Interaction with Virtual Space using Real-time Gesture Recognition

Tsuyoshi KIMURA, Takeshi NAEMURA, Masahide KANEKO† and Hiroshi HARASHIMA

School of Engineering, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113, Japan
TEL : 03-3812-2111 (ext.6781) FAX : 03-5800-5797
tkimura@hc.t.u-tokyo.ac.jp

†KDD R&D Laboratories
2-1-15 Ohara, Kamifukuoka-shi, Saitama 356, Japan

Abstract

We have developed the RICUE (Realistic Immersive CommUnication Environment) system as a 3-D integrated and interactive information environment. This system is furnished with two orthogonal large stereo screens for image output and with three video cameras for image input. To realize the natural and smooth interaction between a user and the RICUE system, we examined a new method for the gesture recognition in which any special sensors are not attached to user's body. However the gesture recognition from image sequences is not so easy because of the multiple joint structure of a human body, as well as the occlusions, and the difficulties in the extraction of particular features from 2-D images. To cope with these problems, this paper examines a method to extract rough information about user's orientation in advance of the gesture recognition. We use this information as well as user's silhouette, a 3-D skeleton model and color information to recognize user's gesture. We adopt a rather simple processing method to realize real-time operations which compares user's silhouette with a simplified silhouette synthesized from a 3-D skeleton model. The experimental results show that real-time gesture recognition of the upper half of the body is successfully carried out even if it turns to any direction.

Key Words: RICUE, Interaction, Gesture Recognition, Real-time, Estimation of Position and Orientation

1 Introduction

In the field of Virtual Reality (VR), it is important to realize a method for the smooth interaction between a user and a virtual space. The recognition of human gesture is one of the important key technology to realize such smooth interaction. In addition to this, the gesture recognition method is useful to construct an automatic input system for human action which makes it easy to generate animated CG characters [1].

The methods to measure human actions or posture can be roughly classified into two categories. One is a method in which some devices such as Fastrack sensors and DataGloves are attached on our bodies. This method is useful to extract human motion precisely. However, it is not comfortable for us to wear such devices on our bodies, and our motion is restricted by them. The other is a method to extract human information from images. In this approach, [2] and [3] extracted human posture from a still image. The 3-D posture is estimated from a 2-D image and the structural information of a human body. However it is difficult to apply them to a human interface system because we need to extract features such as joints by manual operation, and we need a precise 3-D human model. The 3-D posture estimation using plural cameras[4][5][6] is difficult to realize the real-time estimation, because they require much computational cost. As an approach for real-time recognition, Wren has developed the Pfinder system[7]. Pfinder can recognize human gesture in real time by tracking hands, feet and head using col-

ors and contour information. However Pfinder cannot reconstruct 3-D skeletal structure of human body and imposes many restrictions on human motion. Iwasawa has proposed a real-time method for estimating human body posture from thermal images[8]. The method is exempted from the problems of lighting conditions and background colors since they utilize thermal images. It extracts some features using contour and genetic-algorithm-based learning procedure. It achieves high performance in the processing time, but some problems remain when detailed structure is needed because significant features contained inside of contours cannot be extracted.

Recently, an immersive virtual environment such as CAVE has been developed. The posture and gesture estimation methods mentioned above have not been designed for such an immersive type of information environment yet. However, to make the immersive feeling better, it is important to study how to utilize the gesture recognition in such an information environment.

In this paper, we propose a method to recognize hand gesture in real time even when a user turns to any directions. By using the proposed gesture recognition method, we realize the interaction between a user and a VR space in an immersive virtual environment.

2 RICUE as an immersive communication environment

In a system of immersive virtual environment such as CAVE developed in cooperation by EVL of Illinois States University and NCSA, the detection of user's position is usually carried out by Fastrack sensors. However, more flexible communication between a user and a virtual environment should be studied. In this paper, we investigate a new way for the interaction between a user and a virtual environment RICUE (Realistic Immersive CommUnication Environment)[10].

The RICUE is furnished with two orthogonal large stereo screens for image display and with three video cameras for image input. It has two chromakey screens at the opposite sides of display screens to process input video images. The space surrounded by screens and chromakey screens is a two meters square. We can also use the Fastrack sensors in this space. Moreover, some loud speakers are arranged to construct a virtual sound space. The appearance of the RICUE is shown in Fig.1 and its components are shown in Fig.2.

The human gestures estimated from images captured by video cameras will make it possible to utilize the RICUE system as a communication platform between immersive media using ATM, and so on.

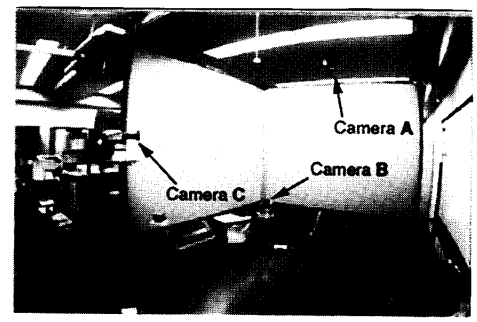


Fig.1 Appearance of RICUE.

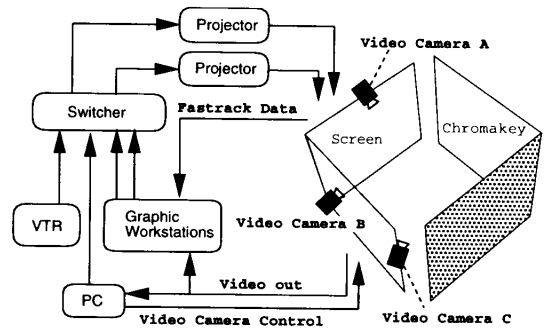


Fig.2 Components of RICUE.

3 Real-time gesture recognition

Estimation and recognition of user's gestures are carried out in two steps. Here, a user is allowed to turn to any directions. First, the orientation of user's body is roughly estimated. Secondly, the gesture is recognized from the estimated orientation of user's body. In both steps, we intended to realize the real-time processing.

3.1 Orientation estimation

3.1.1 Method

In the step of estimating orientation of user's body, the body part is extracted under the assumption that the motion of arms can be regarded as noise.

In this method, we approximate the human body by an ellipse pole model as shown in Fig.3, and estimate a user's orientation from width of the model on a image plane. This method was originally proposed

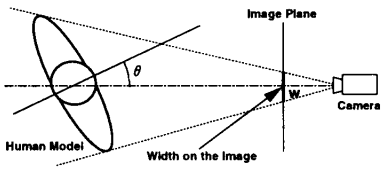


Fig.3 Top view of ellipse pole model of body.

in [9], and we simplified it to achieve the real-time estimation using just a single camera. About this model, the value θ means the orientation of a user. If the distance between the camera and a user is long enough, the width of a body part on the image:w is formulated by Eq.(1). The values a and b are measured at the scaling process.

$$w = \sqrt{a \sin^2 \theta + b \cos^2 \theta} \qquad (1)$$

We cannot judge whether a user turns right or left only from w. Therefore we utilize luminance of the facial area of an image. The flow of processing is composed of the following five steps. Figure 4 shows an example of processed image.

(1) Human body is segmented from the input image. In this process, the current input image and the pre-stored background image are compared.

(2) The scaling parameters are estimated from an image sequence in which a user turns a full circle in front of a camera.

(3) The regions corresponding to human arms are omitted.

(4) Estimate the orientation of a user from the width of the body region.

(5) Determine whether a user turns to right or left by comparing average luminance of right and left sides of facial area.

3.1.2 Experimental results

This estimation method works well and gives the rotational angle at the precision of about 15 degrees for all of the following three cases ; standing up straight, stretching both arms, and moving one's arms freely. As an experiment, a user turns a full circle at a constant rate from the start position 0 degree. Figure 5 shows the results of estimation without applying the right and the left judgement. We find that the orientation of a user is estimated roughly and the error for user's orientation is less than about $\pm 15°$.

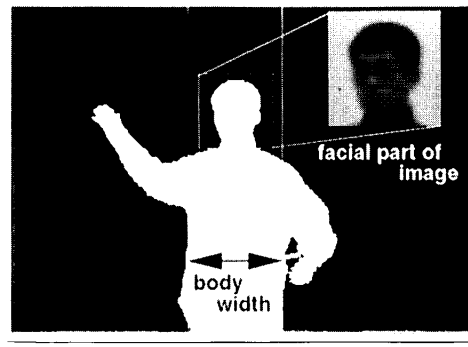


Fig.4 Estimation of orientation.

For the gesture recognition, we need rough information about user's orientation and restrict the range of angle from $-90°$ to $90°$. Thus we quantize the orientation angle to a few levels which contain less error and are needed for gesture recognition. We choose these levels as $0, \pm 30, \pm 45, \pm 60, \pm 90$ degrees. As an experiment for this method, a user turns from $-90°$ to $90°$ at a constant rate. Estimation results are shown in Fig.6. We cannot find the disorder of the errors by quantizing the orientation angle to a few levels. And we find that the orientation of a user is estimated right. Therefore we can utilize this method for the part of the system of gesture recognition.

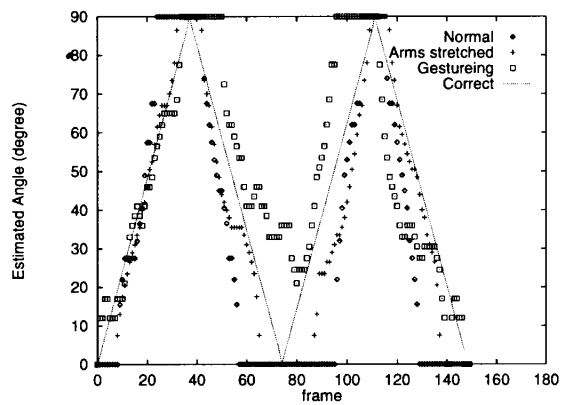


Fig.5 Estimated result without right or left judgement. ($0° \leq \theta \leq 90°$)

3.2 Arm posture estimation

In this paper, the arm posture estimation means to estimate all angles of joints of upper half of user's body from a image sequences.

3.2.1 Method

The skeleton model used in this paper is shown in Fig.7, which is very simple but can express a human posture well. Each joint of this model has two or

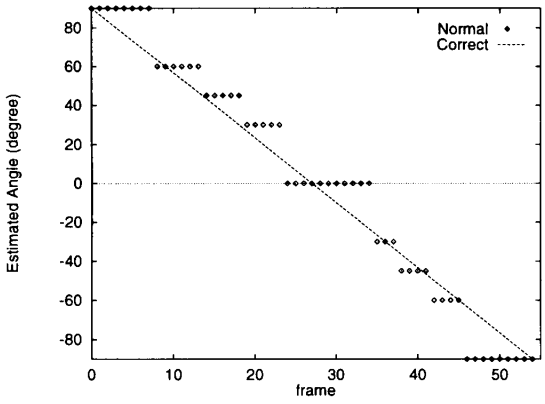


Fig.6 Estimated result for orientation. ($-90° \leq \theta \leq 90°$)

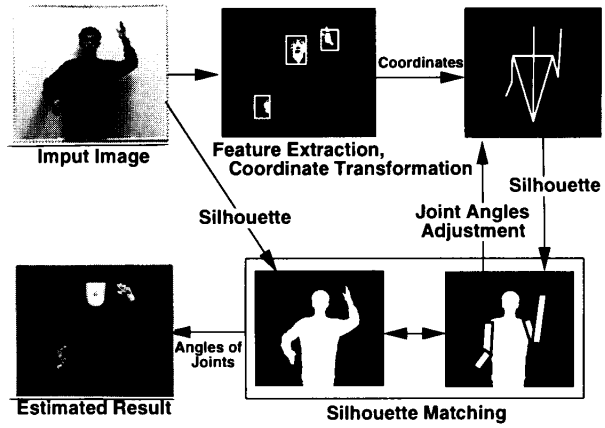


Fig.9 The flow of processes.

three freedom. First of all, we need the images of background and referential pose of a user to construct the skeleton model. The referential pose means the posture in which a user stands upright with stretching his both arms horizontally, which is shown in Fig.8.

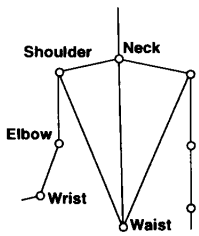


Fig.7 Skeleton model.

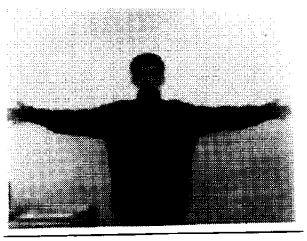


Fig.8 Referential pose.

The arm posture is estimated using a skeleton model of a human body, color information, silhouette and the information about a user's orientation. First, the positions of user's head and hands are tracked using the color information. The rough posture is estimated in advance using the relation of user's head and hands positions. This result will help to carry out the process of arm posture estimation using silhouette matching efficiently. We improve the processing speed by carrying out the extraction of skin color area and silhouette in the restricted image area, which is adjacent to area used for the matching operation in the previous frame. The flow of processes is as follows and illustrated in Fig.9.

(1) The size of skeleton model is adjusted to that of referential posture. The positions of head and both hands are utilized for scaling, estimating relative positions of shoulders to the position of head. The length of each arm (upper and lower arms) are adjusted. The silhouette of body part (shown in Fig.10 (a)) is kept in a memory through this process.

(a) Part of body. (b) Simplified silhouette.

Fig.10 Synthesis of simplified silhouette.

(2) Extract the areas of hands and face by color information in the present frame, when those positions in the previous frame is referred. Calculate the coordinates of hands and face in the image plane and transform to the coordinates in the model space.

(3) Turn the skeleton model by the information of orientation, which is extracted in the process of the orientation estimation.

(4) We call the positions of hands estimated in the second step "targets." Decide the angle of elbow joint so that the distance from a shoulder to a hand may agree with the distance from a target to a shoulder. Next, decide the angle of shoulder joint so that the vector from a shoulder to a hand may agree with the vector from a shoulder to a target. Thus rough arm posture is estimated.

(5) Search the posture, which gives the simplified silhouette approximating the silhouette of human body in an input image best, by moving joints which are independent of user's head and hands

positions. All angles are decided in this step. We generate the simplified silhouette by adding the silhouette of arm part which is estimated by link model of a human body to the silhouette of body part which is extracted from a referential posture (Fig.10 (b)). The silhouette shape of body part must be changed using information of its orientation.

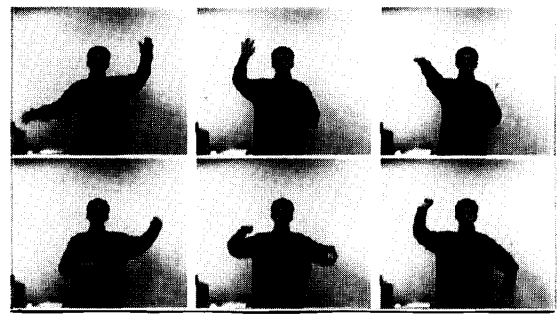
We apply the process of silhouette matching at two different resolution levels ; rough estimation at low resolution and more precise estimation at higher resolution. We can improve the real-time performance by adopting this matching method. Furthermore, since we cannot obtain the depth information from 2-D images, the depth estimation is performed. We divide the depth value into 3 levels because the position of a hand along depth is restricted by a skeletal structure of the upper half of body. The depth is determined to minimize the difference between the user's silhouette in the input image and the simplified silhouette synthesized by the link model.

3.2.2 Experimental results

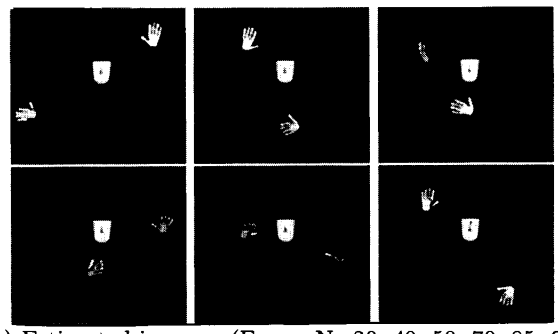
The arm posture estimation has been carried out on the image sequence which starts from the referential pose. Image size is 310 × 230 pixels. Here, a user in this sequence stands with facing to front. Thus the results of orientation estimation are not referred. Figure 11(a) shows input images while Fig.11(b) shows synthesized images based on estimated results. Comparing Fig.11(a) to Fig.11(b), we find that the gestures in the input images are reflected in the estimated images.

To assess the performance, we examined the errors between real value and estimated one. Though we should examine all angles of joints, we pick up the position of elbow in the image, because it is difficult to measure all angles of joints from 2-D image. The result is shown in Fig.12. In these figures, "Estimated" means the position of elbow in the image calculated through the processes of arm posture estimation. "Real" means the position measured manually in the image. "Right" and "Left" means which side is assessed. It can be seen that the position of elbow is estimated in error less than a few pixels, though we don't extract features precisely.

The processing speed of arm posture estimation is about 13 frames/sec using SGI Indigo^2 IMPACT (R10000, 195MHz). We can utilize this system as an

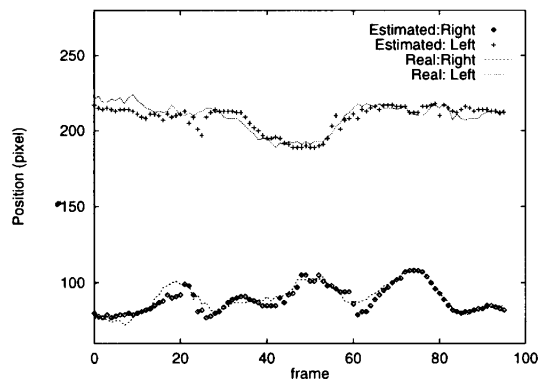


(a) Input images. (Frame No.20, 40, 50, 70, 85, 95)

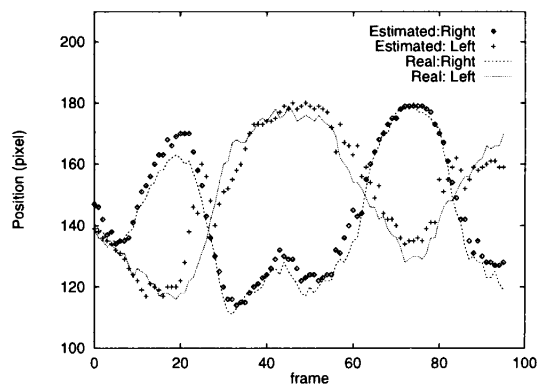


(b) Estimated images. (Frame No.20, 40, 50, 70, 85, 95)

Fig.11 Experimental results.



(a) Horizontal position of elbow.



(b) Vertical position of elbow.

Fig.12 Estimated results for the position of elbow.

interactive interface with this speed. However we have to speed up the processes for more useful real-time applications.

3.3 Gesture recognition

In this section, we describe the gesture recognition method, which means the arm posture estimation based on the result of the orientation estimation. The user's gesture is recognized even if one turns to any directions by combining both processes.

3.3.1 System

The overview of the system for the gesture recognition is shown in Fig.13. Images captured by a camera or stored in the disk are inputted into workstation A in which the orientation estimation process is running and workstation B in which the arm posture estimation process is running. The estimated results of orientation is sent to the next process, that is arm posture estimation, where the processing is carried out using images and information of orientation as resources. Estimated results from the arm posture estimation, which is angles of joints, are sent to a displaying process where 3-D human model is drawn. The 3-D human model as an estimated result is displayed on the screen of monitor or RICUE. We are investigating the integration of this system with the process of position estimation described in Section 4.

The processing speed of the estimation of orientation, position and model drawing is real-time, and that of arm posture estimation is around 13 fps (frames per second). Therefore each process needs to be executed in parallel to pursue the real-time performance. The total speed in the proposed system is nearly 13 fps.

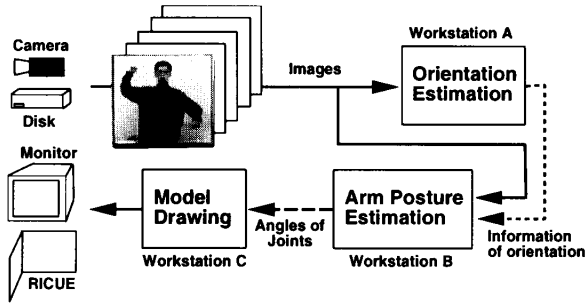


Fig.13 Gesture recognition system.

3.3.2 Experimental results

We have tested the performance of this system through the experiment using an image sequence in which a user makes gestures while turns to some direction. Figure 14 shows the estimated result of orientation from this experimental sequence. Figure 15(a) shows input images and Fig.15(b) shows synthesized images based on estimated results. The processing speed is around 13 fps as described in the previous section. It can be seen that the gesture recognition works well by comparing input and output images. And we can find that 3-D posture is recovered (though approximately) from only 2-D images using a skeletal structure model of a human body.

The results of gesture recognition contain much error when a user turns to 90°. The distinctive features about arms cannot be extracted from an image when a user turns to around ±90°. We have to restrict the angle of user's turning motion within around ±60°.

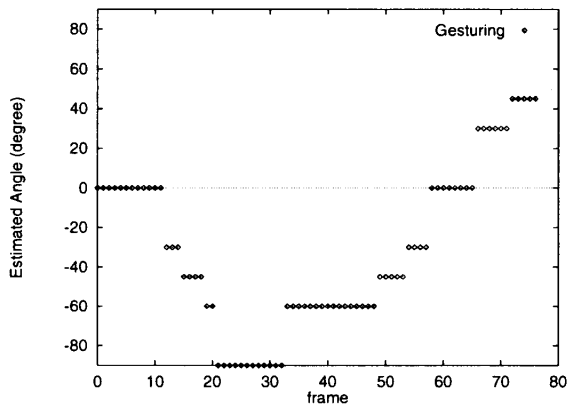
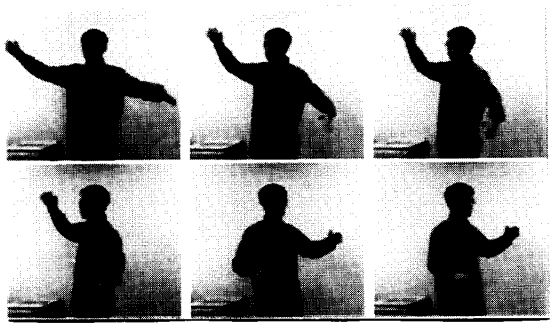


Fig.14 Estimated results of orientation.

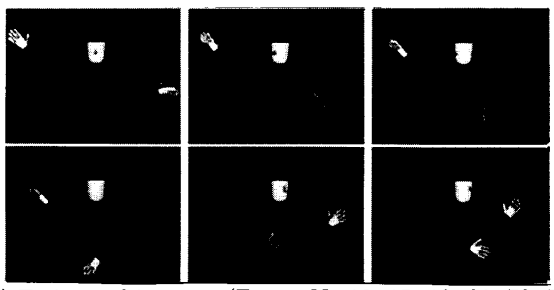
4 Estimation of user's position in the RICUE

The user's position is estimated by using a camera furnished in the RICUE. The coordinates of the head in the image plane is detected by the color information. Then these coordinates are transformed into the coordinates in the RICUE by using camera parameters measured in advance. We use camera A in the RICUE to estimate user's position because it is easy to extract information of user's position for camera A located on the top part of a screen, that is the upper position to a user. The example of the image captured by camera A is shown in Fig.16. The estimation error for user's position is less than 10cm.

The camera parameters are shown in Fig.18. We define the height of camera position as h_0, the angle with a vertical line as θ, the vertical angle of view as ϕ_y, the horizontal angle of view as ϕ_x and the height



(a) Input images. (Frame No.10, 15, 18, 21, 70, 74)



(b) Estimated images. (Frame No.10, 15, 18, 21, 70, 74)

Fig.15 Experimental results.

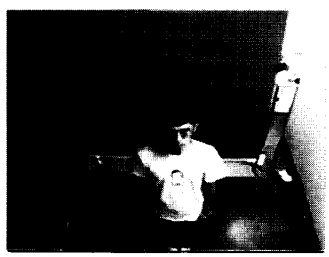


Fig.16 Image captured by camera A.

of user's head position as h_m. About the angle between the center axis of camera and the vector which is from the camera to the head of a user, the vertical and the horizontal angles are defined as α_x, α_y, respectively. We also define a frame of reference in RICUE as X-Y-Z, a frame of reference in the image plane which is captured by camera as x-y. The real position of a user in the frame of reference shown in Fig.17 : Z_{real} and X_{real} are formulated as Eqs. (2) and (3), respectively.

$$Z_{real} = (h_0 - h_m)\tan\alpha_y \quad (2)$$

$$X_{real} = 1.0 + p\tan\alpha_x \quad (3)$$

where $p = \sqrt{(h_0 - h_m)^2 + Z_{real}^2}$.

We define the abscissa and the ordinate of head position in a image plane as x_h, y_h, respectively, and define the horizontal size of image as x_{max} and the

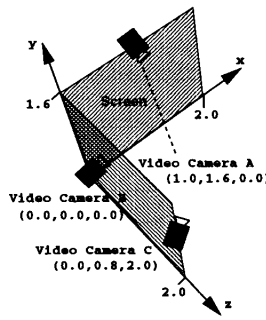


Fig.17 Axes and camera positions in RICUE.

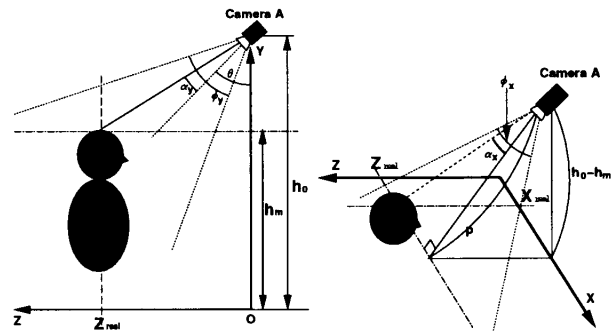


Fig.18 Camera parameters.

vertical size as y_{max}. Then α_x and α_y are represented by the following equations.

$$\tan\alpha_x = \left(1 - \frac{2x_h}{x_{max}}\right)\tan\frac{\phi_x}{2} \quad (4)$$

$$\tan\alpha_y = \left(1 - \frac{2y_h}{y_{max}}\right)\tan\frac{\phi_y}{2} \quad (5)$$

5 Interaction with virtual space

The interaction between a user and a virtual space in the RICUE can be realized by recognizing user's gesture in real time. To confirm whether the user's gesture is recognized correct or not, we construct the virtual environment which displays the image of human character responding to the user's direction and gesture like a mirror. Figure 19 shows the avatar on the screen of the RICUE which plays the same gesture as a user's gesture. We can send the avatar of a user to another immersive virtual environment on the network by utilizing this method. The information to be sent is just the estimated angles of joints. Thus a spatial communication or a cooperative work will be realized by users who are apart from each other.

We realize another application that a user in the RICUE can walk through the VR space using gesture.

In this system, the position of a user is converted into some commands; for example, Proceed, Back, Turn to Right and Turn to Left. We can interact with the VR space without attaching any sensors. We intend to improve this application by assigning more commands to each of different gestures, then we can handle some VR objects interactively.

Furthermore, the rough positions of user's hands in this space can be estimated by the reconstruction of rough 3-D skeletal structure and the estimation of user's position in the RICUE. We can realize the spatial interaction with synthesized virtual world for future by combining these technologies.

Fig.19 Application of gesture recognition.

6 Conclusion

We have developed the RICUE system for the communication between a user and a virtual space. To realize the smooth interaction with this system, we have proposed a new method for real-time arm posture estimation. We can recognize the gesture of a user who turns to any directions using the estimation result of rough orientation of user's body. Experimental results shows the possibility of the interaction between a user and a synthesized virtual world by the proposed real-time gesture recognition method.

For future, we plan to combine the position estimation of a user with the gesture recognition system to realize a spatial interaction between a user and the RICUE. For the total improvement of the system, we also plan to improve the precision of each process, that is orientation estimation, arm posture estimation and position estimation of a user. Thus we want to realize more smooth interaction and useful interface.

References

[1] J.Ohya, K.Ebihara, J.Kurumisawa, R.Nakatsu, "Virtual Kabuki theater: towards the realization of human metamorphosis system," *Proc. of 5th IEEE International Works. on Robot and Human Communication*, pp.416-421, Nov, 1996.

[2] T.Kurokawa and A.Shibata, "Recovery of three-dimensional human body posture from a single view," *Progress in Human Interface*, vol.4, pp.75-84, 1995. (*in Japanese*)

[3] Y.Kameda, M.Minoh and K.Ikeda, "A pose estimation method for an articulated object from its silhouette image," *IEICE Trans.* D-II, vol.J79-D-II, no.1, pp.26-35, 1996. (*in Japanese*)

[4] A.Satoh, S.Kawada, Y.Osaki and M.Yamamoto, "3-D model-based tracking of human actions from multiple image sequences," *IEICE Trans.* D-II, vol.J80-D-II, no.6, pp.1581-1589, 1997. (*in Japanese*)

[5] J.Ohya and F.Kishino, "Human posture estimation from multiple images using genetic algorithm," *12th ICPR*, pp.750-753, 1994.

[6] A.Azarbayejani, C.Wren, A.Pentland, "Real-time 3-D tracking of the human body," *Proc. of IMAGE'COM 96*, May, 1996.

[7] C.Wren, A.Azarbayejani, T.Darrel, A.Pentland, "Pfinder: real-time tracking of the human body," *Proc. of the 2nd International Conf. on Automatic Face and Gesture Recognition*, pp.51-56, Oct, 1996.

[8] S.Iwasawa, K.Ebihara, J.Ohya, R.Nakatsu and S.Morishima, "Real-time estimation of human body postures from monocular thermal images," *ITE Trans.*, vol.51, no.8, pp.1270-1277, 1997. (*in Japanese*)

[9] H.Mori, A.Ustumi, J.Ohya and M.Yachida, "Multiple-camera based estimation of human position and orientation," *ITE Technical Report*, vol.21, no.33, HIR97-43, 1997. (*in Japanese*)

[10] T.Kimura, T.Naemura, M.Kaneko and H.Harashima, "RICUE as a 3-D integrated interactive information environment," *Proc. of VRSJ97*, pp.234-237, Sep, 1997. (*in Japanese*)