MVA '96 IAPR Workshop on Machine Vision Applications, November. 12-14, 1996, Tokyo, Japan

Human-Robot Interface with Appropriate Frame Selection

Kentaro Hayashi Yoshinori Kuno Yoshiaki Shirai
Department of Mechanical Engineering for Computer-Controlled Machinery
Osaka University, Suita, Osaka 565, JAPAN
hayashi@cv.ccm.eng.osaka-u.ac.jp

Abstract

This paper presents a human-robot interface system that enables a user to move a robot by moving his hand. It adopts the technique of the 3D measurement based on the affine invariants from multiple views. In this system, we can choose a reference frame to obtain the 3D hand motion between the user-centered frame and the world-fixed frame. With the former frame, the system interprets the user's relative indication with respect to his body; with the latter frame, it interprets the indication with respect to the world. Selecting one of these two interpretations depending on the situation realizes easier operation of the robot. Experimental results confirm the usefulness of the system.

1 Introduction

Friendly human interfaces are indispensable to realize robot systems working in the real world. Since hand gesture is one of the important means of communication for humans, research interests have been increasing to recognize hand gestures and to use them for human interface[1, 2, 3].

We have also proposed a vision-based human interface system using multiple view affine invariance[4, 5]. It does not need any camera calibration. Moreover, it allows us to choose a reference frame between the one attached to the user's body(user-centered frame) and that fixed to the world(world-fixed frame) for gesture interpretation. If we choose the user-centered frame, the 3D hand motion is interpreted with respect to the user's body. If we move our hand forward, the motion is considered to indicate the forward direction regardless of the body position. However, if we choose the world-fixed frame, the same gesture is interpreted to indicate the direction shown by the hand motion in the world coordinate system.

In the previous system, we mainly moves 3D CG objects by hand gestures. In this case, we can easily realize both frame interpretations because we know all 3D spatial information of target objects. In addition, we can usually choose a reference frame in advance and do not need to change it.

In this paper, we extend the system so that we can move real world objects, such as a mobile robot and a camera with a pan-tilter on the robot by hand gestures. To do this, the system needs to obtain 3D positional information of the targets. Also, such situations often occur that we want to change the reference frame during operation. For example, the world-fixed frame interpretation is appropriate when we ask the robot beside us to go to some direction. We can indicate the desired direction by the hand. However, if the robot goes out of the door and disappears from our view, we cannot use this interpretation. We need to watch images sent from the camera on the robot and to control the robot. In this case, the user-centered frame interpretation is appropriate because we can feel as if we were at the robot's position.

This paper presents the configuration of our human interface system, and the frame selection method. Operation experiments using a mobile robot is also described.

2 Multiple View Affine Invariance

This section briefly describes the multiple view affine invariance, on which our system is based[6, 7]. We adopt weak perspective projection as our camera model. It is a good approximation of general projective transformation when the object-camera distance is much greater than the extent of the object[8].

Suppose we have a set of five 3D points $\mathbf{X}_i, i \in 0, \ldots, 4$. We use four of these points to establish a basis vector $\mathbf{E}_i$ with origin $\mathbf{X}_0$ (see Fig.1). In this basis, the fifth point, and any other point, is given by

$$\mathbf{E}_i = \mathbf{X}_i - \mathbf{X}_0, \quad i \in 1,2,3 \tag{1}$$

$$\mathbf{X}_4 = \mathbf{X}_0 + \alpha\mathbf{E}_1 + \beta\mathbf{E}_2 + \gamma\mathbf{E}_3 \tag{2}$$

for some α, β, γ. The coefficients α, β, γ do not change under any 3D affine transformations, thus called *affine invariants*. They can be viewed as the invariant coordinates of point $\mathbf{X}_4$ in the basis and are related to the 3D structure of the object. Since weak perspective projection is a linear transformation, it is possible to find α, β, γ in terms of the projected points. For a particular view, we obtain

$$\mathbf{x}_4 - \mathbf{x}_0 = \alpha\mathbf{e}_1 + \beta\mathbf{e}_2 + \gamma\mathbf{e}_3 \tag{3}$$

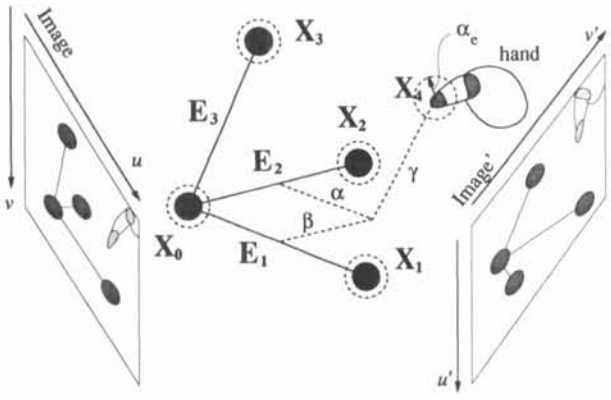


Figure 1: Affine basis

where $\mathbf{x}_0, \mathbf{x}_4$ are the projected points and $\mathbf{e}_1, \ldots, \mathbf{e}_3$ are the projections of the basis vectors. This means that we can derive two equations with three unknowns from the five point locations in a single 2D image. If we obtain a second view with known correspondence to the first view, we can get another equation as follows.

$$\mathbf{x}'_4 - \mathbf{x}'_0 = \alpha\mathbf{e}'_1 + \beta\mathbf{e}'_2 + \gamma\mathbf{e}'_3 \qquad (4)$$

The symbols with primes denote the same vectors in Eq. (3) for the second view. Assuming that this equation (4) is linearly independent to Eq. (3), we have four overdetermined equations In this condition, we can obtain the invariants α, β, γ with the standard method of least squares. Details are found in [9, 4].

3 System Configuration

We have developed an experimental human-robot interface system using multiple view invariants. Fig.2 shows the system configuration. The system consists of an engineering work station (SUN WS), an image processing board(Tracking Vision[10]), an active camera system composed of two cameras on a computer-controlled mobile platform, and a robot. A camera with a computer-controlled pan-tilter is equipped on the robot. We can move the robot and the camera by hand gestures.

We use special marks for the features to calculate invariants to make feature tracking easy and reliable. Fig.3(a) shows the marks on the user. The four marks on the body are used for the reference frame and the two on the arm for the indication of direction. Fig.3(b) shows the marks on the robot. The system obtains the position and orientation of the robot by tracking these marks. The robot and the camera on the robot are connected to the system with radio modems. The input image from the camera is transmitted wireless to the monitor.

Fig.4 shows the overview of the system. The mobile platform moves to keep the marks on the user in the camera fields of view.

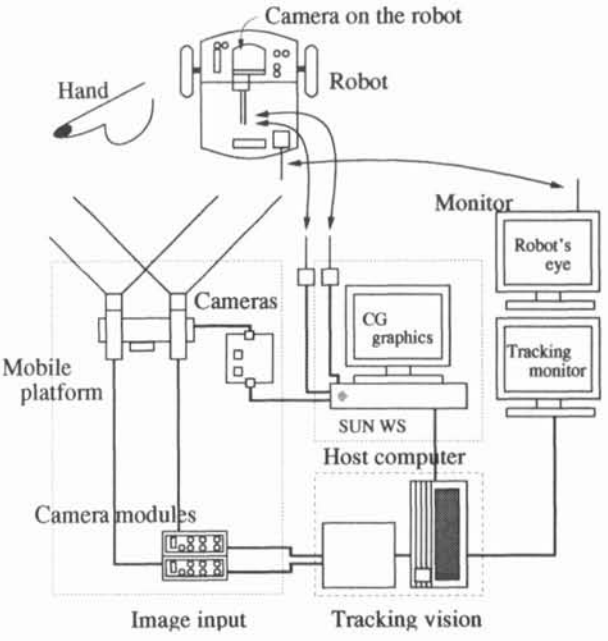


Figure 2: System configuration

4 Operation of the Robot

4.1 Frame selection

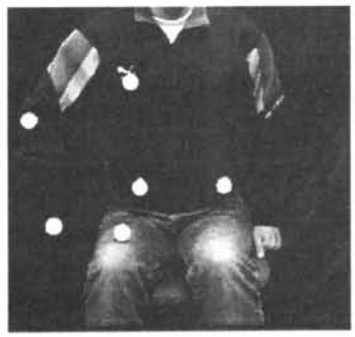
As described in Introduction, the appropriate frame should be selected depending on the situation. Table 1 shows the possible relationships between the user and the robot, and the appropriate frames corresponding to these situations. When the user cannot see the robot(a), it is impossible to use the world-fixed frame because the user cannot seize the relation of the robot to himself. On the other hand, when the robot and the user are close(c), it is natural to operate the robot in the world-fixed frame regardless of the orientation of the robot because the user can easily seize the position and orientation of the robot in the world. The situation (b), in which the user can see the robot, but the robot is distant from the user, is the case between the above two situations. In this case, the selection depends on the preference of the user.

We use the following frame selection method based on Table 1. In the case that the user and the robot are in the camera field of view, the system selects the world-fixed frame. This corresponds to the situation where the robot is close to the user. In the other cases, such as that the robot is not in the camera field of view, the system selects the user-centered frame.

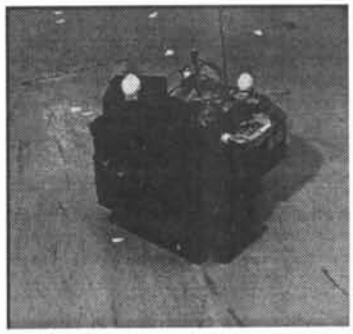
The following sections describe the details of robot operation in each frame.

4.2 Operation in the world-fixed frame

Let $\mathbf{U}_p$ denote the direction vector indicated by the user's arm, and $\mathbf{R}_p$ the heading direction vector



(a) User



(b) Robot

Figure 3: Features on the user and the robot

of the robot projected on the X-Y plane(see Fig 5). These vectors are obtained by extracting and tracking the two marks on the arm and the two marks on the robot, respectively. The system examines the angle between the vectors. If the angle exceeds some threshold, the system rotates the robot so that $\mathbf{R}_p$ coincides with $\mathbf{U}_p$.

When we use the world-fixed frame, we need to establish four reference points in the world. However, we can use the four points on the user's body for this purpose if the following two conditions are satisfied.

1. X-Y plane of the basis runs parallel to the plane on which the robot moves.

2. The basis does not change while the system controls the robot.

The first condition is satisfied if the user is on a chair as in Fig.4. The second condition does not mean that the user cannot move during operation. The world-fixed frame interpretation requires that both arm's and robot's directions should be measured with respect to the same reference frame when a certain hand gesture is interpreted. However, it does not require that the reference frame should be the same through all interpretation times. Thus, the condition is always satisfied if the system takes the

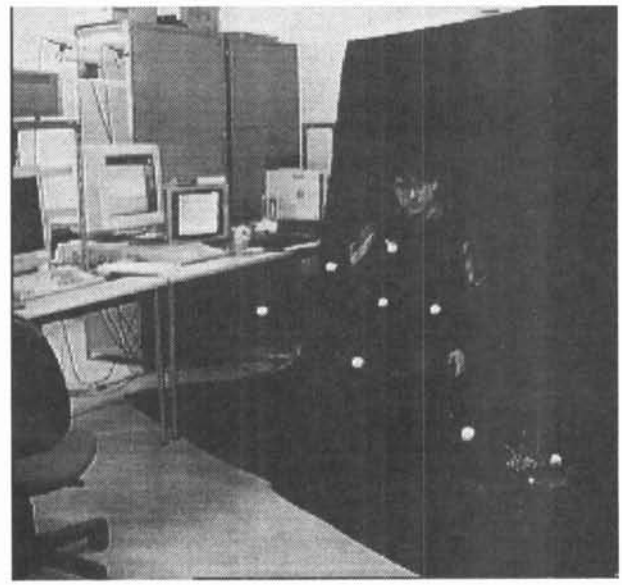


Figure 4: Overview of robot operation

Table 1: Operating frame selection

Operating situation	Operating frame
(a) The user cannot see the robot	user-centered frame
(b) The user can see the robot, but the robot is distant from the user	either frame preferred by the user
(c) The user can see the robot and the robot is close to the user	world-fixed frame

images of all the features on the body, the arm and the robot at the same time.

4.3 Operation in the user-centered frame

In the user-centered frame case, we can operate the robot and the camera on it. We tessellate regions around the right arm as shown in Fig.6. We can select one of the actions of the robot and the camera by indicating one of these partitioned 3D regions by the hand.

When the hand position is nearer to the origin than r_{c2}, the indication is recognized for controlling the camera platform. The octant spherical region is divided into 9 parts by $\theta_1, \theta_2, \psi_1$ and ψ_2. Each part is associated to a certain action of the camera platform. The circular region farther than r_{c2} is allocated for controlling the robot. The quadrant is divided into 6 parts by $\theta_1, \theta_2, r_{r1}$ and r_{r2}. Each part is associated to a certain action of the robot. The user can operate the robot and the camera by

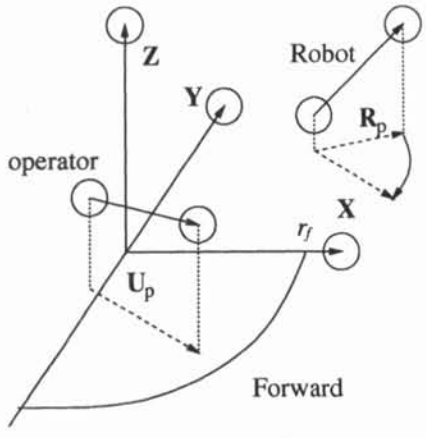


Figure 5: Positions and orientations of the hand and the robot in the world-fixed frame

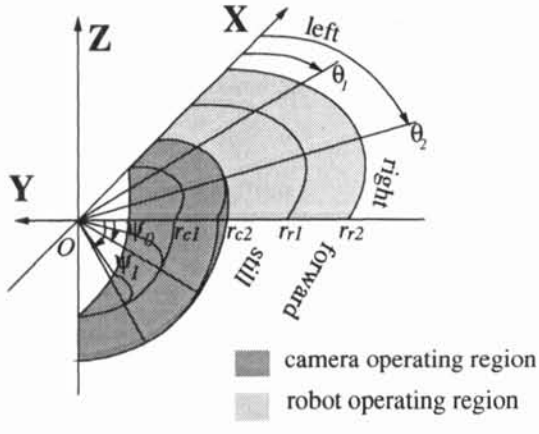


Figure 6: Tessellated regions for operating both the robot and the camera

indicating these parts by the hand.

5 Experiments

We made several experiments of robot operation of the following sequence. First we rotated the robot close to the user to the direction indicated by the arm's direction. The system interpreted the gesture in the world-fixed frame. In the experiments, the system rotated the robot in ± 10 degrees' error. Then, the robot moved to the indicated direction. When the robot went out of the system's camera field of view, the reference frame was automatically changed to the user-centered frame. The user was able to operate the robot and the camera by hand gestures in the user-centered frame. Experimental results show that the frame selection is useful in realizing a user-friendly human-robot interface.

6 Conclusion

We have proposed a human interface system to operate a robot by hand gestures. The system selects one of the two frames which we call the user-centered frame and the world-fixed frame, according to the situation as a reference frame for 3D gesture interpretation. The experimental results show that the proposed system can offer a useful means of communication between humans and robots.

This work was supported in part by the Ministry of Education, Science, Sports and Culture under the Grand-in-Aid for Scientific Research, No.07670492, Artificial Intelligence Research Promotion Foundation, Electro-Mechanic Technology Advancing Foundation, and International Communication Foundation.

References

[1] Roberto Cipolla and Nicholas J. Hollinghurst. Human-robot interface by pointing with uncalibrated stereo vision. *Image and Vision Computing*, pp. 171–178, 1996.

[2] Roberto Cipolla, Yasukasu Okamoto, and Yoshinori Kuno. Robust structure from motion using motion parallax. In *Fourth Intl. Conf. on Computer Vision*, pp. 374–382, 1993.

[3] Thomas Baudel and Michel Beaudouin-Lafon. CHARADE: Remote control of objects using free-hand gestures. *Communications of the ACM*, Vol. 36, No. 7, pp. 28–35, July 1993.

[4] Yoshinori Kuno, Kentaro Hayashi, Kang Hyun Jo, and Yoshiaki Shirai. Human-robot interface using uncalibrated stereo vision. In *Intl. Conf. on Intelligent Robots and Systems*, pp. 525–530, 1995.

[5] Kang-Hyun Jo, Kentaro Hayashi, Yoshinori Kuno, and Yoshiaki Shirai. Vision-based human interface system with world-fixed and human-centerd frames using multiple view invariance. *IEICE Trans. Information and Systems*, Vol. E79-D, No. 6, pp. 219–228, June 1996.

[6] Joseph L. Mundy and Andrew Zisserman, editors. *Geometric Invariance in Computer Vision*. MIT Press, 1992.

[7] Sven Vinther and Roberto Cipolla. Towards 3D object model acquisition and recognition using 3D affine invariants. Technical Report CUED/F-INFENG TR136, Cambridge University Engineering Department, England, July 1993.

[8] D. W. Thompson and J. L. Mundy. Three-dimensional model matching from an unconstrained viewpoint. *IEEE International Conference on Robotics and Automation*, pp. 208–220, 1987.

[9] Yoshinori Kuno, Masao Sakamoto, Ken'ichiro Sakata, and Yoshiaki Shirai. Vision-based human interface with user-centered frame. In *Intelligent Robots and Systems*, pp. 2023–2029, September 1994.

[10] Hirochika Inoue, Masayuki Inaba, Taketoshi Mori, and Tetsuya Tachikawa. Real-time vision system based on correlation processing(in Japanese). *Journal of the Robotics Society of Japan*, Vol. 13, No. 1, pp. 134–140, 1995.