ICAT 2005
December 5-8, Christchurch, New Zealand

Real World Video Avatar : Real-time and Real-size Transmission and Presentation of Human Figure

Tomohiro Tanikawa[2)] **Yasuhiro Suzuki**[3)] **Koichi Hirota**[3)] **Michitaka Hirose**[3)]
Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo, Japan
{*tani,thzuki,hirota,hirose*}*@cyber.rcast.u-tokyo.ac.jp*

Abstract

The presentation of a human figure is an important topic of research with respect to the application of virtual reality technology to support communication and collaboration. In this paper, an approach to transmitting and presenting the figure of a person at a remote location in real time, an implementation of a prototype system based on this approach, and the evaluation of the prototype system using the approach are described. In our approach, images of a person are captured from all around using multiple cameras, transmitted through a network, and displayed on a revolving flat panel display that is capable of presenting different images according to the orientation of the viewing position; the revolving display presents each image so that it is visible exclusively at the orientation from which it was taken, and consequently, an image of the person that can be viewed from various directions is realized. Through the implementation of the prototype system and experiments, it was confirmed that the proposed approach is feasible and that the prototype system functions effectively.

Keywords: Video avatar, Real World, Real-time, Real-size

1. Introduction

In recent years, research on remote communication has been conducted in various fields such as virtual reality and CSCW (computer-supported cooperative work). In remote communication, the presentation of the human figure is one of the important issues; by presenting the figure of a person at a remote location with photoreality, it is expected that participants will become able to communicate in a manner similar to face-to-face communication in the real world, including such nonverbal interactions as pointing, gestures, facial expressions, and eye contact.

Video avatar[6] is one methodology of interaction with people at a remote location. By using such video-based real-time human figures, participants can interact using nonverbal information such as gestures and eye contact. In traditional video avatar interaction, however, participants can interact only in "virtual" space. If the avatar can be presented in the real world, it will become able to support communication in daily life rather than for specialized purposes in a virtual environment; such an avatar is called a real-world avatar. Figure 1 shows a conceptual image of interaction with such real-world avatars. Similarly to communication using the avatar in the virtual environment, it is essential for the avatar to be visible from all around, as is a person in the real world, because participants may move around in their environment while communicating with the avatar, which causes the spatial relationship among participants to change.

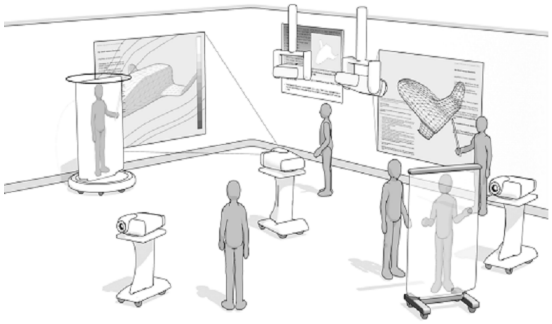


Figure 1: Avatar in the Real World

We have proposed the concept of a "real-world video avatar", that is, the concept of video avatar presentation in "real" space. One requirement of such a system is that the presented figure must be viewable from various directions, similarly to a real human. In this paper such a view is called "multiview". By presenting a real-time human figure with "multiview", many participants can interact with the figure from all directions, similarly to interaction in the real world. We have developed a display system which supports "multiview"[1]. In this paper, we discuss the evaluation of real-time presentation using the display system.

2. Previous Work

The video conference is a primitive implementation of remote communication in which communication that is similar to a face-to-face meeting is realized by using a set of cameras and displays. However, in such systems, the spatial element of communication is ignored, which frequently causes problems in conveying spatial interactions such as pointing to objects and identifying the line of sight of remote participants. The problem has been intensively studied in the field of CSCW and a number of solutions have been proposed[2, 4]. A drawback of conference-style communication is that it forces participants to remain in front of the camera and display set. This is a serious restriction to communication if we consider that, in actual face-to-face communication, people can communicate with one another while freely changing their relative positions.

An ultimate means of supporting spatial communication is the construction of a communication space in a shared virtual environment[3]. There have been a number of studies on implementing such an environment using immersive projection displays[5, 6]. Towards the realization and im-

plementation of such a communication environment, the transmission and representation of human figures that serve as representatives of remote participants is a fundamental topic of research. The human figure presented in the virtual space is called an avatar. The motion of remote participants is reflected in the motion of avatars in the virtual space, and participants are therefore allowed to move freely in this space.

A fundamental difference in implementing an avatar, in contrast to other CG applications, is that the fidelity of avatars is essential from the communication point of view. This is why most approaches to avatar implementation use video images as source data. Immersive Video[7] uses the view volume intersection method to create a rough 3D voxel model of a human figure and projects a video image onto the voxel model. In the Virtualized Reality approach[8], a 3D model of a person is generated by using the images of multiple cameras and computer vision technique.

In a primitive implementation of a video avatar[9], namely, a 2D video avatar, the video image is projected onto a virtual plane at the location of the remote participants in the virtual environment; in the implementation of a 3D video avatar, a video image is projected onto a 3D surface model obtained using a stereo camera or a range finder. There has also been an investigation on a method of capturing images of a person in an immersive virtual environment using multiple cameras[10].

As stated in the introduction, we aim at the representation of an avatar in the real world, and several approaches are conceivable for this purpose. One is to have participants wear a HMD (head-mounted display) operated with the AR (augmented reality) technique[11], and another is to present avatar images by using projectors or other display devices. The method of wearing the HMD is advantageous in that it is capable of presenting avatars at any place with the scene of the environment. However, it has serious technical problems in registration and tracking time delay.

The use of a projector is a common approach to presenting visual information in the real world. The Office of the Future[12] concept involves projecting the scene of a remote office onto walls or other objects in the real world. However, this approach is not suitable for our purpose, because it is not applicable to the case where many people are in a room and communicating with the avatar, and hence different images must be presented to the people depending on their viewing positions.

There are some display devices or systems that provide 3D images visible from various orientations or all around. The multiplex hologram is a kind of holographic display that consists of a cylindrical hologram screen, and it can present different images depending on the orientation of the viewer. However, since the technology to update images in real time has not been established, it is not suitable for the real-time presentation of avatars.

A sophisticated means of presenting an all-around image of a 3D object has been proposed based on the concept of reconstructing ray space, and a prototype system called a multiview 3D display has been developed[13]. There have also been a number of other studies on the implementation of volumetric displays. These displays generate a 3D image by scanning a volume with 2D display surfaces using rotation or translation of the display surface[14, 15]. A feature of this type of display is that the 3D shape is presented as a translucent image. However, although this feature is advantageous for the visualization of volume data, which is a primary application of these displays, it is not desirable for the presentation of a solid image of an avatar.

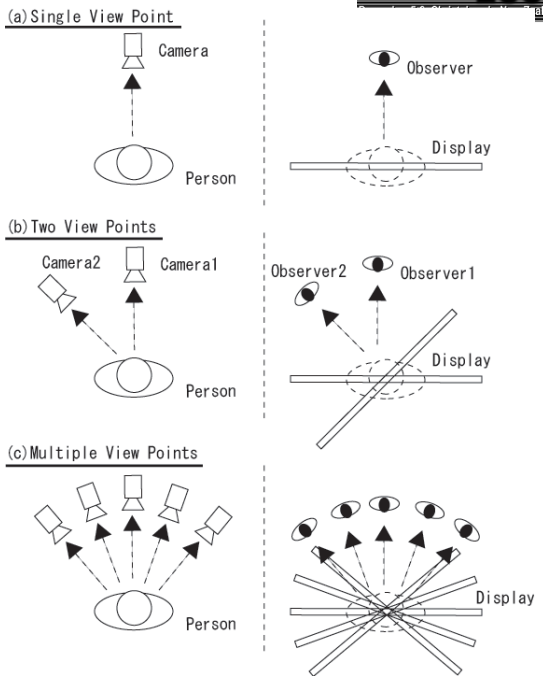


Figure 2: Capturing and Displaying Images

There have been a number of studies in which robots have been used as representatives of remote participants. Although these studies resulted in the successful transmission of spatial pointing actions, gestures[16], head motion and facial expressions[17], the amount and type of information that can be transmitted by robots is still limited in comparison with real-time video images.

3. Real World Video Avatar

We propose a method of presenting a human figure that is visible from all around. In the method, all-around video images of a person are captured, and each image is presented to the viewer at the orientation at which the image was captured.

First, assume the case in which a person is viewed from one position (see Figure 2(a)); a video camera takes an image of the person and the image is presented to a remote viewer. It should be noted that the image of the person obtained by the camera is presented without distortion only when the configurations of the capture system and display system are geometrically similar or congruent with each other.

Next, assume the case in which a person is viewed from two different positions (see Figure 2(b)). Also in this case, for distortion-free viewing, the configurations of the camera system and the display system must be geometrically similar or congruent with each other. However, it becomes clear that a problem of interference between the images taken from these two positions arises; the image on each display panel should be visible only from its corresponding viewing position. For implementation, it is problematic in that the configuration of the display system must be designed such that one display panel does not interfere with

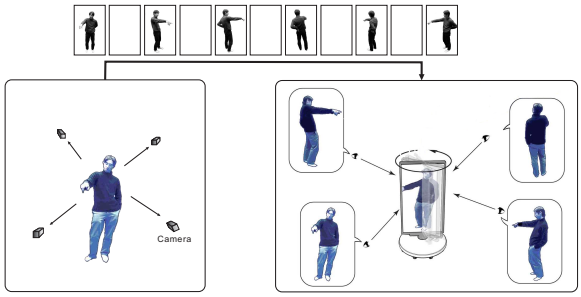


Figure 3: Conceptual Image of our System

Figure 4: Exterior of the Cylindrical Room

the other display panel.

The presentation of images viewed from multiple positions is supported by expanding the system as shown in Figure 2(c); the visible image is switched sequentially according to the motion of the viewer along the viewing positions around the display, and an approximate presentation of the all-around image of a person is implemented. Increase in the number of sets of camera and display panel enables a smoother change of images according to the change of the viewing position.

Since it is impractical to use multiple displays, we employed a time-division approach where a flat display panel is rotated and the image presented by the display is changed according to the orientation of the display panel (see Figure 3). The image of the display ideally must be visible only when the display is directly facing the viewing position, or the display must transmit the ray that goes toward the viewing position while cutting out others. It is important to note that the horizontal orientation of the ray transmitted by the display varies depending on the horizontal location of the pixel on the display panel. As we state below in section 4.2, we use a "privacy filter" sheet to approximately attain this effect.

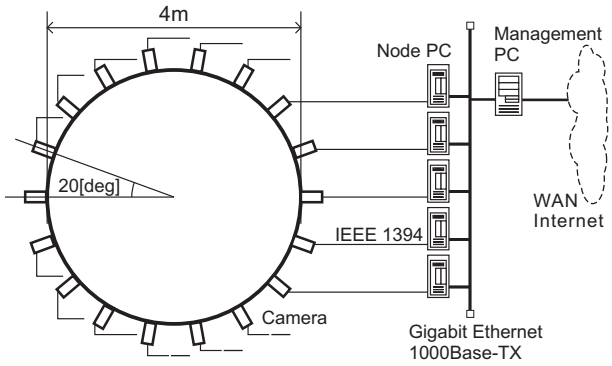


Figure 5: Diagram of the Capture System

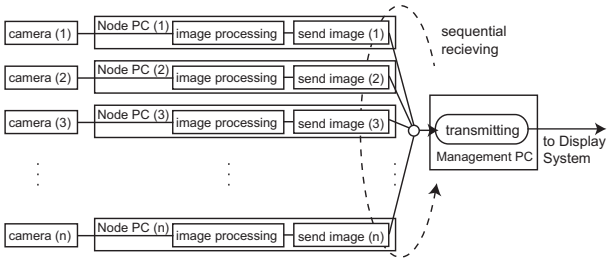
4. Prototype System

In this work, we implemented a prototype system based on the approach. The system consists of a capture system and a display system. In the capture system, images of a person are captured from all around using multiple cameras. In the display system, the images are displayed on a revolving display panel according to the direction.

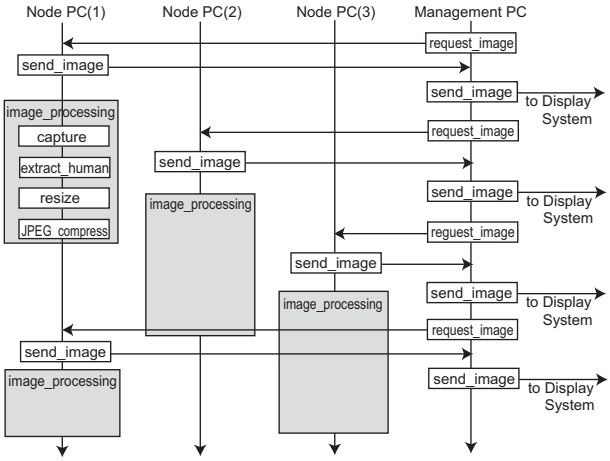
4.1. Capture System

The capture system consists of a cylindrical room that is equipped with eighteen cameras, node PCs and a management PC (see Figure 4, 5). The radius of the room is 2000[mm], and the interior wall and floor of the room are covered by a blue screen sheet for ease of chroma-key processing. The cameras (DFW-X700, SONY) are located around the wall of the room at intervals of 20 degrees at a height of 1200[mm] from the floor, and are fixed so that the optical axes of the cameras are horizontal and cross at the center of the room. The resolution of the cameras is 1024×768 pixels and their frame rate is 15[Hz]. To maintain the performance of the system, the cameras should not be manage by one PC, and therefore we use node PCs which are connected to each camera individually. Each node PC (Pentium 4 2.0GHz) is connected to one camera through IEEE 1394 cable (400Mbps) and receives image data (uncommpressed XGA bitmap) from the camera. All node PCs are connected with the management PC (Pentium 4 XEON 2.8GHz) through a gigabit network (1000BASE-TX). Each node PC has only to control one camera, while the management PC manages the overall process of gathering images from the node PCs and sends the gathered images to the display system via an ethernet.

For the transmission of images of the human figure, it is necessary to minimize the delay time per frame. To transmit images of a human figure with a high update rate, we adopt the following process. Figure 6 is flow diagram of the transmission process of the capture system. The management PC sends request commands to each node PC sequentially. The management PC transmits the image received from the node PC to the display system, and sends a request to the next node PC. In each node PC, the image processing is carried out independently. Each node PC captures an image from each camera, and extracts a human figure from the captured image within the frame rate of the cameras (15 fps). Also, for the transmission, the generated images are compressed. When the node PC receives an image request, it transmits the generated image to the management PC.



(a) Overview of Transmission Process



(b) Time Line of Transmission Process

Figure 6: Transmission Process of the Capture System

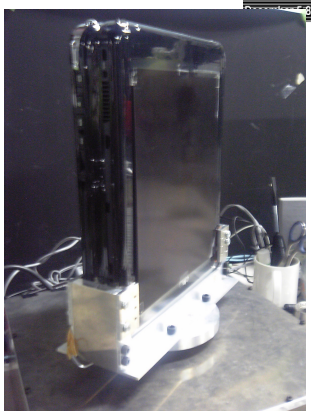


Figure 7: Prototype of Display System

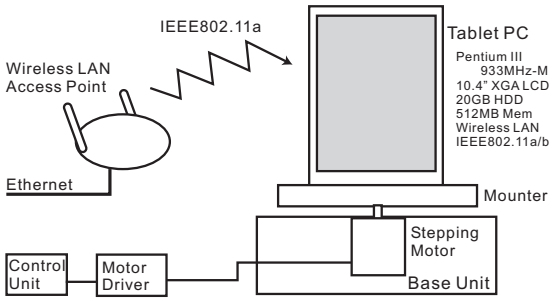


Figure 8: Diagram of the Display System

4.2. Display System

4.2.1. 1/10-scale display system

The display system consists of two tablet PCs (Tablet PC TC1100, HP), and a revolving mechanism (see Figure 9, 10). The tablet PC is equipped with a 10.4-inch XGA LCD, whose refresh rate is fixed at 60[Hz]. The CPU is a Mobile Pentium M 753 1.20[GHz] and the main memory is 512[MB]. To double the update rate, we sticked two tablet PCs together as shown in Figure 9. To reduce the display's viewing angle, we used a privacy filter (Privacy Computer Filter, 3M). The revolving mechanism consists of a stepping motor and a motor driver that controls the frequency of rotation.

The display system is designed as a 1/10-scale system of the capture system; the offset of the display panel from the rotation axis is approximately 0[mm] and the distance from the revolving axis to the ideal viewing position is 200[mm], which is 1/10 the radius of the camera arrangement in the capture system. The display panel covers only a part of the field of view; images from the capture system are cropped in the rendering process by the tablet PC. As shown in Figure 9, the filter sheet is formed so that it is part of the cylindrical surface; the radius of curvature is 200[mm].

The images from the capture system are received through a wireless LAN (IEEE803.11g, 54Mbps), buffered in the main memory, and rendered on the display panel. The rendering process was implemented using DirectX API.

4.2.2. 1/1-scale display system

To present life-size video avatar, we constracted life-size display system, consists of a Plasma Display Panel (42 inch PDP, SONY), Note PC, and a revolving mechanism (see Figure 9, 10). The display system is a system of about 1800mm in height and a display from the knee, the life-size can be displayed enough. The size of the PDP is 920mm $\times$ 518mm and the refresh rate is 85Hz. The PDP and the note PCs are supplied 3500W in the power supply through the slip ring. To take synchronization, each note PC is networked by LAN. The privacy filter (Privacy Computer Filter, 3M) is pasted to the each face of PDP as well as the tablet PC type prototype.

The critical frequency of rotation for the sympathetic vibration was requested about 3.3 [Hz] from approximate expression of the Dan curry about the axial circumference of the system. 3.3 [rps] is provided as a limitation cycle of this system. In this case, the number of presented direction becomes 85/3.3=26 directions.

The rendered images should be synchronized with the revolution of the Note PC. The timing of receiving images from the capture system is unpredictable. To render and update images according to the direction of the Note PC, we adopt separation of the receiving and drawing processes. Figure 11 is the flow diagram of the receiving and drawing processes of the display system. We implemented a double-buffer mechanism for each camera for preserving and for drawing. The receiving process refers to the buffer for preservation and the drawing process refers to the buffer for drawing. The camera number is recorded in the packet

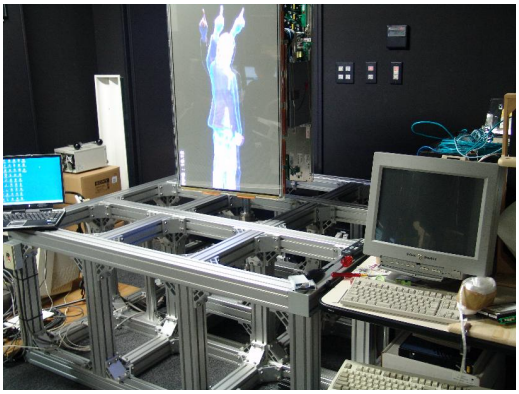


Figure 9: Prototype of Display System

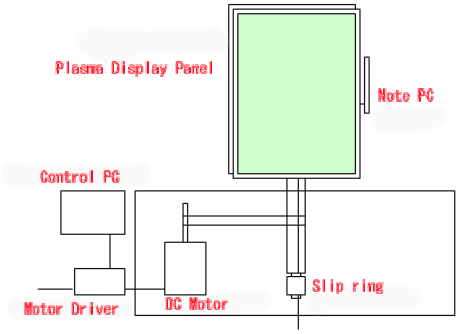


Figure 10: Diagram of the Display System

header of each forwarded image.

The update time of transmitted images depends on the network infrastructure. The receiving process stores the received images in the buffer according to the camera number. The each buffers for preservation and drawing are swapped immediately, after receiving and preserving the image. As a result, the latest received image is always in the buffer for drawing. The drawing process sequentially refers to the buffer for drawing of each camera and renders the newest image according to the direction of display.

5. Experiments and Results

In our prototype system, we connected the capture system and the display systems via the internet. To evaluate the performance of the capture and display systems, we measure the update rate of transmission. And, to apply them as a remote communication system, we measure the delay time of transmission.

5.1. Update Rate

The update rate is an important issue for the display system. Let the update rate of the image on the display panel be f_{disp}[Hz], the rotational frequency F[Hz], and the number of presenting directions n. These values have the relationship $f_{disp} = F \times n$. Each direction's image is updated once per rotation, and thus it should be noted that F is relevant to the temporal resolution of the presented image,

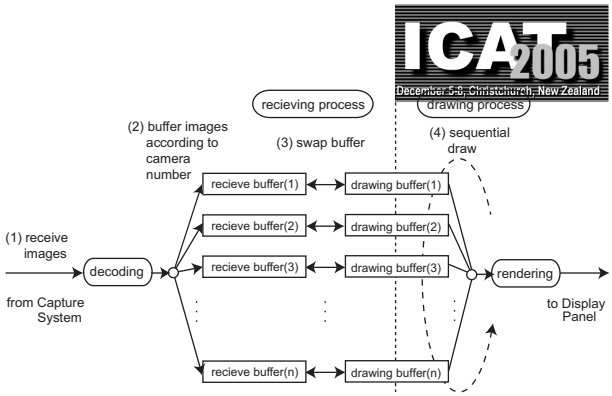


Figure 11: Process of the Display System

while n is relevant to the angular resolution of the orientation of view point. Since the refresh rate of the display panel is limited, the function suggests a trade-off relationship between the temporal and the spatial resolutions of the display system.

For example, to maintain the update rate of the display system, we determined that the resolution of transmitted frames at 120×90 pixels, and the update rate at 30.0 [Hz]. In this case, based on the above mentioned relationship between the update rate and the frequency of rotation, we decided the number of presenting directions of our prototype display system n is 6 and the frequency F is 5 [rps]. In the case that the number of presenting directions is 6, the capture system uses six cameras at 60-degree intervals and transmits images from the cameras.

5.2. Delay Time

The delay time is an important issue for the remote communication system. Hence, we measure the delay time of our transmission and display system. The delay time of the system consists of the processing time in the capture system, the processing time in the display system, and the time for communication between the systems. In addition, the processing time of the display system is divided into waiting time and drawing time. The waiting time until the drawing is initiated changes at random according to the direction the recieved frame should be displayed in and that which the display is facing in. For example, when the direction the display is facing is opposite from in which the received frame should be displayed, the recieved frame must wait and not initiate drawing until the display has half rotated.

This waiting time can be estimated statistically. When F is the rotational frequency of the display, the elapsed time for each revolution of the display is $1/F$ [s]. The number of frames received for each revolution of the display is expressed by f_{trans}/f_{disp}, where the refresh rate of the display panel is f_{disp}[Hz], and the update rate of transmission is f_{trans}[fps]. Because the latest received frame is drawing, the maximum waiting time t_{max} is

$$t_{max} = \frac{1}{F} / \frac{f_{trans}}{f_{disp}} = \frac{f_{disp}}{F \cdot f_{trans}} \quad (1)$$

The number of presenting directions is $n = f_{disp}/F$, and t_{max} is shown below.

$$t_{max} = \frac{n}{f_{trans}} \quad (t_{max} \leq \frac{1}{F}) \quad (2)$$

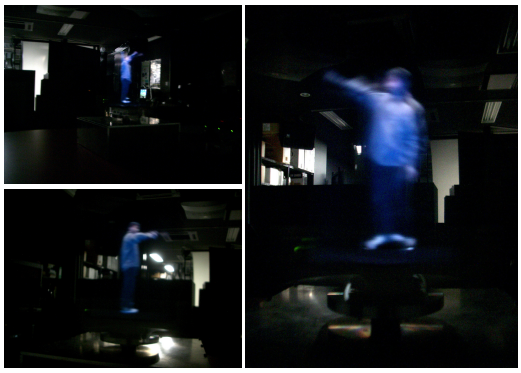


Figure 12: The 1/10-scale Presentation

The average waiting time t_{avg} is

$$t_{avg} = \frac{n}{2f_{trans}} \quad (t_{avg} \leq \frac{1}{2F}) \tag{3}$$

The calculated value and the measured value of the maximum and average waiting times were almost in correspondence. By calculating based on these equation (2), (3), we can estimate waiting time in case of any n and f_{trans}.

5.3. Real-Time Presentation

Based on the measurement of the system's frame rate and delay time, we determined the resolution of the transmitted and drawn images to be 120×90, and the refresh rate f_{disp} to be 40 [Hz]. Also, we decided that the number of presenting directions n should be 9. Hence in the capture system, images were captured from nine viewing positions using nine cameras set at 30-degree intervals.

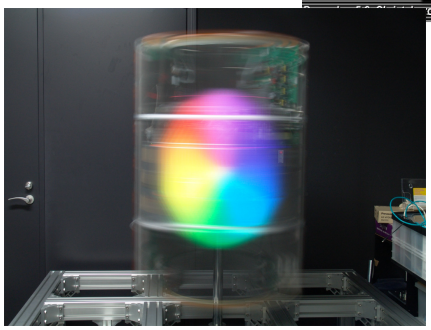
5.3.1. 1/10-scale display system

The display's update rate of the tablet PC was determined to be 40[Hz]; the update rate was reduced from that in the experiment on the off-line presentation, because of the more lengthy process of decoding JPEG images. The frequency of rotation of the display was approximately 4.5[Hz], which is the maximam speed of the motor. By using two tablet PCs, the system double the update rate, 8.88[Hz].

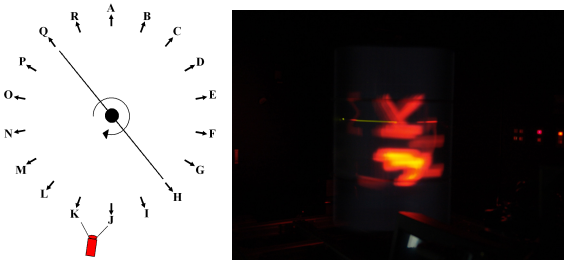
Figure 12 shows the image presented by the display system. The number of images transmitted from the capture system to the display system per second was about 68.7 and the delay time of transmission was about 86.515[ms] on average. It was sufficiently possible to observe the remote participant's behavior including gesture, and other information shuc as clothes and the hairstyle. The systems are located 10km apart, and we connected and integrated them via the Internet.

5.3.2. Prelimary presentation in 1/1-scale display system

To evaluate the PDP type prototype system, we presented color pattern and 18 characters to A-R in 18 directions. Figure 13 (b) shows the presentation method and the result. As a result, we can see only the character required according to the direction. Figure 14 shows the scene of presenting a human figure in our prototype system. The frequency of rotation of the display was 0.8[rps] for safety reasons. In



(a) Color Pattern



(b) 18 caharacters

Figure 13: Test pattern in the PDP type prototype

Figure 14: The 1/1-scale Presentation

this case, the frame rate of the system is 1.6Hz. As shown in figure 14, it was sufficiently possible to observe the all-around image of a person.

Through the experiments discussed above, it was proved that our approach is feasible and the prototype system is capable of presenting the image of a person at a remote location as an image visible from all around in real time. At the same time, a number of problems were revealed by the experiments.

One of the problems is that, when a viewer moves around the display system, there are some positions from which the viewer sees images from two adjacent viewing positions overlapping each other. The overlap is due to the characteristic of the display panel, namely, that the viewing angle of the privacy filter sheet is not sufficiently small; if the image on the display panel is updated while the viewer is within the viewing angle of the display panel, the viewer sees the images both before and after the update.

To transmit and present a remote participant as if the participant is stand nearby, and allow the participant to communicate with the remote participant in the real world, it is necessary that there be a greater number of presented directions and that the update rate of the human figure is higher.

6. Conclusion

An approach to capturing and transmitting a human figure in real time and presenting it in the real world was proposed. A prototype system that consists of a capture system and a 1/1-scale display system was implemented. In experiments with these systems, a 120x90-resolution human figure which can be viewed from six directions was presented. The average delay time was 177[ms]. It was confirmed that the display system is capable of presenting a human figure that is visible from all around, and that the human figure can be updated in real time using images from the capture system. The results proved that our approach is effective for the presentation of a human figure in the real world, and that it is possible to transmit a remote person in real time.

In our future work, we will apply our approach to the further development of remote communication in the real world. By realizing spatial remote communication in the real world, it will become possible for participants to talk about the real environment and real objects. Since the presentation of a real-size human image is essential for spatial communication, we will improve the presentation of a real-size (i.e., an actual scale) display system by using 4 or more panels. We are also interested in integrating a mechanism for supporting locomotion to enable spatial translation of the avatar in the environment.

References

[1] H. Maeda, K. Hirose, J. Yamashita, K. Hirota and M. Hirose, "All-Around Display for Video Avatar in Real World", Proceedings of The Second IEEE and ACM International Sysmposium on Mixed and Augumented Reality (ISMAR03), pp. 288-289, 2003.

[2] K. Okada, F. Maeda, Y. Ichikawa and Y. Matsushita, Multiparty Videoconferencing at Virtual Social Distance: MAJIC Design, Proc. CSCW'94, pp. 385-393, 1994.

[3] H. Takemura and F. Kishino, Cooperative Work Environment Using Virtual Workspace, Proc. CSCW'92, pp.226-232, 1992.

[4] H. Ishii and M. Kobayashi, ClearBoard: A Seamless Medium for Shared Drawing and Conversation with Eye Contact, Proc. CHI'92, pp. 525-532, 1992.

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

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

[7] S. Moezzi, A. Katkere, D. Kuramura and R. Jain, Immersive Video, Proc. VRAIS'96, pp. 17-24, 1996.

[8] T. Kanade, P.J. Narayanan and P. Rander, Virtualized Reality: Concepts and Early Results, IEEE Workshop on the Representation of Visual Scenes, June, pp. 69-76, 1995.

[9] K. Tamagawa, T. Yamada, T.Ogi and M. Hirose, Developing a 2.5-D Video Avatar, IEEE Signal Processing Magazine, Vol. 18, No. 3, pp. 35-42, 2001.

[10] O. G. Staadt, A. Kunz, M. Meier and M. H. Gross, The Blue-C: Integrating Real Humans into a Networked Immersive Environment, in ACM Collaborative Virtual Environments 2000.

[11] S.J.D. Prince, A.D. Cheok, F. Farbiz, T. Williamson, N. Johnson, M. Billinghurst and H. Kato. 3D Live: Real Time Captured Content for Mixed Reality, International Symposium on Mixed and Augmented Reality, 2002.

[12] R. Rasker, G. Welch, M. Cutts, A. Lake, L. Stesin and H. Fuchs, The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays, Proc. SIGGRAPH'98, pp. 179-188, 1998

[13] T. Endo, Y. Kajiki, T. Honda and M. Sato, Cylindrical 3D Display Observable from All Directions, Proceeding of PG'00, pp. 300-306, 2000.

[14] K. Kameyama and K. Ohtomi, A Direct 3-D Shape Modeling System, Proc. VRAIS'93, pp. 519-524, 1993.

[15] J. Napoli, D. M. Hall, R. K. Dorval, M. G. Giovinco, M. J. Richmond and W. S. Chun, 100 Million-Voxel Volumetric Display, AeroSense 2002 - for Cockpit Displays IX: Displays for Defense Applications, 2002

[16] H. Kuzuoka, T. Kosuge and M. Tanaka, GestureCam: A Video Communication System for Sympathetic Remote Collaboration, Proc. CSCW'94, pp. 35-43, 1994.

[17] N. P. Jouppi, First Steps Towards Mutually-Immersive Mobile Telepresence, the Proceedings of the ACM Conference on Computer Supported Cooperative Work, 2002.