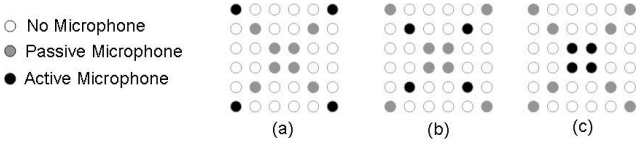

Fig. 2. Dynamic microphone array configuration (a) Low frequency range (b) Middle frequency range (c) High frequency range

The interval among microphones, to capture the sound waves in different frequency ranges using the microphone array is proportional to the capturing frequency range. Therefore, each SImage is generated by using an array of dynamic microphone (MA) in multi frequency layers, as shown in Fig. 1. Each layer is captured by the active microphones in the MA for a given frequency range which makes the correct interval among microphones to capture that frequency layer (i.e. Fig. 2). Each layer of an SImage has the same size of an image and contains of blocks of sound wave with duration of one image-frame.

In order to generate an SImage, the light equivalent of the sound is determined for every pixel or pixel-group in each layer of SImage. The sound wave for duration of a frame is captured, which is proportional to a pixel location. To generate a multi frequency layered SImage, we have to scan the viewing range of an image corresponded to the SImage, which is done by beam-forming of the active microphone in a dynamic MA. Note that the resolution of the SImage depends on the scanning accuracy or the beam width of the MA. The larger number of microphone in an array with same microphone interval, the higher resolution of SImage can be obtained. By placing the microphones in two dimensions, we can generate 2D×1D SImage.

Fig. 2 shows a dynamic microphone array configuration to capture a multi frequency layered SImage with 3 layers. As it is shown in Fig. 2(a), the active microphones are set apart to capture a low frequency range of an SImage. For the higher frequency range of SImage, the closer microphones are activated to capture, as it shown in Fig. 2(b) and Fig.2(c). The bigger the dynamic microphone array can generate more frequency layers of an SImage. Note that all microphones are listening in a same time and generate the multi frequency layered SImage.

3. SUB-BAND SOUND WAVE RAY-SPACE

Ray-space was originally proposed as a common data format for 3D image communication [3]. A similar idea has been proposed in computer graphics field for generating photo-realistic images into computer generated virtual world [11]. It is one of the "image-based-rendering techniques", and is called Lumigraph [4]. It has been widely used to create photo-realistic virtual worlds. Both of them are based on the idea that a view image from an arbitrary viewpoint can be generated from a collection of real view images. Ray-space method describes 3D spatial information as the information of the ray, which transmits in the space.

Due to the similarity of light rays and sound wave rays, we proposed the sound wave representation based on ray-space [7]. Fig. 3 shows an example of the definition of ray-space in a frequency layer of SImage. The same ray-space is generated for all frequency layer of an SImage. The parameterization of all ray-space layers are the same. Let (x,y,z) be three space coordinates, and (θ,φ) be the parameters of direction. x and y are the intersection of the sound ray with XY-plane. θ and φ are the angles of the sound rays passing through XY-plane with Z-axis in horizontal and vertical directions, respectively as shown in Fig 3(a). In free space, a sound ray going through the space is uniquely designated by the intersection (x,y) with plane $z=0$ and its direction (θ,φ). These sound ray parameters construct a 4D space. In fact, this is a 4D subspace of 5D ray-space (x,y,z,θ,φ) of SImages. In this sound ray-parameter, we define a function f whose value corresponds to an intensity of the specified sound ray in a given time (t). Thus, all the intensity data of rays can be expressed by equation (1). This ray parameter is called the ray-space.

$$f(x,y,\theta,\varphi)\ ,\ -\pi\le\theta<\pi \text{ and } -\pi/2\le\varphi<\pi/2 \qquad (1)$$

The sound rays that pass through the specific plane can be captured using a MA, as it shown in section 2. When we set a MA at (x_0,y_0,z_0), the intensity data of sound rays are given by equation (2).

$$f(x,y,z,\theta,\varphi),\quad x=x_0,y=y_0,z=z_0 \qquad (2)$$

For simplicity, we consider only 2D subspace $f(x,\theta)$ of 5D SImage ray-space, which the vertical parallax (φ) and vertical position (y) are neglected. We call this 2D plane of the ray-space data as "Epipolar Plane SImage (EPSI)", which is the cut of ray-space data of Fig 3(b), parallel to xu-plane for a given y. In another word, EPSI is the intersection of epiploar plane with the corresponded camera plane in stereovision in the location of MA, which is well-known in computer graphic field as Epipolar Plane Image (EPI). Therefore, if there is an array of MAs in which all have a common epipolar plane, the intersection of the epipolar

plane with all MA planes can be shown in one SImage, which is EPSI. Let *X*, *Z* be real-space coordinates, and *x*, *u* be the ray-space coordinates as shown in Fig 3(b). The sound rays that pass through a point *(X,Z)* in the real-space form a line in the ray-space is given by equation (3).

$$X = x + uZ, \quad u = \tan\theta \tag{3}$$

Note that symbol *u* is a function of time. Therefore, the ray-space data of SImages is *f(x,y,u(t))*. It gives the most important and interesting characteristic of the 4D ray-space representation that a view SImage corresponds to a cross section of the ray-space data.

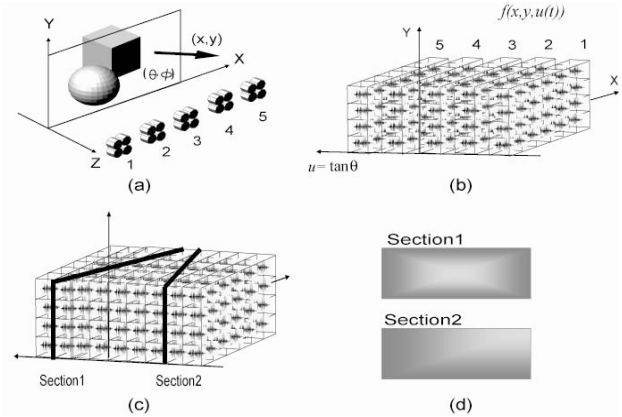

Fig. 3. Ray-space method of each frequency layer of an SImages (a) Ray-space recording (b) Recorded ray-space (c) SImage generation (d) Generated SImage

Therefore, the acquisition/display process of each frequency layer of an SImage can be considered as the recording/extracting process of the ray-space data along the locus. In the acquisition process, we sample the sound ray data in the real space and record them as the ray-space data along the locus as shown in Fig 3(c). In the display process, we cut the ray-space along the locus and extract the section SImage of the ray-space as shown in Fig 3(d). However, in the case of not-dense ray-space data, interpolation is needed. It generates the missing sound rays between two EPSI lines. The EPSI lines are in *(x)* direction for a given *u* as 1D subspace of 5D ray-space data. The interpolation task should be done on 2D×1D SImage ray-space data or EPSI. The interpolation generates an EPSI line between two EPSI lines. In the next section the ray-space interpolation of SImage is explained under EPSI constraint.

4. ARBITRARY LISTENING-POINT GENERATION

The ray-space interpolation technique of images is based on the adaptive filtering interpolation [9],[10] as shown in Fig. 4. This technique was designed to replace a simple linear interpolation technique, which is suitable for input images with viewpoints parallel to the same plane. The main advantage of this method is the approximate geometric models are not required prior any interpolation. Moreover, the scene complexity does not have any effect on interpolation speed.

As interpolation is done for images, we would like to use the same scheme for each layer of an SImages. To generate middle EPSI line, we have to look for the corresponded pixels or blocks of sound wave between two EPSI lines. The best match or the best disparity map can be obtained by each of the following ways.

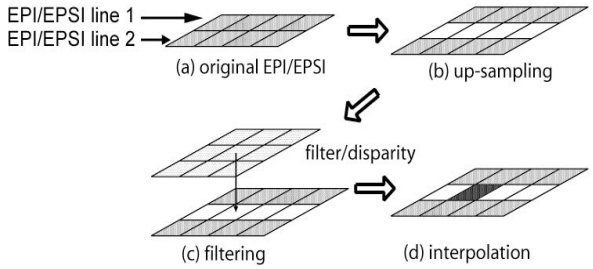

Fig. 4. Interpolation under EPI/EPSI constraint

(a) Performing block matching scheme on each frequency layer of an SImages or EPIS lines, as it is done for images or EPI lines, and generating a disparity map using SImages. In this method, after finding the disparity map, interpolation can be performed by averaging the found corresponded pixels or blocks of sound wave. Note that the best match for a given location is found by a minimization solution of sound wave level difference in two SImages or EPSI, for duration of a frame, in a given maximum disparity.

(b) Using the disparity map obtained by the images captured by the cameras in the location of MA. In this method, the intermediate SImage or EPSI is generated by averaging the found corresponded pixel or block of sound waves. The corresponded location in two SImages or EPSI lines are found by the disparity map of the captured images or EPI lines by cameras in the same location of MAs. After having a virtual SImage (Fig 3(d)) of each frequency layer, the sound corresponded sound is generated by averaging the sound wave in each pixel or group of pixel. The generated sound of all layers for certain duration of time are transferred to frequency domain and combined, then the inverse transform is applied to the combined frequency domain signal, and the desired listening-point is generated.

5. EXPERIMENT

Experimental results using SImages for arbitrary listening-point generation is shown in the following for three layers of an SImage. The experimental setup has 3 MAs on an arc (r=2.9m). Each array has 3 microphones with 183mm distance. Distance between MAs is 183mm. The sound source has 2.3m distance from the middle MA. Three SImages in three frequency layers and 2KHz bandwidth at center frequencies of f=5KHz, f=3KHz, f=1KHz are captured with the size of 6×1, 4×1, and 1×1, respectively.

Through the comparison of the middle SImage of each layer with the original one captured by the middle MA, SNR of 29.06dB, 28.78, and 26.97 are obtained, and the generated sounds in each layer by synthesized SImages have SNR equal to 16.32dB, 14.95dB, 13.64 in comparison with the original sound generated by SImage of the middle MA, respectively. Note that the experiment on multi frequency layered SImages has been done using the same interval of microphones in MA for all layers so that the resolution of SImages are different in different layers, however for the same resolution in all layers dynamic MA can be used.

6. DISCUSSION

There are two challenging issues for 3D sound wave representation and rendering methods.

The first one is the sampling rate. Maximum allowable disparity or sampling rate of each frequency layer of an SImage should be investigated as it has been done in plenoptic sampling theorem [12] for multi-view images. Due to static characteristic of the sound wave for small changes in space, the sampling rate or the interval of MAs to capture SImages is larger than that of images.

The second one is the disparity used for interpolation. In section 4 two methods are proposed. Method (a) is easy to use and in the case of just generating free listening-point, the system does not need to be equipped by cameras, however the accuracy of the generation is lower than method (b). On the other hand, method (b) can perform better than method (a). In addition, the free viewpoint/listening-point generation of video/audio can be synchronized. However, the method (b) is more complex. 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.

7. CONCLUSION

This paper proposed a theory to represent the 3D sound field using ray-space method and sub-band processing of SImages, and introduced an algorithm to generate free listening-point. The proposed methods are independent of sound sources location. In addition, the proposed theory can solve the problem of 3D media integration. The proposed methods are currently being developed on a practical system. In our future research, we will work on sampling theory for capturing SImages, and will perform an efficient integration of 3D audio/video.

8. ACKNOWLEDGEMENT

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

9. 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, Santa Clara, CA, USA, Vol. 5006-66, 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] M. Levoy, P. Hanrahan, “Light field rendering” ACM SIGGRAPH ’96, pp. 31-42 , 1996.

[5] 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.

[6] 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.

[7] M.P. Tehrani, Y. Hirano, T. Fujii, S. Kajita, S. Takeda, M. Tanimoto, K. Mase, “The Sound Wave Ray-Space”, IEEE International Conference on Multimedia and Expo, ICME 2005, Amsterdam, The Netherlands, 2005.

[8] W.H. Neo, B. Farhang-Boroujeny, “Robust Microphone Arrays Using Subband Adaptive Filters”, IEE Proc. Visual Image Signal Processing, vol. 149, no. 1, pp. 17-25, 2002.

[9] M. Droese, T. Fujii, M. Tanimoto, “Ray-Space Interpolation Constraining Smooth Disparities Based on Loopy Belief Propagation”, Proc. of 11th International Workshop on Systems, Signals and Image Processing ambient multimedia (IWSSIP), Poland, pp. 247-250, 2004.

[10] M.P. Tehrani, T. Fujii, M. Tanimoto, “Offset Block Matching of Multi-view Images for Ray-Space Interpolation” The journal of the institute of Image information and Television Engineers –ITE, Vol. 58, No. 4, pp. 86-94 , 2004.

[11] R. Szeliski, “Video mosaics for virtual environments”, IEEE Computer Graphic Application, Vol. 16, pp. 22-30, 1996.

[12] J.X. Chai, S.C. Chan, H.Y Shum, X. Tong, “Plenoptic Sampling”, ACM SIGGRAPH, pp. 307-318, 2000.