Multi-Resolution Stereoscopic Immersive Communication Using a Set of Four Cameras

Takeshi Naemura, Kaoru Sugita, Takahide Takano, and Hiroshi Harashima

Dept. of Inform. & Commun. Eng., The Univ. of Tokyo,
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN

ABSTRACT

In projection-based virtual reality systems, such as the CAVE, users can observe immersive stereoscopic images. To date, most of the images, projected onto the screens, are synthesized from polygonal models, which represent the virtual world. This is because the resolution and the viewing angle of a real image are not enough for such large screen systems. In this paper, the authors propose a novel approach to avoid the problem by exploiting the human visual systems. In the proposed system, the resolution of the center of view is very high, while that of the rest is not so high. The authors constructed a four-camera system, in which the pairs of NTSC cameras are prepared for both left and right eyes. Four video streams are combined into one video stream and captured by a graphics computer. Wide-angle multi-resolution images are synthesized in real-time from the combined video stream. Thus, we can observe the wide-angle stereoscopic video, while the resolution of the center of view is high enough. Moreover, this paper proposes another configuration of the four-camera system. Experimental results show that we can observe three levels of viewing angle and resolution by the stereoscopic effects, while images for each eye has just two levels. The discontinuities in the multi-resolution images are effectively suppressed by this new lens configuration.

Keywords: Stereoscopic Communication, Camera Array, Real-Time Image Processing, Large Screen System, Texture Blending, Video Texture

1. INTRODUCTION

The present technology makes it possible to surmount distance and time barriers, when transmitting information. Nevertheless, a variety of technical issues must still be addressed to enable people, separated by distance, to engage in collaborative work with a clear understanding of each other's intentions. Much research in recent years has consequently focused on "shared space communications" that would give people, that are physically far from each other, the sensation of actually meeting together.[1]

In the field of computer graphics (CG), highly realistic virtual space can now be created, and systems are being used that can achieve a high sense of presence by displaying high-resolution stereoscopic images on large screens in a "surround" format.[2] Compared to such high-resolution CG images on large screens, however, the input of real video images is inferior in terms of resolution even if HDTV cameras are used. In other words, the introduction of real images will not necessarily provide a higher sense of presence than that of CG. In addition, communications systems based on real images require broadband transmission paths to transfer video, and a number of technical issues remain to be solved in this regard.

Against this background, the authors have been studying a stereoscopic communication system using a set of four NTSC cameras that achieves both wide-angle images and picture resolution on par with that of large-screen displays.[3,4] This report reviews a means of improving the quality of multi-resolution stereoscopic images by making texture semitransparent, a solution to the problem of positional gaps in video sequences when transmitting video over ATM networks, and screen partitioning ratio and associated screen resolution when transmitting video from four cameras.

Further author information : naemura@hc.t.u-tokyo.ac.jp, http://www.hc.t.u-tokyo.ac.jp/~naemura

2. BACKGROUND

Current image-based communication services possess only the minimum capabilities necessary to transmit visual information. As a consequence, such services are not necessarily satisfactory when compared to face-to-face discussions, since problems, associated with screen size, resolution, stereoscopic effect, etc., prevent a high sense of presence from being achieved. Here, to faithfully transmit the intent of meeting participants and the state of the conference-room scene between remotely located parties and to produce the feeling of an actual in-person meeting, a communication system must emphasize the sharing of space itself in addition to the transmission of images. Achieving such a service requires specialized technologies for presenting wide-angle and high-resolution images and for transferring images efficiently using two-camera and multi-camera parallax encoding. This section overviews current research results in the various fields associated with these technologies.

2.1. Three-dimensional communication

To date, a variety of systems dealing with three-dimensional communication have been studied ranging from two-camera stereoscopic communication to multi-camera image communications.[5] At present, encoding techniques based on parallax compensation and prediction encoding are being established in conjunction with MPEG standardization and various types of related hardware are being developed. This research, however, has centered on NTSC-level images and there have been almost no studies on levels of resolution exceeding HDTV.

2.2. Three-dimensional display technology

There has been much talk recently about systems that can provide an immersed visual experience with a high sense of presence using high-resolution, large-screen displays that surround users.[2] The resolution of one large screen is normally about 1000-by-1000 pixels, which means that the resolution of real NTSC video images is insufficient for such a screen. Similarly, the resolution of HDTV images is insufficient for the horizontal resolution of 4000 pixels achieved by four surrounding screens. As a result, these systems have up to now been mainly used for the display of computer-generated graphics.

2.3. Computer-based processing/synthesizing of real images

Conventionally, the general approach to processing and synthesizing real images in real time is to use optical equipment like half mirrors. Advances in computer processing power, however, have made it possible to transform or overlay NTSC video-signal input in real time as desired and display the results. In short, computers enable advanced processing and synthesizing to be performed compared to the use of optical equipment. An example of such advanced technology is Video-Based Rendering that can synthesize images from any line of sight based on the input of real multi-camera video images.[6] Real-time, real-image-processing technology by computer is also an important elemental technology of this research.

3. PROPOSAL

In this research, the authors have constructed an image-based communication system consisting of an image input section (image pickup system), transmission section (ATM networks), and output section (large-screen display), as shown in Fig. 1.

3.1. Obtaining wide-angle stereoscopic video

The image-pickup system shown in Fig. 2 has also been constructed to obtain wide-angle stereoscopic video in which importance is attached to the resolution of the center field of view.

The system features two cameras placed in optically conjugate positions on either side of a half mirror oriented at an angle of 45°. This arrangement makes it possible to pick up images having the same optical axis but different angles of view. Two sets of such a combination, moreover, are provided to form two pairs of "eyes." Stereoscopic pickup can therefore be achieved by placing left/right cameras in parallel with each other and separating them by the same distance between human eyes (65 mm).

The cameras used here are SONY XC-999 12-inch-CCD ones, and the lenses attached to these cameras come in two combinations as shown in Table 1. These two sets of lens combinations are compared in section **4.2**.

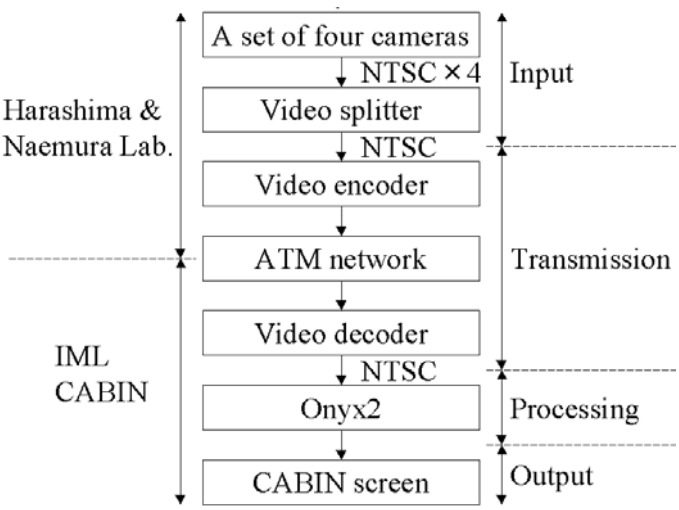

Figure 1. System configuration

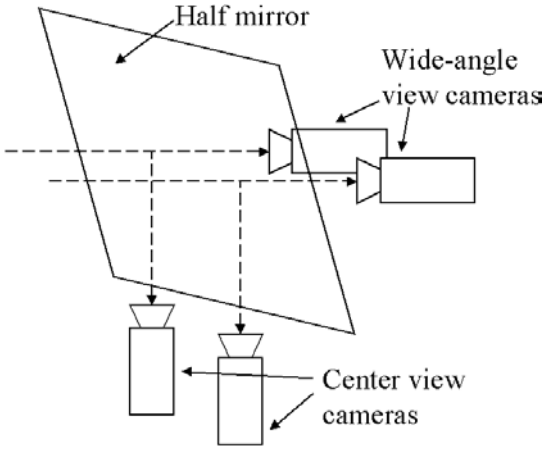

Figure 2. A set of four cameras

Furthermore, to prevent lens focus from fluctuating after being adjusted, all lenses are set to ∞, and to maintain a fixed brightness among the video signals from the four cameras, the automatic gain control function on each camera is not employed.

Figure 3 shows the result of combining video from the four cameras in the pickup system into one screen using a video splitter. The center-view video on the lower half of the screen becomes inverted due to the half mirror.

Table 1. Sets of lenses

	Left		Right	
1	12.0 mm	(narrow)	12.0 mm	(narrow)
	3.5 mm	(wide)	3.5 mm	(wide)
2	6.0 mm	(normal)	12.0 mm	(narrow)
	3.5 mm	(wide)	3.5 mm	(wide)

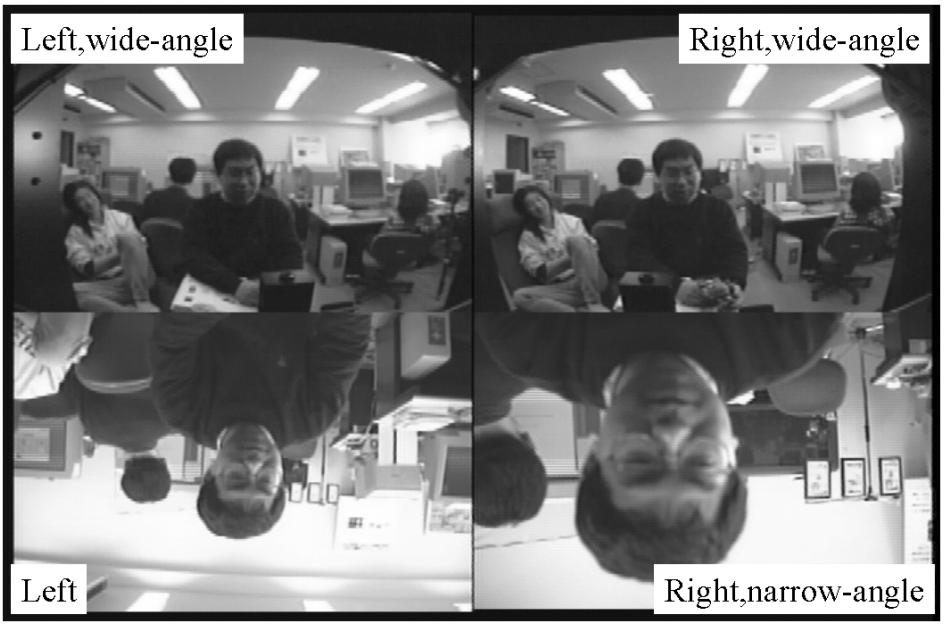

Figure 3. Input video image

3.2. Video transmission over ATM networks

To transmit video over an ATM network, the authors use K-Net Corporation's CellStack video having an ATM 155Mbit/s SDH/SONET interface. CellStack video enables the transmission of 30 frames per second at a resolution of 720-by-480 pixels by encoding NTSC signals in Motion-JPEG.

Specifically, we set up sending-side and receiving-side CellStack video units at two separate locations at the University of Tokyo, and transmitted the video from four cameras.

3.3. Real-time synthesis of stereoscopic video

To synthesize video in real time, the system uses the Onyx2 graphic workstation from SGI featuring the DIgital Video Option (DIVO) for video input. This configuration enables NTSC input in the form of 720-by-486-pixel, 24-bpp full-color images to be stored in memory at a rate of 30 images per second.

Images stored in memory can then be mapped to polygons as video texture. The use of such a rendering-engine function speeds up warping processing in image synthesis, enabling real-time displaying to be performed.

In this research, the system is configured with only one DIVO, and input to the Onyx2 consists of a single video signal achieved by integrating the video sequences of four cameras by a video splitter. Although the resolution of each video sequence drops as a result of this splitter, sufficient resolution is obtained when displaying the video on the screen, as described in section **4.4**.

3.4. Stereoscopic display on the screen

The Computer Augmented Booth for Image Navigation (CABIN) is large-screen, three-dimensional display equipment installed within the Intelligent Modeling Laboratory (IML) at The University of Tokyo.[7] It features five 2.5-meter-square screens, each of which can display 750-by-750-pixel stereoscopic images. Considering the difficulty of obtaining sufficient resolution by real-video input, CABIN has been used mainly for the presentation of CG-based virtual space.

4. EXPERIMENT

4.1. System calibration

4.1.1. Survey of lens characteristics

When using more than one camera, attention must be paid to differences between the lens characteristics of each camera. In this experiment, the authors first attached a number of lenses of the same specifications to the same camera one at a time and took video with each. Then, after comparing the video so obtained, we used only those lens having nearly the same characteristics.

4.1.2. Extracting texture from input video

As shown in Fig. 3, input video to the Onyx2 workstation combines the video sequences from four cameras on one screen. To therefore synthesize a single picture, single streams of video must be appropriately extracted from the input video. Moreover, since a variety of devices will be connected to the video transmission path when transmitting motion pictures in CellStack video, positional gaps in the video input to Onyx2 may occur as connected devices change. For this reason, the boundaries of each picture in the partitioned video image are automatically detected in order to deal flexibly with positional gaps in video input.

4.1.3. Adjustment of wide-angle lens distortion

This research employs wide-angle lens to pick up video with a wide field of vision. One feature of wide-angle lens, however, is distortion in the corners of the picture. This problem can be solved by mapping the texture of the wide-angle image to a spherical surface. To this end, the authors created a program that allows us to interactively modify the mapping area on the sphere and determine which parameters with respect to input video would eliminate a visual perception of this distortion. This adjustment need by performed only once for the lens in question; thereafter, distortion adjustment can be performed automatically in real time.[3] The result of performing such adjustment on the right wide-angle picture in Fig. 3 is shown in Fig. 4.

Figure 4. Example of distortion adjustment

4.1.4. Image synthesis through warping

When overlaying wide-angle video and center-view video, discontinuities in their boundaries can hinder a sense of presence. To therefore synthesize a picture in which discontinuities are suppressed, a 4-by-4 mesh is established on the center-view image and mesh vertices are manipulated using a GUI that the authors have created (Fig. 5).

Figure 5. GUI for central-view warping

Here, the mesh deformation parameters are characteristic values of the camera pickup system, and they need be adjusted only once. This software-based two-dimensional warping makes it possible to compensate for differences in camera and lens characteristics and offset of the optical axis between cameras. However, if two cameras with different angles of view are not strict optical conjugates, the above positional gaps cannot, in principle, be compensated for. This means that even if mesh deformation is optimized for a certain still picture captured by camera, discontinuities will again appear on the screen for an ever-changing motion picture, especially one in which the depth of a subject changes. To therefore reduce such discontinuities, it becomes necessary to place an object at various depths and to perform fine adjustments to camera position (hardware calibration) and adjustments to warping parameters (software calibration) at each depth.[4]

In addition to the above positional gaps, there is also the problem of gaps in picture quality that arise due to sudden changes in resolution near boundaries. In this paper, resolution at the boundaries between pictures is made to change smoothly from a visual viewpoint. Specifically, a texture blending function of a rendering engine is used to make the ends of a center-view image semi-transparent and to produce a linear change in the mixing ratio of wide-angle and center-view images near the boundaries. As a result, no discontinuities can be noticed at boundaries, although a slight sacrifice must be made with regard to detail in the center-view image.

4.1.5. Effects of video transmission over ATM networks

After implementing the above processing, video was transmitted over ATM networks and displayed by the CABIN system. At this time, the effects of compression and transmission bit rate on picture degradation, noise, delay, etc., were examined, but the most serious problem uncovered was positional gaps in the input video as described above. Even if various parameters are adjusted after integrating the video sequences from four cameras into one video sequence and inputting this directly into Onyx2, a gap of several pixels can still occur due to ATM transmission. This situation is dealt with by readjusting parameters in a semi-automatic manner. The cause of this problem is

considered to be repeated A/D and D/A conversion processing, and until all processing from input to output is performed digitally, this problem cannot be prevented.

4.2. Visual effect of multi-resolution stereoscopic images

This system picks up multiple video images using multiple lenses having different angles of view, and eventually synthesizes these images into a single stereoscopic image. Here the authors examine the visual effect resulting from different combinations of lenses.[3] The authors used in particular the two sets of lenses listed in Table 1 and compared the stereoscopic sensations of the stereoscopic images produced by each set on the screen.

For lens set (1), the size of the high-resolution center-view area is the same for both left and right images, and resolution itself is the same (Fig. 6). When combined into a stereoscopic image, both a high-resolution area and a low-resolution area can be observed.

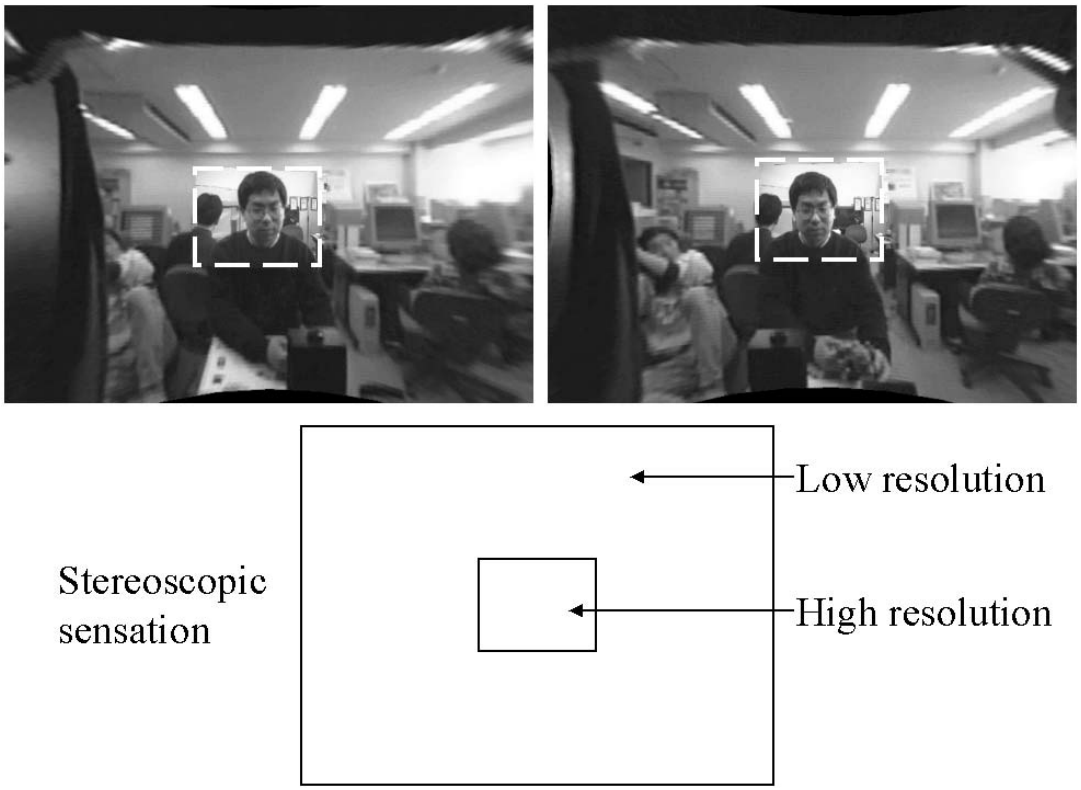

Figure 6. Example of the first set of lenses

For lens set (2), the observers of the resulting stereoscopic image felt as if a picture with three levels of resolution was being presented (Fig. 7). Here, the fact that there are only two levels of picture resolution presented to each eye would not be understood unless one eye was covered. Compared to the perception of two levels of resolution, that of three levels of resolution makes it difficult to notice the boundary sections (between video images) where resolution changes.[3]

A characteristic of human vision when viewing a stereoscopic image in which left/right resolution differs is the feeling that a high-resolution picture is being observed. The field of picture compression makes use of this characteristic.[8] More studies, however, should be performed on stereoscopic characteristics in an environment having a mixture of various degrees of resolution. This research has achieved a means of suppressing the perception of sudden changes in resolution by using four cameras to attain three levels of resolution.

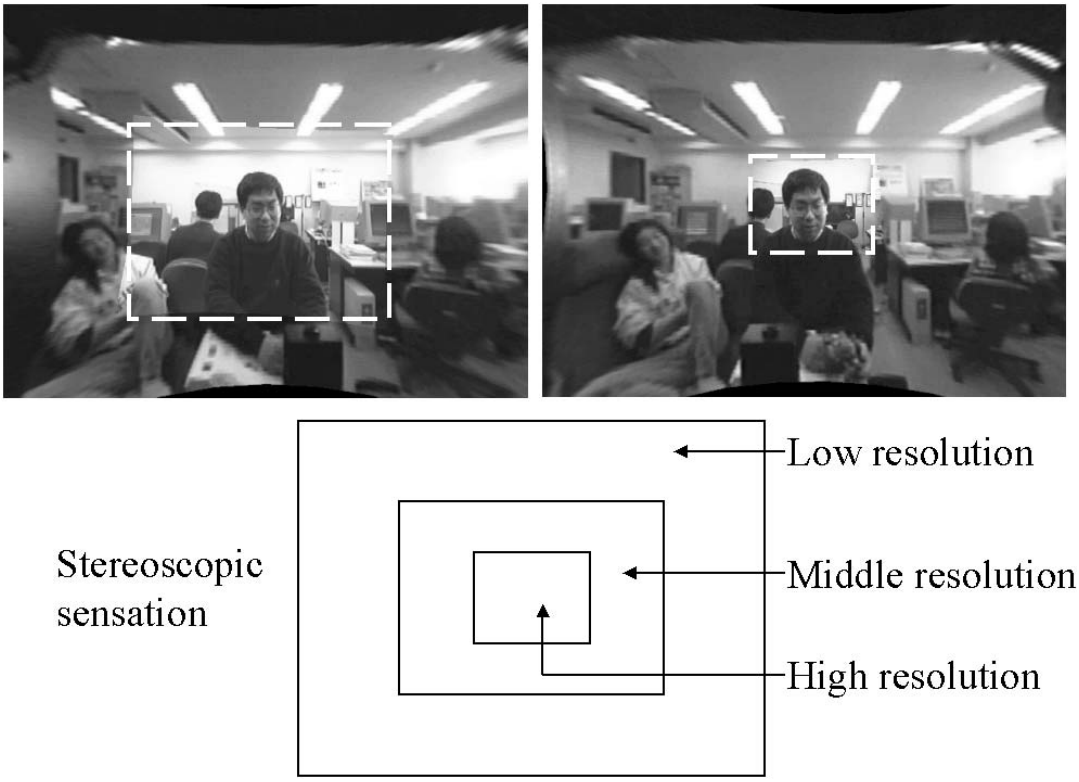

Figure 7. Example of the second set of lenses.

4.3. Visual effect of input-video partitioning ratio

The video splitter used in this research can be set to any partitioning ratio, which means that the resolution of constituent video can be changed as desired. The authors therefore varied the partitioning ratio and evaluated the results of picture synthesis with the aim of determining the most appropriate ratio.

First, for the case of inputting only a wide-angle view (Fig. 8), while resolution is uniform across the entire picture, resolution of the center section feels inadequate.

Next, when partitioning the screen in a 1:1 ratio between wide-angle video and center-view video (Fig. 9), the resolution of wide-angle video turns out to be less than that of Fig. 8. The resolution of center-view video, however, is higher, which makes it easier to read the facial expressions of people in the center of the picture.

Finally, when partitioning the screen in a 1:3 ratio between wide-angle video and center-view video (Fig. 10), the resolution of center-view video becomes even higher, but the drop in resolution in wide-angle video at the edges of the screen becomes noticeable.

The above results indicate that a partitioning ratio of 1:1 between wide-angle video and center-view video provides a favorable visual sensation.

4.4. Relationship between input video and screen resolution

Based on the results presented in section **4.3**, the resolution of one camera's video at input to Onyx2 becomes about 340-by-220 pixels for a partitioning ration of 1:1 between wide-angle video and center-view video. Furthermore, based on the results of picture synthesis, the angle-of-view ratio becomes 4:2:1 when picking up video with three types of lenses of focal lengths 3.5 mm, 6.0 mm, and 12.0 mm.

If we therefore set the viewing position of the observer so that the field of vision of one CABIN screen is the same as that of wide-angle video, the output resolutions of wide-angle, middle-angle, and narrow-angle video become as shown in Table 2.

Figure 8. Wide-angle view only

From the table, we can see that the output resolution of narrow-angle video is actually less than its input resolution. This indicates that transmission of video at sufficiently high resolution (with respect to performance limitations of the display system) can be achieved in the vicinity of the center view despite a sacrifice in resolution that must be made when integrating information from four cameras.

In addition, as the results of the above study depend on the view angles of the lenses used and display resolution of the CABIN system, as well as on viewing position of observers, there is still room for additional studies targeting various combinations of these parameters.

Figure 9. Wide-angle:center = 1:1

Table 2. Resolution comparison

View angle	Wide	Middle	Narrow
Focal length	3.5 mm	6.0 mm	12.0 mm
Input resolution	340×220	340×220	340×220
Output resolution	750×562	375×281	187×140

Figure 10. Wide-angle:center = 1:3

5. CONCLUSIONS

This paper has examined a stereoscopic communication system that provides both high picture quality commensurate with the resolution of large screens and a wide field of view. This system was achieved by constructing a four-camera pickup system using optically conjugate cameras and obtaining spatial information with cameras having different angles of view. It was shown that overlaying stereoscopic images having different left/right resolutions and view angles can produce a multi-resolution stereoscopic sensation with little sense of incongruity.

Particular results are summarized below.

- The quality of multi-resolution stereoscopic images was improved by using a texture-blending function.
- The problem of positional gaps due to ATM transmission was pointed out and handled by automating the detection of picture boundaries.
- The feasibility of employing a video splitter was investigated from the viewpoints of picture partitioning ratio and display resolution when transmitting four video streams.

Future research subjects include automatic warping and the effects of picture compression on multi-resolution display in stereoscopic communication. In addition, we expect the design of a high-performance pickup system using HDTV cameras for stereoscopic communication to become a topic of interest.

REFERENCES

1. T. Naemura, *Ray Based Coding of Real Space and Its Application to Augmented Spatial Communication*. PhD thesis, Dept. of Elec. Eng., The Univ. of Tokyo, 1996. (in Japanese).
2. C. Cruz-Naira, D. J. Sandin, and T. A. DeFanti, "Surround-screen projection-based virtual reality: The design and implementation of the CAVE," in *SIGGRAPH'93*, pp. 135 – 142, 1993.
3. T.Naemura, K. Sugita, T. Takano, and H. Harashima, "Wide-angle stereoscopic communication with four cameras," in *the 3rd Image Media Processing Symposium (IMPS98)*, pp. 109 – 110, 1998. (in Japanese).
4. K. Sugita, T. Takano, T.Naemura, and H. Harashima, "Wide-angle stereoscopic communication using a set of four cameras," in *3D Image Conference '99*, pp. 37 – 42, 1999. (in Japanese).
5. T. Naemura, M. Kaneko, and H. Harashima, "Compression and representation of 3-D images," *IEICE Trans. Inf. & Syst. (Invited Survey Paper)* **E82-D**(3), pp. 558 – 567, 1999.
6. T. Naemura and H. Harashima, "Real-time video-based rendering for augmented spatial communication," in *SPIE Visual Commun. and Image Process. (VCIP'99)*, vol. 3653, pp. 620 – 631, 1999.
7. M. Hirose, "CABIN - a multiscreen display for computer experiments," in *Int'l Conf. Virtual Systems and MultiMedia (VSMM'97)*, pp. 78 – 83, 1997.
8. M. G. Perkins, "Data compression of stereopairs," *IEEE Trans. Commun.* **40**(4), pp. 684 – 696, 1992.