Usage of Video Avatar Technology for Immersive Communication

Tetsuro Ogi[1, 2, 3], Toshio Yamada[1], Yuji Kurita[1], Yoichi Hattori[1], Michitaka Hirose[2]
[1]*Telecommunications Advancement Organization of Japan*
[2]*The University of Tokyo*
[3]*Mitsubishi Research Institute, Inc.*
tetsu@iml.u-tokyo.ac.jp

Abstract

Representation of the realistic human image has been required in the virtual reality applications due to the improvement of the high presence virtual reality display systems. In this study, various kinds of video avatar techniques that integrate the person's live video image into the virtual world were developed. In order to use the effective video avatar methods in the virtual reality applications, it is desirable to choose the suitable geometric model and the suitable capture method according to the purpose and the system environment. This paper also discusses the video avatar studio and the video avatar server technologies that are being developed in order to use the various kinds of video avatar techniques properly.

1. Introduction

Recently, immersive projection display such as the CAVE [1] or the CABIN [2] has become very popular, and high presence virtual worlds are generated in it. In these systems, representation of the realistic human image is required. In addition, since several immersive projection environments have been connected through the broadband networks, the realistic human image is also required as a high presence communication tool between remote places [3].

In order to meet such a demand, video avatar technology has been studied. The video avatar is a technique to represent a high presence human image by integrating the live video image of the human into the three-dimensional virtual world. In order to generate a video avatar, various kinds of methods to make a geometric model or to segment a human image, etc. can be used. The authors have developed the various video avatar techniques in the MVL (Multimedia Virtual Laboratory) project aiming to construct a virtual laboratory on the broadband network.

This paper discusses the features and the usage of various video avatar techniques that were developed in this project. Moreover, we will explain the design and the implementation of the video avatar studio and the video avatar server that are being developed to use the video avatar techniques more effectively.

2. MVL (Multimedia Virtual Laboratory)

2.1. MVL environment

MVL is the concept of the virtual laboratory constructed on the broadband network, and it is being promoted by the Ministry of Public Management, Home Affairs, Posts and Telecommunications. Figure 1 illustrates the concept of the MVL. In this figure, a scientist, an engineer and a designer are discussing with each other, while sharing the design model and the experimental data that exist in the remote places. Namely, the MVL aims at realizing high presence collaborative work among remote people sharing some information.

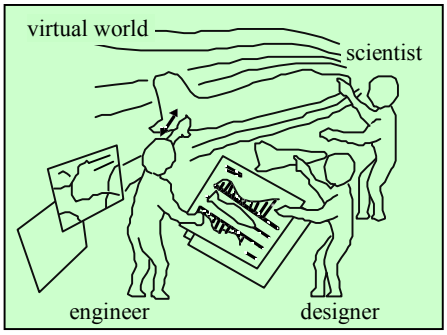

Figure 1. Concept of multimedia virtual laboratory

In order to realize this concept, a networked research environment was constructed by connecting several immersive projection environments, such as the CABIN, the COSMOS [4], the multi-screen workbench and the real work places, through the JGN (Japan Gigabit Network) network of the TAO (Telecommunications Advancement Organization of Japan).

CABIN and COSMOS are CAVE-like multi-screen displays developed at the University of Tokyo and the

Gifu Techno-plaza respectively. These displays can represent the immersive virtual worlds by projecting the synchronized stereo images on the surrounding screens. The multi-screen workbench is a desk side work space that consists of three screens. In addition, the real work place is used as a mixed reality environment by installing the transparent multi-screen system and projecting the stereo images of the virtual world and the communication partner on it. Figure 2 illustrates the networked research environment of the MVL.

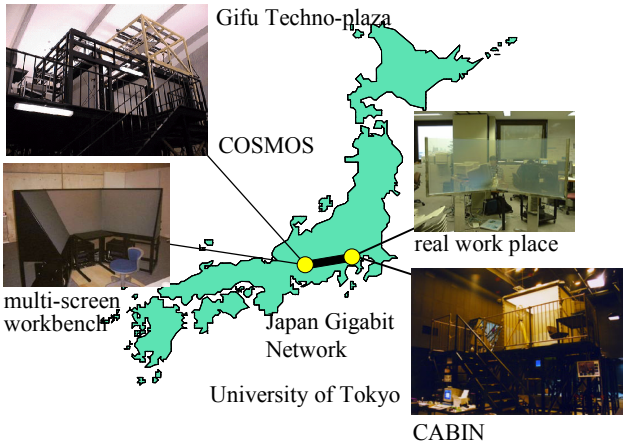

Figure 2. Networked research environment of MVL

2.2. Video avatar

In the MVL project, the video avatar technology that utilized the live video image of the user in the shared virtual world was developed. In order to generate a video avatar, various methods can be considered. Figure 3 shows the basic process of generating the video avatar.

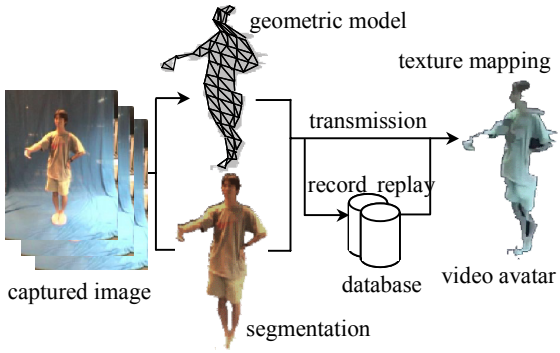

Figure 3. Basic process of generating video avatar

In this process, a user's image is first captured by the video camera, and only the user's figure is segmented from the background as well as the geometric model of the user is created from the captured image. The geometric model and the segmented user's image are transmitted through the network or are stored in the database system as video avatar data. At the opposite site, these data is received or retrieved from the database, and the video avatar is generated by texture mapping the segmented user's video image onto the geometric model. In particular, the video avatar can be used as a communication tool between remote places by performing these processes at both sites.

In order to generate a high presence video avatar, it is a technical subject how an accurate geometric model is created or how a clear image of the user is segmented in the above-mentioned process. In the following chapter, the features of the various video avatar techniques that have been developed in the MVL project are discussed from the viewpoints of creating a geometric model and of segmenting a person's image.

3. Video avatar techniques

3.1. Geometric model

Although the human's body has a three-dimensional shape, the video image itself is two-dimensional. Therefore, in generating a video avatar, it is an important problem to create a geometric model that has three-dimensional information.

3.1.1. Plane model. The simplest geometric model that can be used to generate a video avatar is a two-dimensional plane model [5]. Since this method integrates the video image as a rectangular billboard into the virtual world, it cannot express the three-dimensional person's gesture such as moving hand in front of the body.

Therefore, in this study, the camera switching method was introduced to add the three-dimensional information to the plane model. In this method, multiple cameras are placed surrounding the person to capture the person's image from various directions, and the selected camera is changed according to the movement of the viewer's viewpoint as shown in Figure 4. Although the video image captured by one camera is two-dimensional, the effect of motion parallax can be utilized by switching the video images captured from various directions.

Since this method needs to place the multi-viewpoint camera system around the person, it cannot be used in the CABIN or the COSMOS. But when the video avatar data is recorded beforehand to replay them in the applications, the multi-camera system can be used in the studio. In this study, the multiple video cameras were placed surrounding the person at intervals of 10 degrees or 20 degrees in the blue background room. The video images captured by all cameras were stored in the database, and

the nearest frame to the user's viewpoint was retrieved and it was integrated in the virtual world. Although the plane model itself has only two-dimensional information, it can be used effectively in various demonstration systems because it represents a clear image of the person in the virtual world. Figure 5 shows the example of the plane model video avatar that is used to navigate the virtual world in the demonstration program.

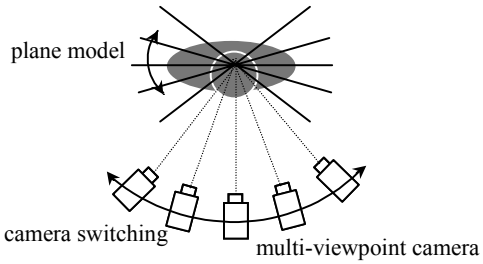

Figure 4. Plane model using camera switching

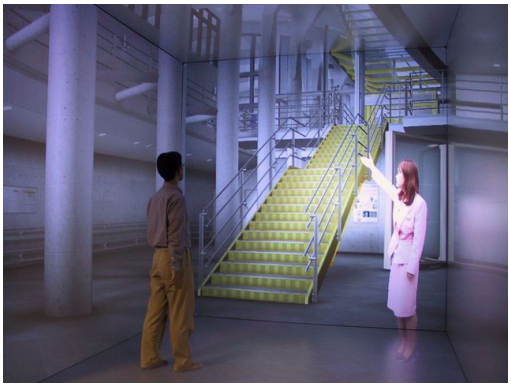

Figure 5. Plane model video avatar used for the navigation

3.1.2. Depth model. In order to generate a geometric model that has effects of the binocular parallax and the motion parallax, the stereo matching algorithm can be applied. In this method, the distance value for each pixel is calculated by computing the gaps between the corresponding points in the stereo images captured by the stereo camera. By texture mapping the video image on the polygon surface model of the person that is created using the distance information, a depth model video avatar can be generated. Since this method creates a geometric surface model only for the camera direction, it is called 2.5-dimensional video avatar.

Although the visualized image of the 2.5-dimensional video avatar is well formed when it is seen from the camera direction, it becomes largely distorted when the viewer's position moves away from the camera direction. Therefore, in this study, the camera switching method using the several stereo cameras was introduced so that the well-formed three-dimensional image can be seen from various directions by switching the depth model created by each stereo camera as shown in Figure 6. In the immersive communication system used in the MVL project, the Triclops Color Stereo Vision systems made by Point Grey Research Inc. were placed in the corners of the CABIN and the COSMOS, and they were used to create the 2.5-dimensional depth model video avatar [6].

Since the 2.5-dimensional video avatar can represent the three-dimensional fingertip position using the depth information, the user can effectively discuss with the video avatar while pointing at the design model or the visualized data in the shared virtual world. Figure 7 shows the example in which the user is talking with the depth model video avatar in the scientific visualization application.

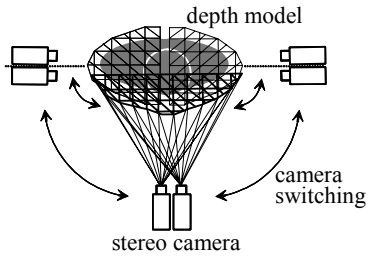

Figure 6. Depth model using camera switching

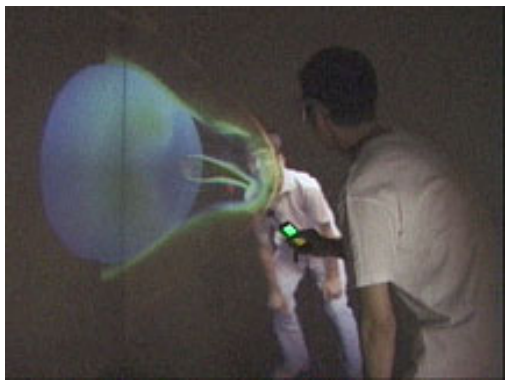

Figure 7. Depth model video avatar used for the scientific visualization

3.1.3. Voxel model. In order to compensate the imperfection of the 2.5-dimensional video avatar, it is necessary to generate a complete three-dimensional model from the captured video images. Although several methods have been proposed to create a three-dimensional model from multiple video images such as the multi-baseline stereo matching [7] or the integration of several depth models [8], in this study, the shape from silhouette method was used by taking the processing time

into consideration [9]. Figure 8 illustrates the process of the shape from silhouette method. In this method, the scene space is divided into voxel cubes, and the three-dimensional voexl model of the avatar is created by computing whether each voxel cube is inside or outside the silhouette of the person's image captured by each camera. This voxel model can be easily converted to the geometric polygon model using the Marchine cubes method, and then the three-dimensional video avatar can be created with relatively small calculation load.

Although this model is not so effective for the mutual communication because it needs a studio where the multiple cameras are installed, it is an effective method to record the three-dimensional video avatar data and replay them in the application program. Especially, when the video avatar data is used in the shared virtual world among several sites, the representation using the three-dimensional model is required, because it may be seen from various directions simultaneously. In this study, the voxel model video avatar technique was used to record the person's action in the real work place. Figure 9 shows that the three-dimensional video avatar is used effectively in the psychological experiment on the person's gesture recognition in the shared virtual world.

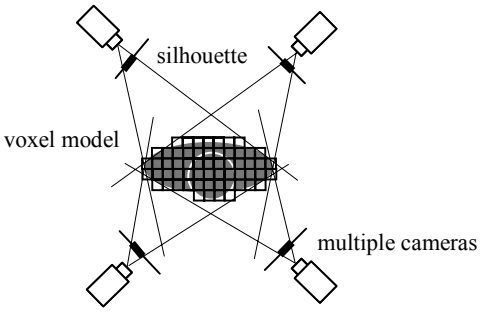

Figure 8. Voxel model using shape from silhouette

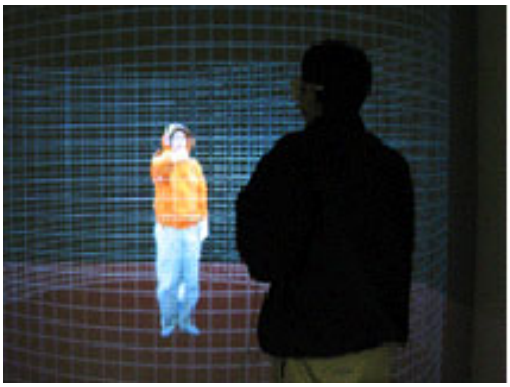

Figure 9: Voxel model video avatar used for the psychological experiment

3.1.4. Face model. In the above-mentioned methods, since the geometric model is created from the captured video images in real-time, the accuracy of the person's geometric model depends on the camera conditions. Therefore, another approach of the offline creation of the accurate geometric model can be considered. In this study, the accurate face model was created beforehand in the offline process, and the video images captured from two directions were texture mapped on it in real-time while tracking the person's head movement using the OPTOTRAK sensor of the Northern Digital Inc. to represent the facial expression of the video avatar. Figure 10 shows the example of the offline face model video avatar used in the virtual conference room.

Although this offline model can be effectively used for the face model, it cannot be used for the person's body because the shape of the body is largely changed according to the person's action. In this example, the part of the body under the neck is created using the depth model, and the three-dimensional face model and the 2.5-dimensional body model are combined. In addition, when the user moves largely in the virtual world, it is difficult to adjust the positions between the offline face model and the real-time video texture. But when the user is sitting on the chair using the workbench or talking to the other people in a short distance in the shared virtual world, this model can be effectively used because the user does not move so largely.

Figure 10. Face model video avatar used in the conference room

Table 1 shows the features of the above-mentioned several geometric models of the video avatar. In this table, each item is evaluated according to a three-grade system. In this table, the subjective items, such as image quality, obstacle to image, calculation load, and system cost, were judged by the relative evaluation among the methods. Although the two-dimensional model is desirable when a high quality person's image is required as a component of the virtual world, the 2.5-dimensional model or the three-dimensional model is more desirable when it is used for

the collaborative work in the three-dimensional space. In the case of using the immersive projection display, the camera switching methods are not desirable because the projected images are disturbed by the multi-camera system. In addition, the available model is also restricted by the calculation load and the system cost. From these results, we can understand that when the video avatar is used in the virtual reality application, the suitable model of the video avatar should be selected and used according to the purpose of the application and the system environment.

Table 1. Features of several video avatar models

	avatar model	binocular parallax	motion parallax	whole body image	image quality	obstacle to image	calculation load	system cost
2D	plane model	×	×	○	○	○	○	○
	switching plane model	×	○	○	○	×	○	△
2.5D	depth model	○	△	○	△	△	○	○
	switching depth model	○	○	○	△	×	○	△
3D	voxel model	○	○	○	△	×	△	×
	face model	○	○	×	○	○	○	○

3.2. Person's image

Next, the segmentation method of the person's image and the camera condition should be considered to make clear texture data of the video avatar.

3.2.1. Segmentation method. In order to segment a person's image from the background, several segmentation methods can be used. The most typical segmentation technique is a chroma-key method [10]. Although this method needs a special room with a blue background, the high quality person's image can be clipped out with small segmentation error. Therefore, in this study, the chroma-key method is used when the offline video avatar data is recorded in the blue-back studio or the blue backdrop can be used in the immersive projection display.

Next, a background image difference method is used as a typical segmentation method without using a blue background. In this method, the background image without the person is captured beforehand, and only the person's image is segmented from the background by comparing the person's image with the background image [11]. This method can also be used as a relatively stable technique with small segmentation error in various environments that include the real work place. Figure 11 shows that the user is communicating with the video avatar that is generated using the background image difference method in the shared real work place. In this example, the remote users are sharing the augmented reality space by projecting the user's image on the transparent screen in the real work place mutually.

However, the background image difference method cannot be used when the user is experiencing the virtual world in the multi-screen immersive projection display because the background image also changes. In this condition, the image segmentation method using a depth-key is often used.

This method clips out the person's image using the threshold value of the distance between the camera position and the user, and it can be used effectively when the depth model video avatar is used. However, since the quality and the error of the segmented person's image depend on the accuracy of the measured depth data, it needs to prepare the good lighting and camera conditions.

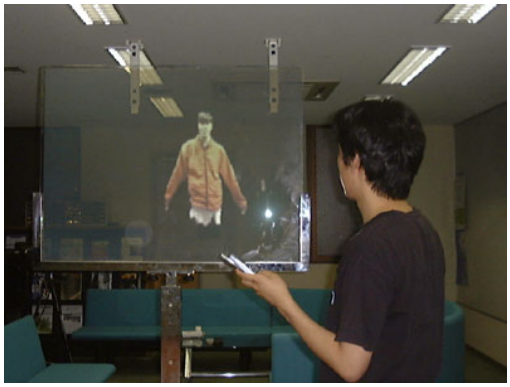

Figure 11. Segmented video avatar displayed in the real work place

3.2.2. Camera condition. In order to create a clear image of the video avatar, the camera condition must be taken into consideration as well as the image segmentation method. In particular, several techniques have been proposed to capture the clear image of the user using the high sensitive and wide field of view camera system in the immersive projection display.

For example, it is necessary that the user's figure is captured in the whole area of the display space even when the user walks around in it. In this study, a camera head control system was developed to capture the image of the moving person [12]. In this method, the user's position was tracked using the electromagnetic sensor, and the camera head was controlled so that it followed the user's movement. Figure 12 illustrates the system configuration of the camera head control system. In this system, the background image is changed when the camera head is rotated. Therefore, the wide area of the background image was captured beforehand, and the selected area of the background image was dynamically changed according to

the camera head movement to segment the person's image using the background image difference method.

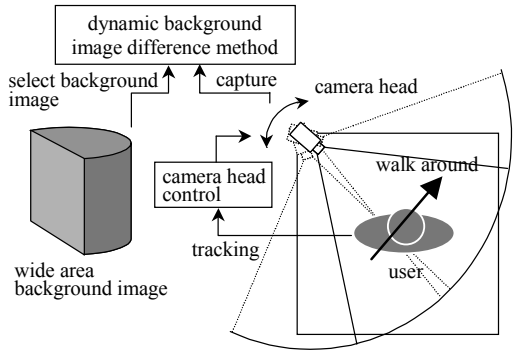

Figure 12. System configuration of camera head control system

Moreover, when the user's image is captured in the immersive projection display such as the CABIN or the COSMOS, the lighting condition is also important. Although a darker room is desirable to project a clear image onto the screen, a brighter lighting condition is required to capture a vivid image of the user. Therefore, in this study, a synchronized strobe light method was developed as shown in Figure 13 [13]. In this system, when the liquid crystal shutter glasses close both eyes, the strobe light is flashed, a blue back image is projected on the back screen and the user's image is captured by the video camera synchronously. Conversely, when the liquid crystal shutter glasses open, the strobe light is stopped and the image of the virtual world is projected on the screen. Therefore, the user's image can be captured in the bright condition using the chroma-key method, as well as the user can experience the immersive virtual world in the dark conditions. Although the synchronized strobe light method has an influence of flickering on the observers, it can generate a clear image of the video avatar in the immersive projection display.

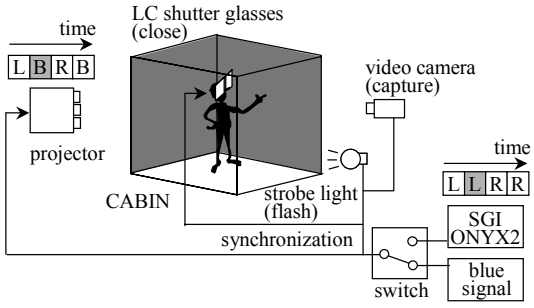

Figure 13. Principle of the synchronized strobe light method

Thus, several video avatar techniques have been developed in the MVL project. But in the application systems, the available methods are restricted by the camera condition and the display environment. Therefore, we can understand that the suitable video avatar technique should be selected and used according to the purpose of the application and the system configuration of the available virtual reality system.

4. Video avatar transmission system

4.1. Video avatar studio

In order to use the suitable video avatar method properly, a video avatar generation system that can create and transmit various kinds of video avatar data is required. The authors are now developing a video avatar studio to meet these requirements. Figure 14 illustrates the system construction of the video avatar studio, and Figure 15 shows the studio room with the multi-camera system.

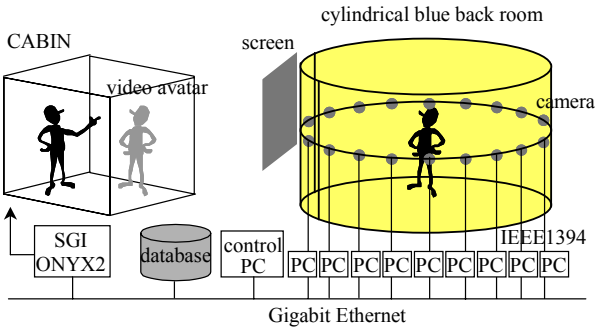

Figure 14. System construction of the video avatar studio

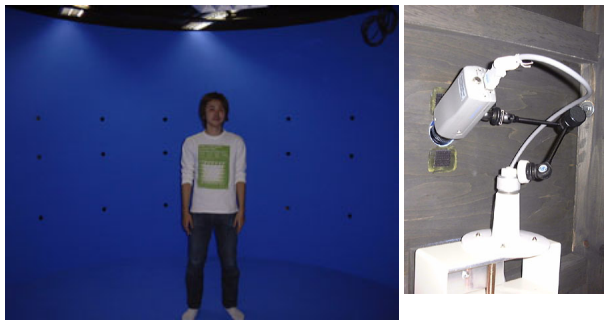

Figure 15. Video avatar studio with multi-camera system

The video avatar studio is a cylindrical room with a diameter of 4,000 mm and a height of 2,570 mm, and eighteen digital CCD cameras (SONY DFW-X700) are embedded at intervals of 20 degrees in the blue back wall.

The heights of the camera positions can be changed among 600 mm, 1,200 mm, and 1,575 mm from the floor. Each camera has 66x88 degrees angle of view using Cosmicar H416 lens, and it can capture XGA (1024x768) resolution images at 15 fps frame rate. These captured images are transmitted to the PCs through the IEEE 1394 connections, and they are used to generate various kinds of video avatar data.

Moreover, the part of the front wall of the cylindrical room can be removed, and a 100 inch screen is equipped outside the front wall. When the offline video avatar data is recorded or the video avatar data is sent one way from the studio to the virtual world, this wall is closed to capture the user's images from all directions using the surrounding cameras. On the other hand, when the video avatar is used for the mutual communication in the shared virtual world, the front wall is removed to display the the partner's figure and the virtual world on the front screen.

Although several studios have been built to generate a specific video avatar data, this system has a feature of generating various kinds of video avatar data by processing images captured by multiple cameras in real-time. For example, in order to generate the two-dimensional plane model video avatar, the selected image is switched among eighteen cameras according to the movement of the user's viewpoint. The 2.5-dimensional depth model video avatar is generated by using the neighboring cameras as a stereo camera module, and the three-dimensional voxel model video avatar is generated by using the all images simultaneously.

In addition, the multiple video cameras can be used in order to increase the time resolution as well as to increase the space resolution. In this system, when the eighteen cameras are used simultaneously to generate the voxel model, the transmission of the video images becomes the bottleneck for the real-time performance. Therefore, in this study, a camera grouping method was developed to increase the update rate of the video avatar model. In this method, every third camera is grouped and each group of the cameras captures the video image of the user with time shifting. Then, the video avatar model can be generated in appearance three times faster than the usual use.

4.2. Video avatar server

Next, the video avatar server is being developed in order to send the video avatar data to the multiple sites. In early research, the video avatar technique had been used for the communication between two sites where the same kinds of systems were equipped. The video avatar server aims at realizing the communication among multiple sites, where different kinds of video avatar models can be used in different ways such as the real-time communication or the record and replay representation. Therefore, this system is designed so that it can distribute the various kinds of video avatar data in efficient way. Figure 16 illustrates the software construction of the video avatar server.

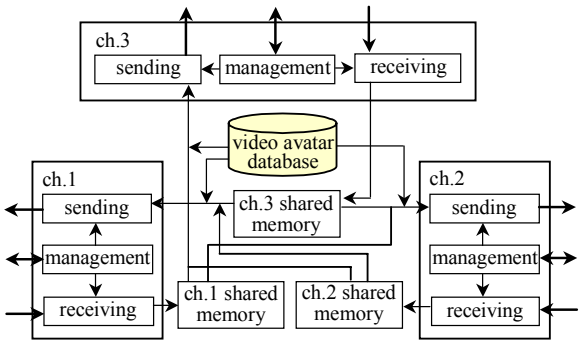

Figure 16. Software construction of video avatar server

In the communication using the video avatar server, each site first sends a demand of connection to the server. At the server site, avatar management process, avatar receiving process and avatar sending process are started, and the information about the other client sites and the available video avatar models are exchanged with each client site. The avatar data is once transmitted from each client to the server, and it is distributed to the other client sites.

This video avatar server can be used in order to not only support the online communication, but also to replay the recorded video avatar data. In this case, at the server site, the avatar sending process retrieves the corresponding video avatar data from the database according to the request from the client, and it transmits the video avatar data to the client sites by controlling the replay speed. Therefore, the clients can receive and replay the video avatar data without considering whether it is an online data or a recorded data.

Figure 17 shows the example of the video avatar communication using the video avatar server among three sites. In this experiment, the CABIN at the Hongo campus of University of Tokyo, the cylindrical screen at the Komaba campus of University of Tokyo, and the COSMOS at the Gifu Techno-plaza were connected through the JGN network, and the virtual showroom was visualized. Even when each user walked through the virtual showroom freely, they were able to discuss with each other while looking at the other users' video avatars in the three-dimensional shared virtual world.

Thus, it is expected that we can use the effective video avatar technique according to the purposes of the applications, by using the video avatar studio and video avatar server technologies.

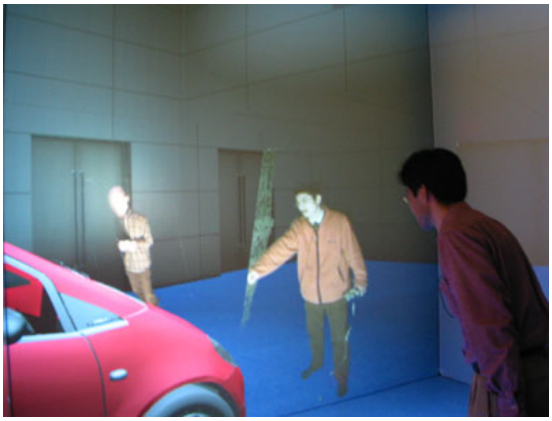

Figure 17. Video avatar communication among three sites

5. Conclusions

In this paper, various kinds of video avatar techniques have been developed to present a realistic human image in the immersive virtual world, and the features of these methods were discussed. In order to create an effective video avatar, it is desirable to choose the suitable geometric model and the suitable segmentation method according to the purpose and the system environment. Therefore, it is also required to build a video avatar generation environment in which various kinds of models and capture methods can be used. Therefore, this paper also discussed the video avatar studio and the video avatar server that are being developed to meet these requirements. Future work will include completing several functions of the video avatar studio and the video avatar server, and evaluating the effectiveness of these techniques.

6. References

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

[2] M. Hirose, T. Ogi, S. Ishiwata, T. Yamada, "Development and Evaluation of the Immersive Multiscreen Display CABIN", Systems and Computers in Japan, Vol.30, No.1, pp.13-22, 1999.

[3] J. Leigh, T.A. DeFanti, A.E. Johnson, M.D. Brown, D.J. Sandin, "Global Tele-Immersion: Better than Being There", ICAT'97, pp.10-17, 1997.

[4] T. Yamada, M. Hirose, M. Iida, "Development of Complete Immersive Display: COSMOS", VSMM98, pp.522-527, 1998.

[5] J. Insley, J. Sandin, T. DeFanti, "Using Video to Create Avatars in Virtual Reality", Visual Proceedings of 1997SIGGRAPH, pp.128, 1997.

[6] T. Ogi, T. Yamada, K. Tamagawa, M. Kano, M. Hirose, "Immersive Telecommunication Using Stereo Video Avatar", IEEE VR2001, pp.45-51, 2001.

[7] T. Kanade, P.W. Rander, "Virtualized Reality: Being Mobile in a Visual Scene", ICAT/VRST'95, pp.133-142, 1995.

[8] M.M. Sein, Y. Suzuki, H. Kakeya, Y. Arakawa, "Generating the 3D Model of an Object for Realizing the 3D Shape-shared Communication over the Network", IEEE KMN2002, 2002.

[9] S. Moezzi, A. Katkere, D.Y. Kuramura, R. Jain, "Immersive Video", VRAIS'96, pp.17-24, 1996.

[10] F. Hasenbrink, V. Lalioti, "Towards Immersive Telepresence SCHLOSSTAG'97", IPT98, 1998.

[11] V. Rajan, S. Subramanian, D. Keenan, A. Johnson, D. Sandin, T. DeFanti, "A Realistic Video Avatar System for Networked Virtual Environments", IPT2002, 2002.

[12] K. Hirose, T. Yamada, T. Ogi, K. Hirota, M. Hirose, "A Video Avatar in Wide Viewing Field by Camera Tracking Method", Vol.3, No.2, pp.33-36, 2001 (Japanese)

[13] M. Hirose, T. Ogi, M. Kanou, T. Yamada, "Synchronized Chroma-key Method for Communication between Immersive Projection Environment", Vol.2, No.2, pp.49-52, 2000 (Japanese).