In Proceedings of the 11th International Conference on Artificial Reality
and Telexistence, Tokyo, Japan, pp. 42-47, December 2001.

# A Head-Tracked, Live-Video-Based Telexistence System Using a Fixed Screen

Yasuyuki Yanagida[1], Shintaro Saito[2], Seiichiro Yano[3],
Taro Maeda[3], and Susumu Tachi[3]

1) ATR Media Information Science Laboratories
2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288 Japan
*yanagida@atr.co.jp*

2) IBM Japan, Ltd.

3) School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033 Japan

## Abstract

Visual display systems using fixed screens, including Immersive Projection Technology (IPT) displays, have a merit that they can provide a stable world against the user's head motion. In spite of this merit, such display systems have not been used for "telexistence in real worlds", which requires accurate stereoscopic view of live video image. We have proposed a method to realize a lived-video-based, real-time telexistence visual system with a fixed screen: to keep the orientation of the camera constant while following the user's eye position, and to control the position and size of the video image for each eye on the screen in real time. We have also designed technical elements to compose the system, i.e., a constant-orientation camera system and real-time 2D image manipulation (shifting and resizing) subsystem. In this paper, we describe the design and implementation of the entire system that realizes fixed-screen-based telexistence, which inherently has an ability of showing a stable remote world.

**Key words**: Telexistence, Stereoscopic Display, Immersive Projection Technology, Fixed Screen, Motion Parallax

## 1. Introduction

There are various types of visual displays used for Virtual Reality (VR) applications. If we focus on the spatial relationship between the users' eyes and screens, these displays could be categorized in two types: head-mounted displays (HMD) [1] and fixed-screen-based displays. Head-mounted displays are fit for personal use, as they occupy only little volume in front of the user's face. Meanwhile, fixed-screen-based displays, including ordinary CRT displays, simple combination of a projector and a screen, Immersive Projection Technology (IPT) display systems (such as CAVE [2], CABIN [3], and COSMOS [4]) and various kinds of head-tracked displays (HTD) such as Responsive Work Bench [5], require significant space. However, they have a strong advantage that they can provide the user with a stable world, when the user moves (especially rotates) his/her head.

The stability of the displayed world in IPT systems was shown by Cruz-Neira et al. [2], in the context of the observed angular error of the displayed point, when the "calculated" head position (recognized by the system) differs from the actual one due to the tracking error or the system delay. Though they insist on the merit of using large screen placed at several meters apart from the user's viewpoint, the behavior of conventional CRT and large screens are essentially the same, except for the scaling factor defined by the distance between the user's eye and the screen. We rather focus on the difference of the behavior of IPT systems from that of HMD, especially when the user rotates his/hear head. Now let us consider an HMD-based system with non-negligible time delay. If the user begins to rotate his/her head to the right, the whole world might rotate sticking to the user's head, then after a moment, the image on the screen would begin to flow to the left to cancel the head rotation, and keep flowing to the left for a while after the user stops rotating the head. Thus the user might feel the world is shaking. On the other hand, IPT systems do not require the displayed image to be updated for the pure rotation about the observation point.

Nevertheless, it is still necessary to track the head motion and to display the appropriate stereoscopic image on the screen, if not, the displayed world would be distorted or dynamically deformed according to the user's head motion. Diner et al. [6] and Rolland et al. [7] derived the relationship between the perceived location of the displayed point and the translational motion of the viewpoints. The result shows that the world would be expanded, compressed, or sheared about the virtual plane that coincides with the screen. The user's rotational head motion also affects the perceived world, as it equivalently causes variation in the inter-ocular distance, thus resulting in the difference in depth perception. We calculated such distortion of the world [8], to find that the world would be distorted in radius, though the points in the world hardly move in tangential direction. This result again confirms the stability of the displayed world in IPT systems. Fig. 1 shows an example of distortion of the perceived world when the user moves (translates and rotates) the head, if the image obtained by an ordinary stereo camera set is displayed for each eye.

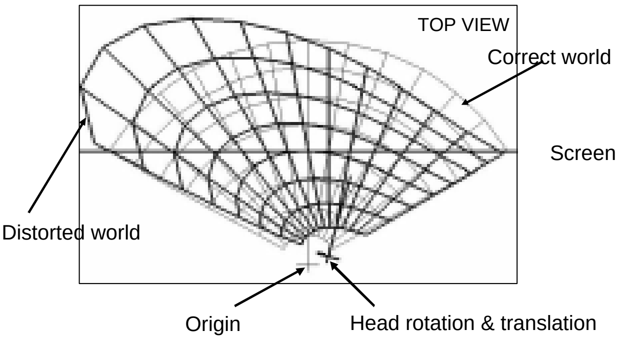


Fig. 1: Distortion of the perceived world when the user moves the head, if an ordinary stereo camera set is used. The screen is located at 2m forward from the original viewpoint, and the radial mesh is 50cm.

Next, we may as well mention the live-video-based approach for telexistence. One might say that we can view the remote world interactively in three-dimension by using image-based modeling and rendering (IBR) techniques [9][10], and insists on the merit of not requiring mechanical part in the system. However, we still think it important to display the remote world "as is", without any complicated computational processing. As we show in Fig. 2, IBR requires the reconstruction of three-dimensional world, or at least the extraction of the depth information explicitly or implicitly, to offer the freedom to specify the viewpoint and the projection parameter when generating the displayed image. This process of reconstruction or depth extraction is an ill-posed problem and it is quite difficult to get the perfect result. If the graphics image for the specified viewpoint was generated based on some erroneous reconstruction result, the user might feel uneasy to face the situation such as dust is floating in the air. Also the user might feel as if the world were made of blocks of discrete sizes, if the spatial resolution of the reconstructed three-dimensional world is not sufficient. In the live-video-based approach, we leave the three-dimensional reconstruction task to the human operator, without being meddled in by computers. This approach is based on a philosophy that man-machine systems should be designed so that the human being can exert their ability of perception and decision-making, while the machine would assist the human user, rather than being a delegate for the tasks.

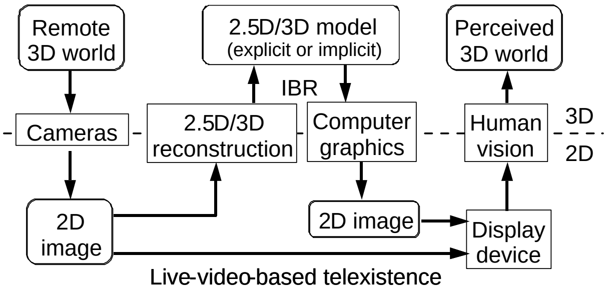


Fig. 2: Two approaches: live-video-based telexistence and image-based modeling/rendering

In spite of the merit of stability mentioned above, fixed-screen-based displays including IPT have not been utilized for live-video-based telexistence or telepresence in real environments, which requires appropriate stereoscopic video images corresponding to the operator's head motion. We found that the time varying, off-axis (or off-centered) projection required in fixed-screen-based displays has prevented these systems from being used for telexistence, because ordinary cameras only have fixed and symmetric fields of view about the optical axis. So far we have proposed a method to solve this problem [8] and designed several technical elements to compose the system [11], including a constant orientation camera system and real-time 2D image manipulation subsystem.

In this paper, we describe an entire prototype system, which can follow both of the user's rotational and translational motion. In Section 2, we briefly review our proposed method to realize a live-video-based telexistence system using a fixed screen. In Section 3, the basic design of the system is described. In Section 4, the implementation detail of the system is described. In Section 5, we report a simple evaluation experiment of the implemented system.

## 2. Principle

As we mentioned before, the problem to realize a live-video-based telexistence using a fixed screen lies in the required feature of the projection for the fixed screen: time variant and off-axis projection. It is easy to control the projection matrix in generating computer graphics images whenever staring to render a new frame (like `glFrustum()` in OpenGL [12]), but it is rather difficult for live video image obtained by an ordinary camera. We have submitted several ways to "equivalently" realize the dynamic off-axis projection. This can be reduced to the real-time control of the position and size of the video image. One way is to realize full functions by optics, i.e., to design and implement a shift and zoom optics that can be controlled in real time. Another way is to manipulate the image after capturing it, by simply discarding a portion of the image or the screen area (Fig. 3). We started with the latter method, as we do not need special camera optics with this method.

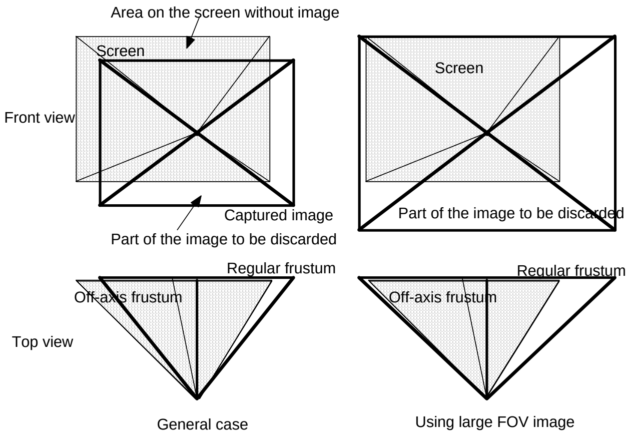
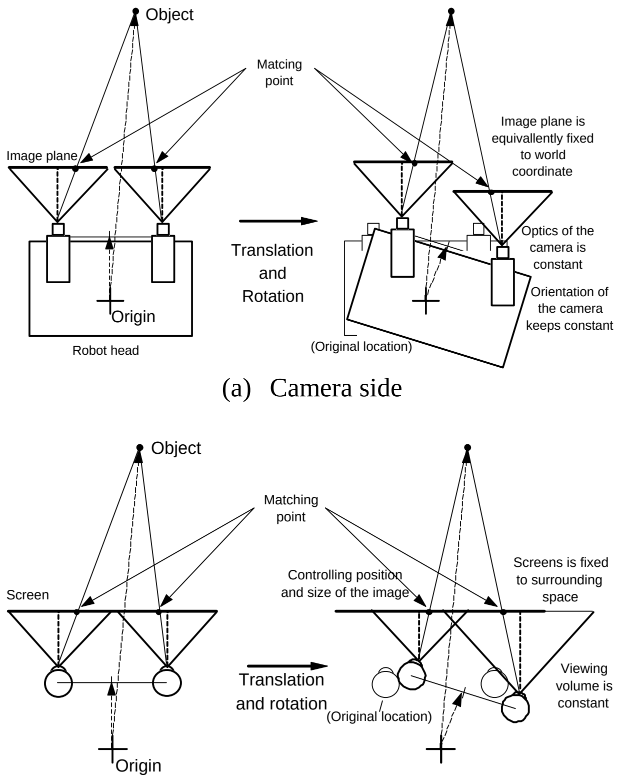


Fig. 3: Substituting a regular pyramid for off-axis projection


Object
Matcing point
Image plane
Image plane is equivallently fixed to world coordinate
Translation and Rotation
Optics of the camera is constant
Orientation of the camera keeps constant
Origin
Robot head
(Original location)
Object
Matching point
Screen
Controlling position and size of the image
Screens is fixed to surrounding space
Translation and rotation
Viewing volume is constant
(Original location)
Origin

(a) Camera side

(b) Display side

Fig. 4: A method of live-video-based telexistence using a fixed screen.

Using this substitution of viewing volume, we can configure the telexistence system as follows:

(1) Keep the orientation of the camera constant, while following the position of the user's each eye.

(2) Control the position and size of the displayed image, so that the resulting viewing volume at the display side is identical to that of the camera side.

This method is shown in Fig. 4. Further detail on the principle is described in [8].

## 3. System Design

Generally speaking, 6 degrees of freedom (DOF) is necessary to specify the user's head position and orientation, but 5 DOF is sufficient for the proposed method, as the orientation of the camera is fixed. In other words, each camera for left and right eye has 3 DOF for translation (no DOF for rotation, as the orientation of the camera is fixed), thus 6 DOF for two. There is one constraint to define the distance between two viewpoints; hence the result is 5 DOF.

So far we have designed and implemented a camera subsystem to follow the change of the user's eye position due to the head rotation [11]. This subsystem has a link mechanism to follow the user's yawing and rolling motion, whereas the orientation of the camera is kept constant. A pitching motion is not necessary, as the two cameras move in the same way.

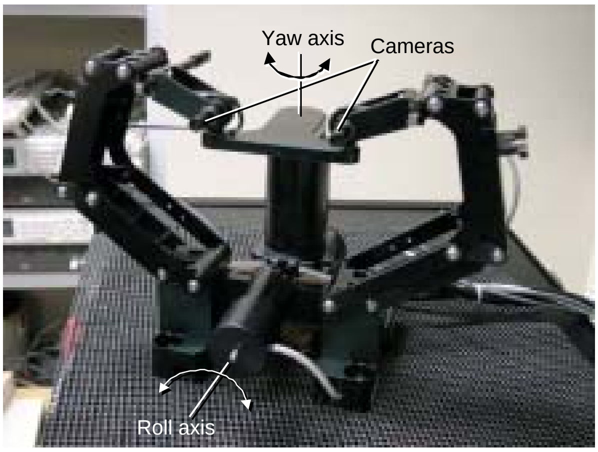


Fig. 5: Constant-orientation link mechanism equipped with CCD cameras

Since the constant-orientation link mechanism has 2 DOF, a complete system can be constructed if this link mechanism is carried on the stage, which can translate itself with 3 DOF. Here, however, we implemented 2 DOF out of 3, to support horizontal motion of the user's head. As the remaining 1 DOF for vertical motion is omitted, the current system cannot support the user's specific motion such as nodding or stretching. Nevertheless, this configuration can cover a wide variety of head motion caused by the trunk motion, such as looking into the object and body sway. By introducing the mechanism to support the user's head translation, we can provide the user with a correct motion parallax.

## 4. System Implementation

The entire system consists of a tracking subsystem, a camera subsystem, an image manipulation subsystem,

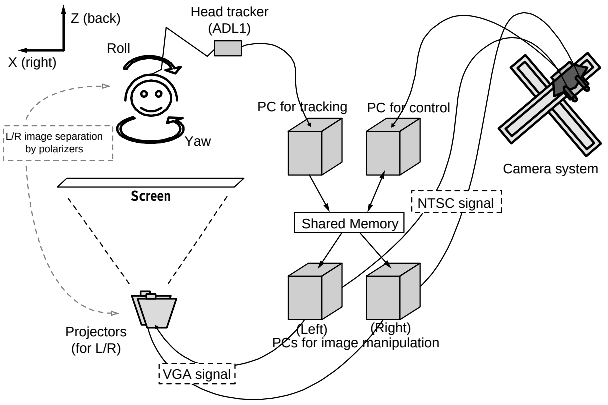


Fig. 6: System configuration.

and a display subsystem. The system configuration is shown in Fig. 6.

### 4.1 Motion tracking subsystem

We used Shooting Star ADL-1, a mechanical link device equipped with potentiometers, for head tracking. The output data from ADL-1 were sent to PC (Pentium-II 266MHz CPU, MS-DOS 6.2) through RS-232C serial communication line by 115.2 kbps. We obtained raw joint angle from ADL-1 and calculated the direct kinematics on the PC, as the CPU of the recent PC is far faster than the embedded one of ADL-1, so that we can obtain more accurate result (no approximation in algorithm) and higher sampling rate. We calculated the Cartesian position and orientation of the coordinate system attached to the user’s head, the origin of which is located at the midpoint of the user’s both eyes. The result was obtained in the form of 4 x 4 transform matrix, relative to the reference coordinate system. After calculating this matrix, we decomposed the rotation matrix to match the joint angle that can be fed to our link mechanism (to be described later). The final result was sent to other PCs in the system, through a shared memory interface board (Interface Corp. MemoLink PCI-4914). The loop frequency of the head tracking and kinematics calculation task was approximately 680Hz.

### 4.2 Camera subsystem

Fig. 7 shows the whole camera subsystem, which consists of 2DOF constant-orientation link mechanism and 2DOF linear sliders, which correspond to horizontal translation and rotation, respectively. The constant-orientation link mechanism mainly supports the variation of the position of each eye due to the user’s head rotation. The roll axis prevents the collapse of the binocular fusion, and the yaw axis plays a role in keeping consistent disparity, avoiding the distortion of the perceived world described in Section 1. The link is designed so that the position of the camera has offset from the axis (5cm forward and 12cm upward), aiming at supporting by itself the viewpoint motion due to the user’s natural head rotation. The link holds two small cameras (Toshiba IK-SM43H: 7mm in diameter, 1/4 inch CCD, 0.41M pixels, NTSC output) with compact lens (Toshiba JK-L04S, focal length f = 4mm), providing 39 degrees and 51 degrees for vertical and horizontal field of view, respectively. The link is driven by DC motors: Maxon Motor RE036-072 (70W) for yaw axis and RE025-055 (20W) for roll axis. The joint angle is detected by optical rotary encoders (Tamagawa Seiki OIH-35; 3000 pulses per rotation).

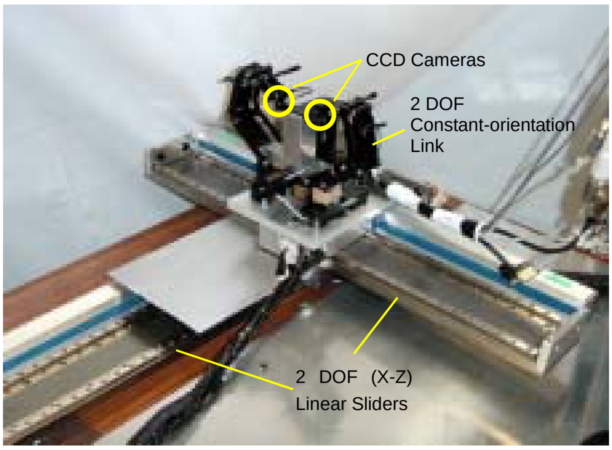


Fig. 7: Camera system.

We mounted this constant-orientation link on a 2DOF translation mechanism, which consists of two linear sliders (Yokogawa Precision LM110), combined in orthogonal direction. The movable range is 50 cm for side-to-side and 70 cm for back-and-forth.

We used a PC (Intel Celeron 333MHz CPU, MS-DOS 6.2) to control the camera system. This PC receives the desired joint angle data from the tracking PC through the shared memory interface (Interface Corp. MemoLink PCI-4913), and controls the mechanical system to follow the desired value. The given data set is calculated to fit the mechanical configuration in advance. Since our system does not require pitch angle, the rotation matrix obtained by the direct kinematics is decomposed to roll-yaw-pitch Euler angle, and then we simply ignore the endmost pitch angle. Roll and yaw angles are used to control the constant-orientation link, to determine the relative position of left and right eyes. Strictly speaking, pitch angle should be reflected to the vertical translation in accordance with the offset from the rotation axis, but this contribution is currently ignored, as we have not implemented vertical translation in the system.

For translation, the data fed to the linear sliders are calculated by subtracting the offset contribution of the constant-orientation link from the head position obtained by the direct kinematics calculation. Here we again ignore the vertical motion for the current system

The status of the mechanical system is measured by optical encoders, which is used for control loop. We used a simple PID algorithm to control each joint of the system, and the control output is sent to the mechanical system through D/A interface board. The sampling rate of the control program was approximately from 3.5 kHz to 4 kHz.

### 4.3 Image manipulation subsystem

We used a PC (Pentium-III 600MHz CPU, Windows 2000) to manipulate the video image for each eye. Each PC has a graphics board with video input and output (ASUSTeK AGP-V3800) so that the video image is processed locally on the graphics board, without passing through the system bus. The task executed here is to shift and resize the video image in real time. We used Microsoft DirectDraw API included in DirectX 7.0a. The video signal is captured by using DirectDraw VideoPort and directly output to the display, adjusting the position and size in real time by overlay function. Here we found it not sufficient to realize this feature by the driver software provided by the board vendor, so we directly controlled the video input processor (Phillips SA711A Enhanced Video Input Processor) implemented on the board, by using the protocol of I$^2$C bus [13]. Thanks to this configuration, we succeeded to control the position and size of the video signal, as well as updating the video image in 60 Hz, the field rate (not frame rate) of NTSC signal.

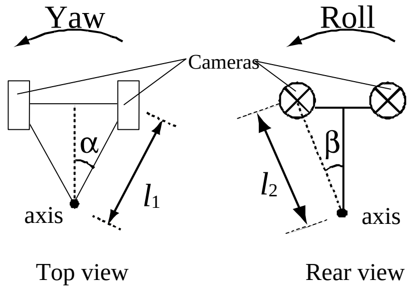


Fig. 8: Constants on length and angle for calculating shift amount and magnification of the video image.

The variables to be controlled are the shift amount and magnification ratio. Let the x-axis be right and z-axis be backward, then the shift amount ($S_x$, $S_y$) on the screen ($x$, $y$) and the magnification ratio $k$ is calculated as:

$$S_x = r_x \mp l_1 \sin(\alpha \pm \theta)$$
$$S_y = -l_1 (\cos\theta - \cos(\alpha \pm \theta))$$
$$k = 1 + \frac{r_z + l_2 (\cos\beta \pm \phi))}{L}$$

where $r_x$ and $r_z$ are the $x$ and $z$ element of the measured head position, $\theta$ and $\phi$ are roll and yaw angle, respectively, $L$ is the distance from the origin (initial midpoint of right and left eyes) to the screen, and $l_1$, $l_2$, $\alpha$, and $\beta$ means the length and angular constant depicted in Fig. 8. The double signs mean the upper corresponds to left eye and lower to right eye.

### 4.4 Display subsystem

Video images for right and left eyes are projected by compact projectors (NEC LT150J, DLP). We used popular polarizers to separate right and left video images. The stereoscopic image projected on the screen is shown in Fig. 9.

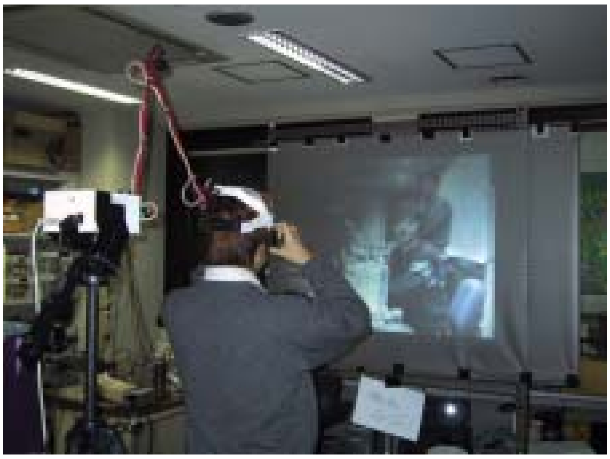


Fig. 9: Stereoscopic image on the screen.

## 5. Results

Fixed-screen-based stereoscopic live-video systems so far, which use ordinary pairs of fixed stereo camera, have forced the user to endure the unnatural motion parallax or distortion of the perceived world. Our prototype system, however, succeeded to eliminate such unnatural behaviors and to present the stable world as stable.

To verify the effect of the proposed method and the implementation result, we made a simple experiment on the perceived world. We asked subjects (8 male twenties) if they could perceive the distortion of the world on the following conditions:

(1) The prototype system is not active: i.e., it is used as an ordinary fixed stereo camera system. The subject is asked to move his head side-to-side and back-and forth.

(2) Identical to condition (1), except that the user is asked to close his eyes while moving.

(3) The system is fully active.

All subjects perceived the distortion of the displayed world with condition (1). With condition (2), most subjects did not perceived the distortion if the motion is relatively small, while some subjects reported the distortion when they wildly sway their trunks. Even though, this perception was not at least prominent compared with the case of condition (1), where the subject was watching the world while moving his head. With condition (3), none of the subjects reported the distortion, while a few subjects noticed somewhat unnatural while moving their heads.

Comparing the result for condition (1) and (2), we can confirm the preferred characteristics of fixed-screen-based systems that they can provide stable worlds, and that the problems of conventional systems lies in that the dynamic deformation of the perceived world, rather than the static perception of world distortion. The result for the condition (3) might be a proof for our proposed method, though our system is still not perfect. Ideally we can offer the visual stimuli perfectly equivalent to that of everyday life when the subject moves the head, but there was some flaw on the prototype system, including mechanical time delay, overshooting/undershooting due to the imperfection of the control parameter adjustment. The result, which the unnatural behavior was perceived only when the subject was moving, shows that the static error of this system was within the human user's threshold level.

## 6. Conclusion

We constructed a live-video-based telexistence system using a fixed screen, which can provide the user with a stable remote world. By supporting head translation as well as head rotation, the dynamic deformation of the perceived world, known as the inverse parallax problem, has been significantly reduced.

Future research will include the implementation of the remaining vertical axis, improvement of the performance, more quantitative evaluation of the proposed method and the prototype system, and applying this method to the field that requires highly stable and accurate 3D display, such as tele-surgery systems.

## Acknowledgement

This study was supported in part by a Grant for "Research for the Future Program #97I00401" from the Japan Society for the Promotion of Science.

## References

1 I. E. Sutherland: "A Head-Mounted Three Dimensional Display," *Proc. Fall Joint Computer Conference, AFIPS Conf. Proc.*, Vol. 33, pp.757–764 (1968).

2 C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti: "Surround-Screen Projection-Based Virtual Reality: The Design and Implementation of the CAVE," *Computer Graphics* (*Proc. SIGGRAPH '93*), pp. 135–142 (1993).

3 M. Hirose, T. Ogi, S. Ishiwata, and T. Yamada: "Development and Evaluation of Immersive Multiscreen Display 'CABIN'," *The Transactions of the Institute of Electronics, Information and Communication Engineers*, Vol. J81-D-II, No. 5, pp. 888–896 (1998).

4 T. Yamada, H. Tanahashi, T. Ogi, and M. Hirose: "Development of Fully Immersive Display "COSMOS" and Evaluation of Virtual Space Navigation," *Transactions of the Virtual Reality Society of Japan*, Vol. 4, No. 3, pp. 531–538 (1999).

5 W. Krüger, C-A. Bohn, B. Fröhlich, H. Schüth, W. Strauss, and G. Wesche: "The Responsive Workbench," *IEEE Computer*, pp. 42–48 (1995).

6 D. B. Diner and D. H. Fender: *Human Engineering in Stereoscopic Viewing Devices*, Plenum Publishing Corporation (1993)

7 J. P. Rolland and W. Gibson, "Towards Quantifying Depth and Size Perception in Virtual Environments," *Presence*, Vol. 4, No.1, pp. 24–49 (1995).

8 Y. Yanagida, T. Maeda and S. Tachi: "A Method of Constructing a Telexistence Visual System Using Fixed Screens," *Proceedings of IEEE Virtual Reality 2000*, pp. 117–125 (2000).

9 L. McMillan and Gary Bishop: "Plenoptic Modeling: An Image-Based Rendering System," *Computer Graphics* (*Proceedings of SIGGRAPH 95*), pp. 39–46 (1995).

10 R. Raskar, 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," *Computer Graphics* (*Proceedings of SIGGRAPH '98*), pp. 179–188 (1998).

11 Y. Yanagida, A. Mitsuhashi, T. Maeda and S. Tachi: "Implementation of Fixed-screen-based Telexistence Visual System," *Proceedings of the ICAT '99* (9th International Conference on Artificial Reality and Telexistence), pp. 123–130 (1999).

12 J. Neider, T. Davis, and M. Woo: *OpenGL Programming Guide*, Addison-Wesley (1993).

13 PHILIPS: *Data Sheet I2C Bus, SAA711A Enhanced Video Input Processor* (*EVIP*), (1998).