A Next Generation Video Streaming System for Intuitive Remote Interaction

Kaoru Sugita Nobuhiro Nakamura Shinichi Baba Masao Yokota
Fukuoka Institute of Technology
sugita@fit.ac.jp, {mgm05008, mgm05011}@ws.ipc.fit.ac.jp, yokoya@fit.ac.jp

Abstract

In this paper, we propose a live video streaming system based on virtual reality technologies for intuitive interaction among people remotely located. This system is one kind of sever-client system and can provide remote users with virtual 3D audiovisual fields in real time via a very high-speed network. The server captures audio and video data from its clients, compiles them into one 3D audiovisual scene at a virtual conference and broadcasts it over the clients. At the present stage, our system captures 2 videos and creates one 3D video at a time. Our system can play 3D audiovisual contents on Windows XP systems as well as on CAVE systems.

1. Introduction

In recent years, live video streaming systems and TV conference systems are very popular [1, 2, 3]. We have also developed a WWW conference system in order to provide WWW browsers with a live video communication facility [4] and evaluated its performance [5, 6]. However, these systems are only to display flat pictures and thereby the users hardly feel talking face-to-face with the other participants at the same place.

On the other hand, there have been proposed several 3D model reconstruction methods [7, 8]. These are applicable to videos as well as to photographs and can be implemented on virtual reality systems such as CAVE system [9] and, as a further extension, can provide the remote users with face-to-face realities.

This paper proposes a VR live video streaming system in order to facilitate intuitive interactions among remote users and shows its implementation and evaluation based on the experimental results.

2. VR live video streaming system

Our VR live video streaming system can provide the remote users with such a communication facility as shown in Fig.1 where 3D audiovisual fields are available on Windows XP systems as well as on CAVE systems. The participants can allocate their images arbitrarily in the video. The sound field, so called, Virtual Sound Field (VSF) is created so as to match the arrangement of the images. Therefore, each participant can communicate with the other users as if they were actually talking face-to-face at the same place.

Our proposed system is one kind of sever-client system as shown in Fig.2 and can provide remote users with virtual 3D audiovisual fields in real time via a very high-speed network. The server captures audio and video data from its clients, compiles them into one 3D audiovisual scene at a virtual conference and broadcasts it over the clients.

The server consists of 5 components as follows:
(S1) User Interface - Provide windows to display 2D videos and to set up video capture parameters
(S2) Video Capturer - Take in 2D video and audio data for 3D contents
(S3) 3D Video Creator - Extract frames, detect parallaxes and create polygons with textures
(S4) 3D Video Stream Controller - Broadcast polygons and textures controlling their numbers, frame sizes and rates
(S5) Sound Stream Controller – Broadcast sounds controlling the sampling rates

Each client consists of 7 components as follows:
(C1) User Interface - Provides windows to display 3D videos and to set up 3D video parameters
(C2) 3D Video Allocator - Arrange 3D videos in the virtual world
(C3) 3D Video Controller - Play 3D video with sounds synchronized
(C4) 3D Video Stream Controller - Receive polygons and textures controlling their numbers, frame sizes and rates
(C5) Virtual Sound Field Creator - Play sounds reflecting the arrangement of the 3D videos
(C6) Sound Controller - Play sounds with 3D videos synchronized
(C7) Sound Stream Controller - Receive sounds controlling the sampling rates

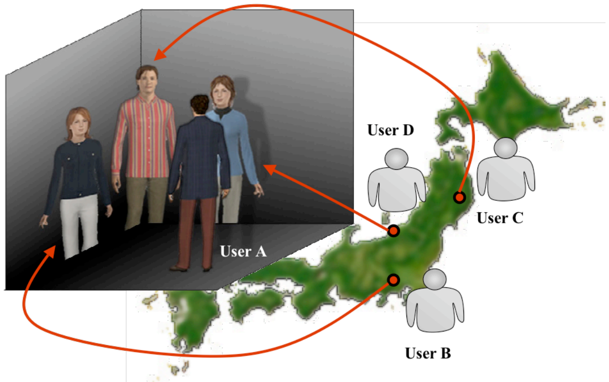


Fig.1. VR live video streaming system.

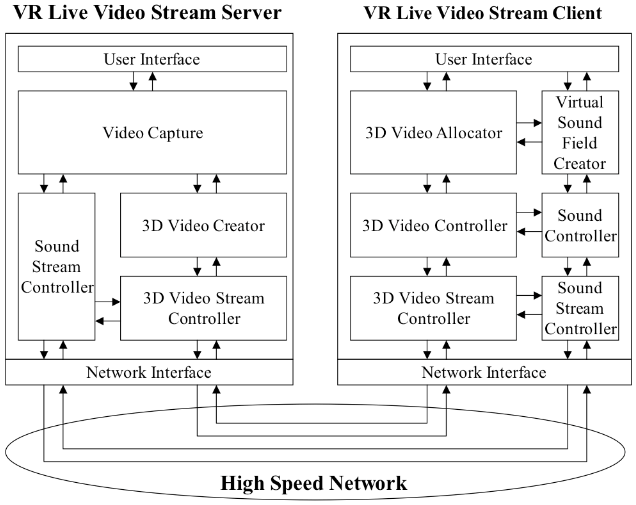


Fig.2. System architecture.

3. 3D video creation method

Figure 3 shows spatial relations of an object and its 2D video frames. As known well, the object's position (*O*) can be computed from the points (*P1* and *P2*) projected on the two different frames according to (1)-(3), where $O(x_0,y_0,z_0)$, $P1(x_1,y_1,z_1)$, and $P2(x_2,y_2,z_2)$, respectively.

$$x_0=(D/2d)(x_1+x_2) \quad (1)$$

$$y_0=(D/d)y_1=(D/d)y_2 \quad (2)$$

$$z_0=(D/d)f \quad (3)$$

A 3D model is composed of polygons each of which consists of vertexes with 3D coordinates. Such a model can be reconstructed by corresponding the 3D coordinates specifying the object's surface with the 2D coordinates in the video frames. A 3D video supplies a time-sequenced set of polygons and texture images. The details of 3D model creation are shown in Fig.4.

In order for modeling in real time, we have developed a new method to control the number of polygons for parallax detection. In this method, the parallax between the frames is calculated efficiently using one set of candidate pairs of corresponding pixels so called 'Template for Candidate Projection Area (TCPA)' as shown in Fig.5. Usually such a candidate pair cannot be determined uniquely due to the ambiguous z coordinate of the object while it is prepared in advance by (4)-(7). Therefore, the pair minimizing the difference between TCPAs is selected as the most certain one.

$$x_1 = f\left(\frac{2x_0 + D}{2z_0}\right) \quad (4)$$

$$y_1 = f\left(\frac{y_0}{z_0}\right) \quad (5)$$

$$x_2 = f\left(\frac{2x_0 - D}{2z_0}\right) \quad (6)$$

$$y_2 = f\left(\frac{y_0}{z_0}\right) \quad (7)$$

4. Virtual sound field

The virtual 3D sound field is intended to play sound reflecting the relation between the 3D video display point and the viewpoint in the virtual world. As shown in Fig.6, the sound source and the sound receivers are located according to the 3D video. In order to realize the vertical stereo effect, each receiver is divided logically into 2 sub-receivers and each speaker plays the same sound twice reflecting the difference in distance and in vertical angle between the sub-receivers. The volume and delay time at the receiver is given by (8) and (9).

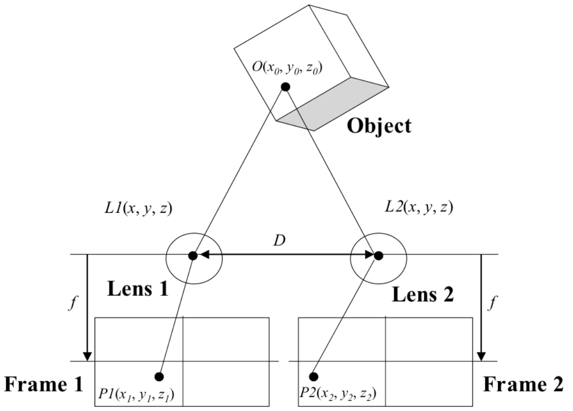


Fig.3. Spatial relations of object and frames.

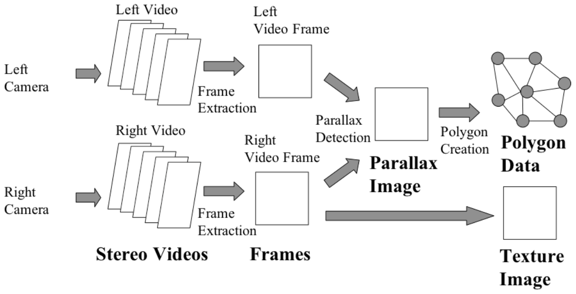


Fig.4. 3D model creation.

$$V_r = f_v(V_{ss}, L_{svx}) \tag{8}$$

$$D_r = \frac{f_d(L_{svx})}{c} \tag{9}$$

where

L_{svx}: the distance between the sound source and the sound receiver,

V_{ss}: the volume at the source,

V_r: the volume at the receiver,

$f_v(V_{ss}, x)$: the volume at the distance x from the source,

D_r: the delay time at the receiver,

c: the speed of sound, and

$f_d(x)$: the delay time at the distance x.

5. Implementation

The server was implemented on Windows XP system while each client was implemented on either Windows XP system or CAVE system, employing the C Language, Intel Open Computer Vision Library (Open CV) [10], WinSock2, Open GL, GLUT and CAVE Library. The CAVE library was used to draw/display 3D videos and to support the head tracking on CAVE system only. The others were utilized for both the systems in the conventional ways except that the Open CV was for capturing videos and extracting video frames.

Figure 7 shows the videos from the left camera (a) and the right camera (b) and the 3D video (c) composed on the server. Logicool Qcam for Notebooks Pros were employed as web cameras put at a space 25cm long enough for us to recognize easily the parallax between Fig.8 (a) and (b). Figure 8-c was viewed at a client on Windows XP System, where the user could control the viewpoint and view angle through the keyboard.

6. Conclusion

The VR live video streaming system and its implementation were described. At the present stage, our system captures 2 videos and creates one 3D video at a time. Our system can play 3D audiovisual contents on Windows XP systems as well as on CAVE systems while the existing live video streaming systems and TV conference system are only to display the flat pictures. Currently, we are evaluating our system in its precision and performance and also implementing the sound modules. In the future, we will introduce some highly precise certainty calculation method.

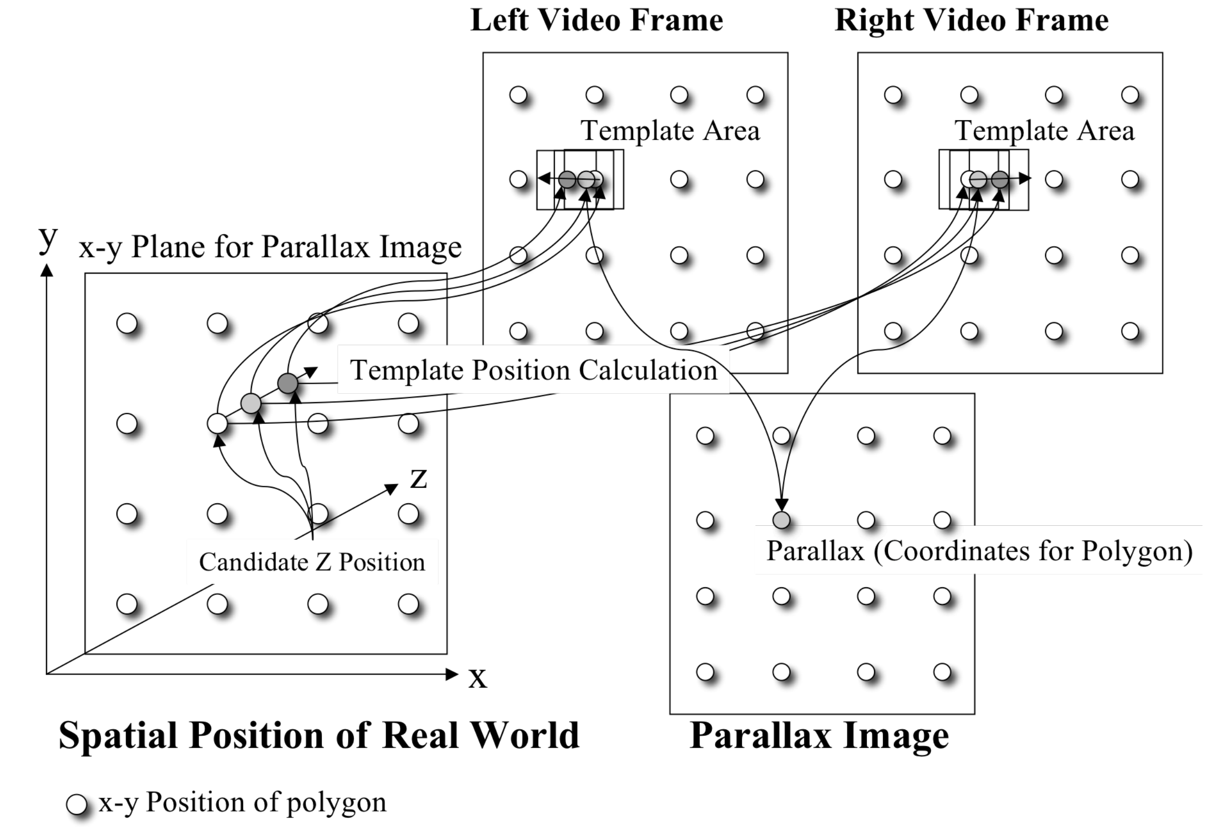

Left Video Frame
Right Video Frame
Template Area
Template Area
y
x-y Plane for Parallax Image
Template Position Calculation
z
Candidate Z Position
x
Parallax (Coordinates for Polygon)
Spatial Position of Real World
Parallax Image
○ x-y Position of polygon

Fig.5. Parallax Detection.

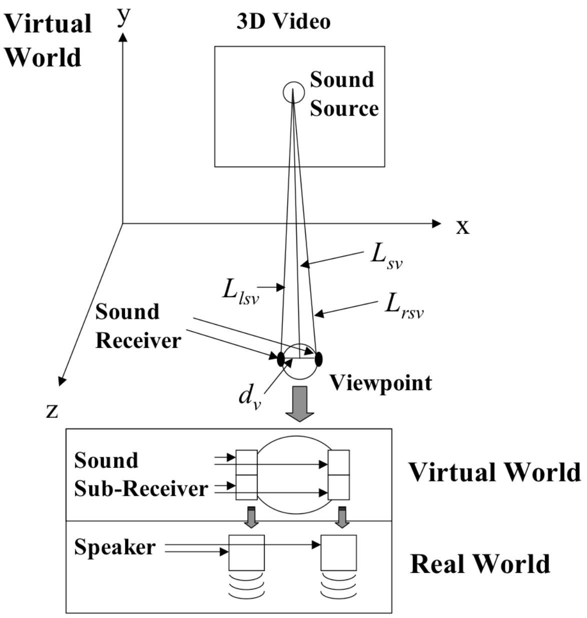
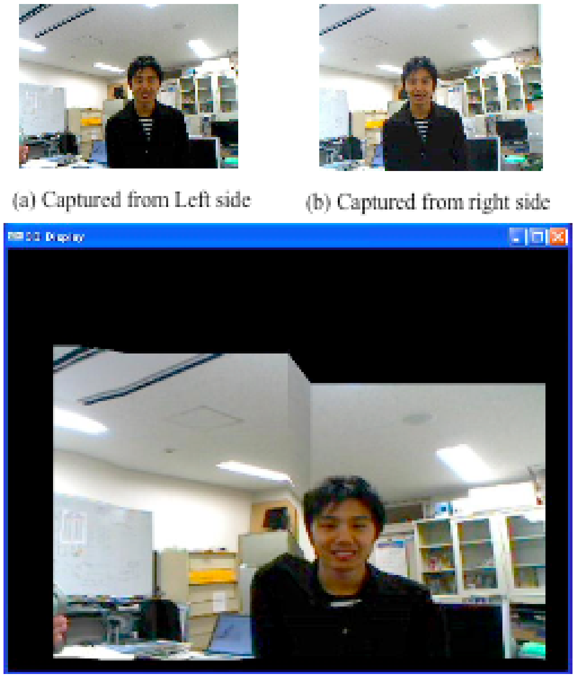


Fig.6. Virtual Sound Field.



(a) Captured from Left side

(b) Captured from right side

(c) Composite 3D Video

Fig.7. Implemented System.

References

[1] D.Nishantha, Y.Hayashida and T.Hayashi: "Application Level Rate Adaptive Motion-JPEG Transmission for Medical Collaboration Systems", Proc. of the 6th International Workshop on Multimedia Network Systems and Applications (MNSA-2004), pp. 64-69, 2004.

[2] Y.Maita, K.Hashimoto, Y.Shibata: "A New TV Conference System with Flexible Middleware for Omni-directional Camera", Proc. of the 7th International Workshop on Network-Based Information System (NBIS-2004), pp. 84-88, 2004.

[3] Y. Kato, D. Jiang and K. Hakozaki: "A Proposal of a Streaming Video System in Best-Effort Networks Using Adaptive QoS Control Rules", Proc. of the 18th International Conference on Advanced Information Networking and Applications (AINA-2004), Vol.2, pp. 54-57, 2004.

[4] Kaoru Sugita, Noriki Uchida, Akihiro Miyakawa, Leonard Barolli: "Implementation of WWW Conference System for Supporting Remote Mental Health Care Education", Proc of the Seventh International Workshop on Multimedia Network Systems and Applications (MNSA2005), pp.686-692, June.2005.

[5] Kaoru Sugita, Noriki Uchida, Akihiro Miyakawa, Giuseppe De Marco, Leonard Barolli: "Performance Evaluation of WWW Conference System for Supporting Remote Mental Health Care Education", Proc. of the 11th International Conference on Parallel and Distributed Systems (ICPADS-2005), Vol.1, pp.271-277, July.2005.

[6] S.Baba, K.Sugita, N.Uchida, G. De Marco, L.Barolli and A.Durrusi: "Performance Evaluation of WWW-based Conference System ", Proc. of the 8th International Workshop on Network-Based Information System (NBIS2005), pp.50-56, Aug.2005.

[7] Yin Li, et al.: "Stereo Reconstruction from Multiperspective Panoramas ", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 26, No. 1, pp.45-62, 2004.

[8] James Davis, et al.: "Spacetime Stereo: A Unifying Framework for Depth from Triangulation ", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 2, pp. 296-302, 2005.

[9] Cruz-Neira, C., Sandin, D.J., and DeFanti, T.A.: "Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE. ", Proc of SIGGRAPH '93 Computer Graphics Conference, pp. 135-142, August 1993.

[10] Intel Open Source Computer Vision Library, http://www.intel.com/technology/computing/opencv/index.htm