

## ACTIVE CAMERA GUIDED MANIPULATION

Jiang Yu Zheng†, Qian Chen‡ and Saburo Tsuji‡

† ATR Communication Systems Research Laboratories  
Seika, Soraku, Kyoto 619-02, Japan

‡ Department of Control Engineering, Osaka University  
Toyonaka, Osaka 560, Japan

### Abstract

In this paper, we introduce a new scheme that uses an active camera to guide object manipulation in an environment that has not been modelled. By actively moving camera to desired position with proper distance and viewing angle from objects of interest, we can acquire good data for robot visual feedback control when locating objects. The environment is described in object centered coordinates system and is measured relatively in consecutive 2D image spaces. This approach is flexible and efficient in manipulation, since it avoids complicate 3D modeling and image processing carried out is driven by tasks. We give an idea that not only robot operation, but also camera motion are guided by visual feedback. We also study the strategies to control this active guidance.

### 1 Introduction

A lot of works have been done on manipulation and path planning based on precise internal world model. If there is no 3D model obtained *in priori*, manipulation has to rely on sensor data<sup>[1]</sup>. In our case, the major operation is to pick up object and then place it at some desired position. Under this circumstance, the robot needs to obtain fine information for manipulation only near the object to operate, and visual feedback is efficient than acquiring perfect 3D structure followed by a blind move of robot. The robot motion is no longer absolute but relative to goal position and is adjustable according to camera output.

Up to now, the effectiveness of the visual feedback control is restricted by sensing ability of a passive camera. Two kinds of camera setting, one is to mount camera on arm and another is to locate camera at a stationary position with an over view of work space, have their shortcomings in understanding global relationship between the arm and other objects, and sensing distant objects on a good spatial resolution, respectively.

We introduce a new scheme that visual feedback is carried out by an active camera which can change its states such as viewing distance and view point. It is important to change camera state if we want visual feedback be robust. There is three-dimensional uncertainty in camera sensing, as well as two-dimensional uncertainty in image analysis. A 3D position computed from image disparity at two stationary view points may has a large error if object is distant. Occlusion may also happen when an object is viewed from some specific

point. In a large view sight, segmentation on an object of interest may yield unsatisfactory result because of poor resolution and inappropriate thresholding over entire view. However, by actively changing view point, viewing distance and resolution of camera, as well as selecting processing window, we can not only solve 2D and 3D uncertainties in image understanding and guiding robot, but also realize efficiency in processing.

Active vision, proposed by Aloimonos<sup>[2]</sup>, acquires structure of objects from reading output of constrained camera motion. The computation problem of 3D shape becomes simple and "well-posed". It is advocated by Ballard in a more general sense as human vision<sup>[3,4]</sup>. The action of human eyes to look and coordinate hand performance gives us a hint in dealing with manipulation problem. Bajcsy and etc. think vision process is a kind of "intelligent control"<sup>[5]</sup>. They attempted to explore scene exhaustively by tuning parameters of processing and sensor in order to obtain interpretation fit with model. We focus on problem of positioning in manipulation, and investigate active viewing paradigm concerning more with problem of how to get a good view according to ongoing task (task driven), as well as the way to survive when extraction of the minimum data necessary for assigned task is failed (data driven).

Our camera has functions of foveation, saccade, and function of changing view point. The camera motion is basically task driven and is qualitatively controlled using visual feedback as a human does. By comparing current view with an ideal state given by task, the camera can move to get a view with proper resolution, viewing direction and viewing aspect. Figure 1 shows the system diagram. After the camera is located, the robot arm moves to the position under the guidance of camera. It then informs the camera to start next task or sub-task. Our motivation is to manipulate objects in an environment with efficiency and accuracy.

In the following sections, we will first give some basic concepts of our system. The robot motion guided by camera is discussed simply in section 3. And we explore camera motion in section 4. Control strategy is described in section 5. Some experiments are shown in section 6.

## 2 Hand-eye Coordination

### 2.1 Focused Parts of Interest

The major tasks are to move a robot hand to pick up an object, or to place an object at a desired position.

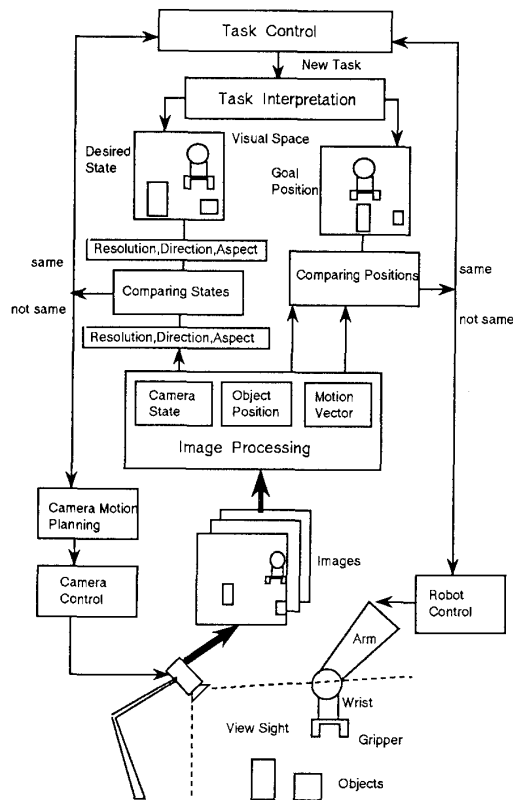


Fig.1 System Overview.

Our robot handles a block world. An active camera driven by another arm takes images for extracting position, size and motion of the focused parts of interest.

There are two parts of interest in guiding robot operation. One is a stationary goal in the environment and another is a moving target. In pick-up process, the goal position is on an object and the moving target is the robot hand. In put-down process, the goal is a specified position and the moving target is the object in robot hand. While the camera is moving to a new point, the focused parts of interest are both dynamic in the image. Both goal position and moving target are described by several major lines and points. These lines and points are tracked in the images while the target is moving in the space, or the camera is changing its state.

**2.2 Basic Steps in Handling Objects**

We describe the environment by an object centered coordinates system at the goal position as figure 2 depicts. The position of the moving target (either hand or object in hand) is denoted as  $D = (X, Y, Z)$ . For camera positions  $C_i, (i = 1, 2, \dots, I)$ , we obtain a series of 2D visual spaces.  $D$  is measured as  $d_i, (i = 1, 2, \dots, I)$  in these spaces. There can have various strategies for a target to approach the origin ( $D \rightarrow 0$ ). We move the hand in such a way that the displacement  $D$  is reduced

by a sequence of moves  $D_i, (i = 1, 2, \dots, I)$  generated at view points  $C_i$ . The consecutive move  $D_i$  not only makes image difference from the goal position  $d_i$  in the  $i$ th image approach to zero, but also keeps differences  $d_1, \dots, d_{i-1}$  at previous view points be zero. By selecting view point  $C_i$  properly, we can obtain distinct measure of differences  $d_i$  so as to control the robot motion. In the 3D environment, a vector can be determined from two aspect views. Therefore, the hand can arrive a position by at least two consecutive moves.

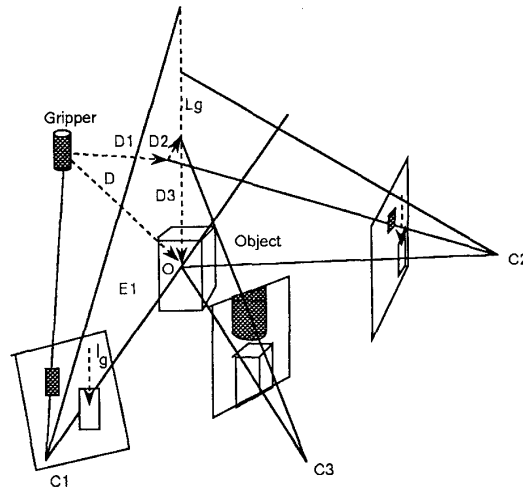


Fig.2 Object centered coordinates system.

**2.3 Task Oriented Image Processing**

Our camera can take images at different resolutions. Image processing works on region segmentation and line extraction, as well as tracking focused region in continuous images by correlation when a proper resolution is selected. Based on these results, a view port that covers objects of interest can be located as shown in figure 3. The distance between the center of the view port and image center yields an angle of viewing direction (camera axis) respect to a desired direction of observation. It is used to rotate camera when the parts of interest are necessary to be kept in the view sight or at the center of image. Tracking moving target in the images is done on low resolution level and measuring its precise displacement is usually on high resolution level. The data in a special window with changeable size and shape is located at focused part and is processed

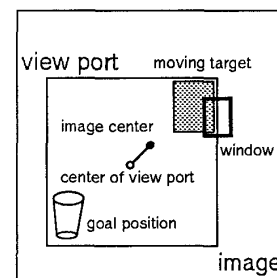


Fig.3 Image representation.

for saving cost in real time tracking and feature detection. Image processing is basically task driven, which switches between camera sub-task and robot sub-task in hand-eye coordination.

For simplicity, object recognition is carried out by using color in the current time. The object description is given by some attributes or approximate position indicated. More complicated description can be one or several aspect views of objects<sup>[7]</sup>. Some brief lines and points on the object are detected as a reference in determining its position relative to the goal position.

### 3 Camera Guided Robot Motion

The robot hand have translations  $(\Delta X, \Delta Y, \Delta Z)$  in the 3D space. Its wrist attached with a gripper has one degree of rotate ( $\theta$ ) around a vertical axis. We measure relative distance in two-dimensional space, which is less expansive than acquiring exact 3D shape and position by stereo vision and other range finder.

Let us look at the steps in picking up an object (see figure 2). Suppose the direction to grasp is given as  $L_g$  (from top), we want the hand to move to the  $L_g$  along which the object is picked up.

- 1: The camera first moves to a position to get an aspect view of the focused object. The projection of  $L_g$  in the image is  $l_g$ . The camera guides the target to arrive plane  $E1$  through camera center and the line  $L_g$ . If we can establish the connection between the object centered coordinates system and the camera coordinates system, the difference measured in the image at  $C_i$  can be converted to the robot motion in the 3D space. The robot hand first has three trial moves which are linear independent to each other in the space. Their image vectors are precisely extracted for estimation of the mapping between two coordinates systems (see Appendix). Then, a 2D vector  $d_1$  towards  $l_g$  is planned, and the direction of the robot motion  $D_1$  is given to be orthogonal to the camera axis. The robot hand moves along it until its projection arrives the line  $l_g$  in the view.
- 2: After the hand arrives the plane  $E1$ , it tries another move within the plane under the guidance of camera. This is attempt to locate and remember the constrained plane  $E1$ , in order to reduce one degree of freedom in locating hand. It is done by adjusting trial move in the 3D space until we get one move whose image vector is on the line  $l_g$ . The wrist rotates until the opened gripper is parallel to the surfaces to grip.
3. At this point, our processing switches to camera task in order to change view point for another aspect view (basically side view). From the second view point, the hand motion towards the object is straightforward. It tries move in the estimated constrained plane  $E1$  using visual feedback. The manipulator moves inversely if the motion detectable in the images is leaving the object.

The process of placing object is the same steps as the pick-up operation, except the goal position and moving target now become a specified position and an

object carried in hand, respectively.

## 4 Acquiring Good Camera States

### 4.1 Defining Camera Motion

A subject we are more interested in is how to move the active camera to acquire good data in guidance of manipulation. Some basic motivations are as follows.

1. Since we use multiple views to yield 3D location, a view reflecting 'distinct' depth is preferable in solving 3D uncertainty. Also, occlusion avoidance demands view point change of camera. The occlusion here means either occluded case where interested object is invisible or occluding case where segmentation may fail because of the low contrast between object of interest and others behind it.
2. The camera needs to change its resolution and processing window either to obtain precise position in locating object, or to obtain high speed in tracking target.
3. We need to change viewing distance to have proper view sight. This happens when the camera has a glance at whole site in finding objects and look into details in locating moving target.

The basic idea to control the moves is visual feedback, given some cues to approach desired state. The camera can translate  $(\Delta X, \Delta Y, \Delta Z)$  in the 3D space, and rotate  $pan_c$  along vertical axis and  $tilt_c$  along a horizontal axis orthogonal to the camera axis. At the beginning of the entire work, the camera axis (vector through the image center and its focal point) is calibrated briefly by a trial move approaching to a 3D point, which keeps the point at the image center during the move. This motion vector of camera denoted as  $Q$  is memorized as viewing direction of the camera. As the camera moves to various view points and changes its viewing directions, the camera axis is transformed according to the undergoing rotation and translation so that the current posture of the camera is always known. Based on this information, the horizontal direction perpendicular to the camera axis can also be figured out when the camera moves around.

The function of foveation that enlarges or reduces the size of view port in the image can be realized by either zooming camera or translating along camera axis towards the object. The saccade function that directs camera to various places is realized by controlling camera tilt and pan. And the aspect of object is updated by moving camera to new point of view. This is performed by a repetitive steps of 1) a short distance translation perpendicular to the current camera axis on horizontal plane, under the condition that any of the interested part is kept in the image, followed by 2) rotating pan to keep the focused point at the center of the image.

For each task, there are several parameters describing the goal state of camera. These parameters include *Size of view sight*, *Viewing direction*, and *Viewing aspect*. We move camera using visual feedback so that the differences between real measured camera state and the ideal goal state tends to zero. It is impossible to predict precisely where the camera can obtain optimal views from current image data, because we have no 3D model and our approach itself is to acquire good im-

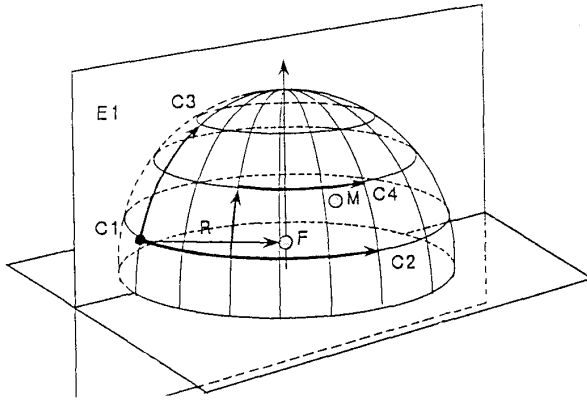


Fig.4 Camera space represented by spherical coordinates system centered at a fixation point on object.

age data from moving camera. Qualitative instructions such as *move forwards or backwards, turn camera to a point, and move along a circle on horizontal or vertical plane for aspect views* are given. A process plans camera motion repeatedly until all of the fields in the camera state satisfy the conditions given.

#### 4.2 Feedback Control of Camera Motion

The feedback control in changing viewing distance (resolution, size of view sight) and viewing direction are straightforward and have explicit goal conditions. In order to change the resolution, the camera moves along the camera axis until the size of the view port that covers the extracted parts of interest satisfies its specified ideal size, say 2/3 of the image in width. The viewing direction is changed by tilt and pan until the center of the view port shifts to the image center.

The feedback control in changing view point is a little complicate. Because this sub-task is for determining 3D location of target with respect to goal position, we can intuitively know the side view will make the difference  $D_2$  between the target and the goal position be distinct in the image. We suggest that the camera moves on horizontal or vertical plane along a circle around the objects. Let us describe the camera position by spherical coordinates system centered at a fixation point on an interested object, as figure 4 depicts. Let  $F$  and  $M$  denote the fixation point and the target, both on the constrained plane  $E1$  at first view point  $C_1$ . The camera can move on a sphere horizontally or vertically to get side view such as  $C_2$  or top view  $C_3$ . Its trace can even be a combination of horizontal and vertical moves to arrive a point such as  $C_4$ , if  $F$  is occluded by other objects during the horizontal move.

Since the 3D distance to the fixation point is not known, our camera motion along the horizontal or vertical circle is approximated by a sequence of translating sideway and then rotating camera axis towards fixation point. The problem now becomes whether such a move will arrive a stopping condition (goal state). As figures 5 show, the camera motion from the first view point to the second is realized as follows:

1. Pan rotation until  $M, F$  (both on the constrained plane  $E1$ ) locate on the central line in the image

(figure 5a). The goal position  $F$  is selected as fixation point. This insures that the translation and pan rotation followed will draw a circle on horizontal plane if step is fine.

2. Translation perpendicular to the camera axis on the horizontal plane and pan rotation that keeps the fixation point on the central vertical line. As the camera has an ideal move on the horizontal circle, the target  $M$  has a trace as an ellipse in the view (figure 5b). Therefore, the stopping condition for the camera motion is at the point where the image distance between  $M$  and the center line changes from increasing to decreasing.
3. Changing the size of view sight to have a close view of objects  $F$  and  $M$ .

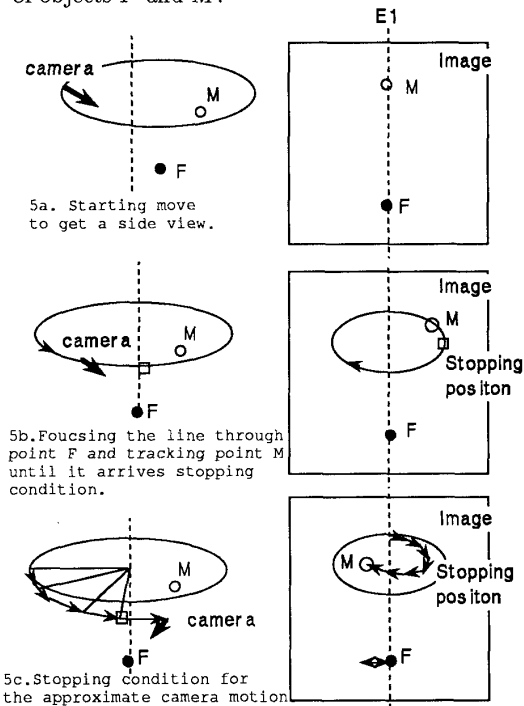


Fig.5 Moving camera to obtain an optimal aspect view.

Even for approximate motion that the translation has large distance, which comes out a trace leaving the ideal circle as figure 5c depicts, we can still find the stopping point in the image. Therefore, our camera pursues a qualitative motion, which is low cost in control.

## 5 Control Strategies

Our strategy to control this active guidance is mainly task driven, which is slightly different from that for recognition<sup>[8]</sup>. The major objective in our case is to fulfill task, instead of an exhaustive look or identification. Therefore, our system only acquires minimum data necessary for the tasks. And active viewing function is embedded in the task sequence generated from task definition. Our tasks and sub-tasks are organized

as a hierarchy as listed in figure 6. There are several levels in this hierarchy, each of them uses functions in the subsequent low level on its right side. Under the task level and its interpretation, there is a sub-task level for camera to look for distinct information and to guide robot. Camera and hand motion are performed at the control level behind it. Finally, they are realized at the level using visual feedback which require information from image processing. The ultimate functions are the extraction of color, region, line and matching. The numbers under each function item in figure 6 indicate the functions employed in its direct right column. As a result, generating task sequence is similar to the depth-first searching in the multi-layer tree of this task hierarchy.

We consider data driven processing only when information needed is failed to be extracted. The extra processing is fired only if information such as an image feature indispensable for finding fine location of object or tracking target becomes invisible. The reasons of failures are usually bad segmentation, bad illumination, occlusion, shape change when moving camera, and lack of contrast between focused object and background, etc.. Since there are many alternatives in camera parameters selection when pursuing tasks from high-level to low-level, our system performs the processing repeatedly by changing each parameters step by step. The criterion in surviving 'accident' in data acquisition is to have the least conflict to the general tasks in upward levels. The consequence of this 'back-tracking' is either the information is obtained so that the tasks are continued, or the processing goes one level above to start change parameters there.

## 6 Experiments

PUMA arms are used in our manipulation system. Image processing was done on sun4 workstation with an image processor. Figure 7-10 show a hand-eye motion sequence in picking up an object. The robot wrist is tracked and its position is computed from the center of the region. Figure 7 shows hand positions in approaching an object viewed at first view point. In figure 8, trace of its center is depicted by segments connecting the start and arrival positions. The first three strokes are the trial moves. At the stopped position, there is another trial move keeping it within the constrained plane **E1**. Figure 9 shows the hand motion in the second aspect view after moving camera. The hand has a random move which was opposite to the goal direction in this case and is reversed as soon as the motion is detected. As it arrived top of the object, the gripper moves down to grasp it, with a fine position adjustment as figure 10 displays.

Placing an object at a given position (or on an object in our experiment) is similar with the pick-up operation. Figure 11 shows the object (taken in hand) motion in the same aspect view as figure 9, which approaches to another constrained plane **E2** through the goal position. Beginning with trail moves, its trace is displayed in figure 12. The moves within the constrained plane **E2** to the top of the object and put down gripped one on it are given in figure 13, and its trace is shown in figure 14.

## 7 Conclusion

In this paper, we introduced a new method in manipulation which employs an active camera to realize visual feedback of the robot control. We have shown how a series of selected 2D view can perfectly solve the locating problem. We explored the issue of how to locate camera to get a good visual space, as well as the robot motion under the guidance of camera. Our camera motion is controlled qualitative by visual feedback, given some cues in the task definition.

Because we avoid computing absolute position of objects and robot hand, as well as planning robot path, and also because our control strategies of camera and image processing are not fell into the conventional hierarchy from low level sensing to high level interpretation and planning, but task driven, the manipulation is very simple and efficient. By actively changing visual space, we can locate objects precisely since the spatial resolution provided by the visual sensor, which was so far not as accurate as the robot motion in some part of the work space, is possible to be improved at any robot acting place.

## Acknowledgement

The first author wishes to give his sincere thanks to Dr. Kohhei Habara, Chairman of the board, Mr. Kohichi Yamashita, President, and Mr. Fumio Kishino, head of ATR Communication Systems Research Labs. for their supports regarding this research.

## Appendix

Because we only need to estimate the motion direction, other than absolute distance which will be determined in visual feedback, rotation transformation between object centered coordinates systems and camera centered coordinates system is sufficient to convey the planned move in the image to the robot motion in the space. Let  $V_o$  and  $V_c$  denote a motion vector in object and camera centered coordinates systems respectively (see figure 15). We can represent the relation between them as:

$$V_o = RV_c \quad (1)$$

where  $R$  is a rotation matrix. Let the trial moves of robot  $T_i (i = 1, 2, 3)$  be unit vectors orthogonal between each other, as figure 15 depicts, we have

$$T = [T_1 \ T_2 \ T_3], \quad |T| = 1, \quad ||T_i|| = 1 \quad (2)$$

and

$$T_3 = T_1 \times T_2, \quad T_1 = T_2 \times T_3, \quad T_2 = T_3 \times T_1 \quad (3)$$

Their corresponding vectors in the camera coordinates system are  $C_i (i = 1, 2, 3)$ , i.e.,

$$C = [C_1 \ C_2 \ C_3], \quad T = RC \quad (4)$$

For the simplicity, we assume the projection from one coordinates system to another is orthogonal projection. Hence, the measured  $x$  and  $y$  coordinates changes in the image due to the motion  $C_i (i = 1, 2, 3)$  become

$$k(C_{1x}, C_{1y}), \quad k(C_{2x}, C_{2y}), \quad k(C_{3x}, C_{3y}), \quad (5)$$

where  $k$  is a scaling coefficient. Because rotation transform will not change length of vector, as well as angle between vectors, we have:

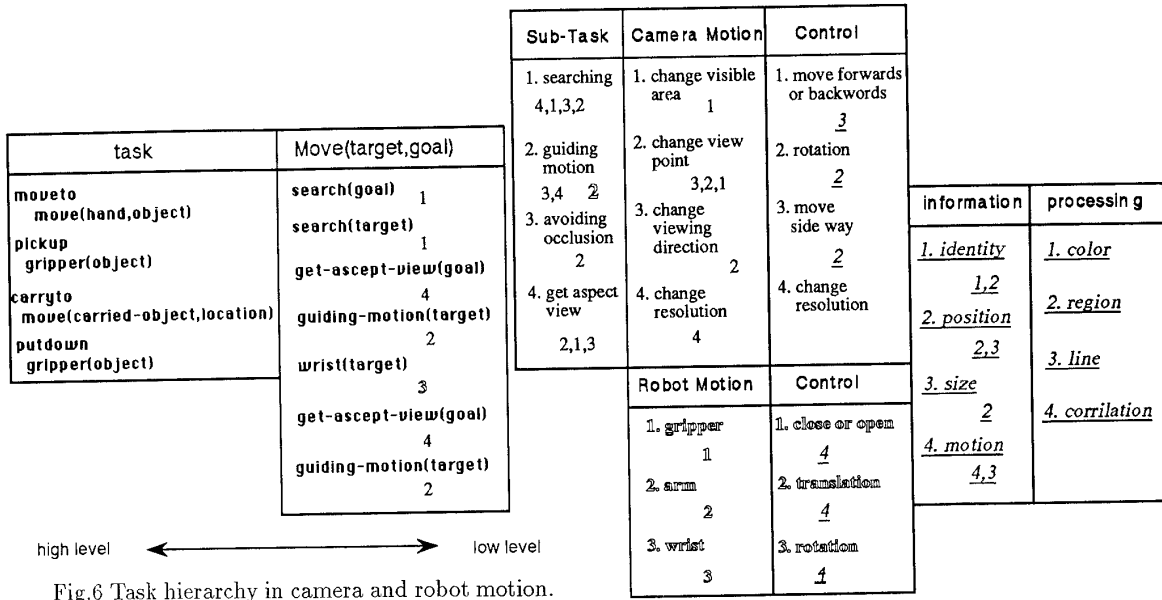


Fig.6 Task hierarchy in camera and robot motion.

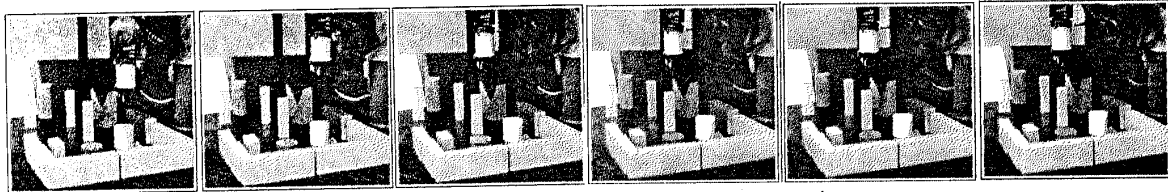


Figure 7. Camera guided hand motion viewed from first point.

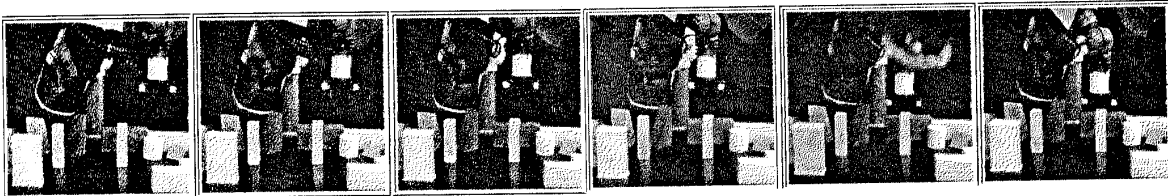


Figure 9. Hand motion viewed from another view point after moving camera.

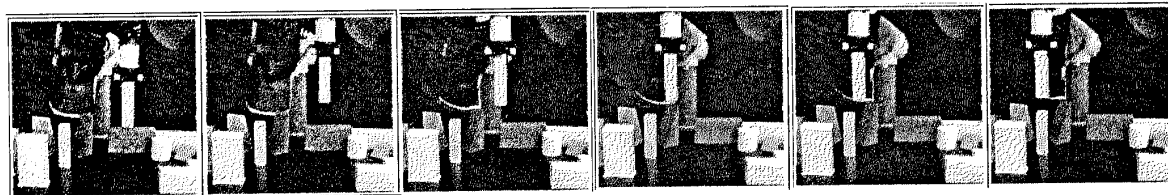


Figure 11. Taking object to another goal position.

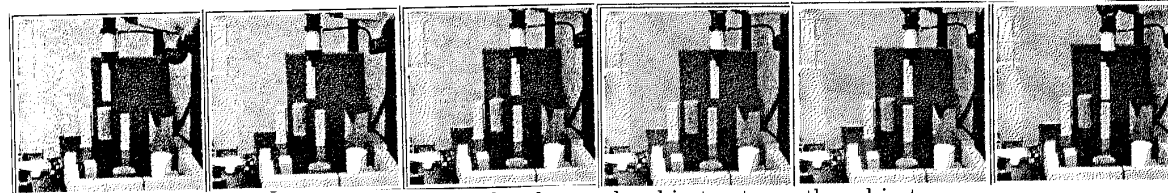


Figure 13. Moving hand to locate the object onto another object.



Figure 8. Trace of robot hand in motion displayed in figure 7.

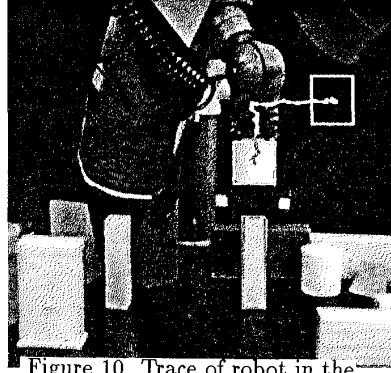


Figure 10. Trace of robot in the constrained plane.

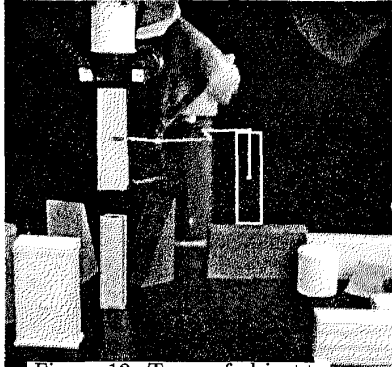


Figure 12. Trace of object to the constrained plane.

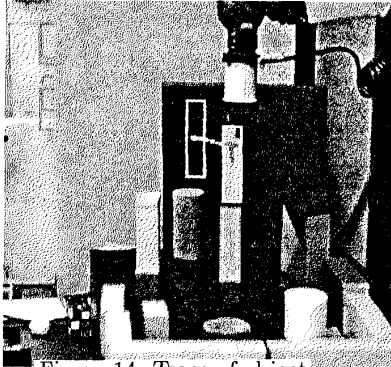


Figure 14. Trace of object when it is put down.

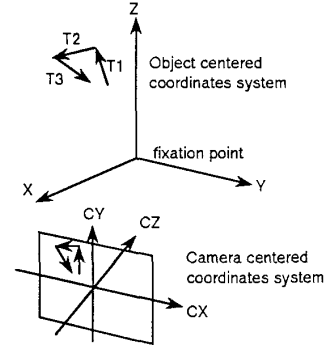


Figure 15. Camera guided move of robot hand.

$$C_3 = C_1 \times C_2, \quad C_1 = C_2 \times C_3, \quad C_2 = C_3 \times C_1 \quad (6)$$

and

$$\|C_1\| = \|T_1\| = 1. \quad (7)$$

according to (2)(3). In more detail, it can be written as:

$$C_{1z} = k^2(C_{2x}C_{3y} - C_{3x}C_{2y}) \quad (8)$$

$$C_{2z} = k^2(C_{3x}C_{1y} - C_{1x}C_{3y}) \quad (9)$$

$$C_{3z} = k^2(C_{1x}C_{2y} - C_{2x}C_{1y}) \quad (10)$$

and

$$C_{1x}^2 + C_{1y}^2 + C_{1z}^2 = 1 \quad (11)$$

from which  $k$  can be computed, and  $C$  is determined completely. Based on this result,  $R$  is determined as

$$R = TC^{-1} \quad (12)$$

We design the real move from the end of trial moves towards a point on  $l_g$  in the image.  $V_{cx}$  and  $V_{cy}$  thus are determined. The  $V_{cz}$  is simply set zero so that the direction in the 3D space is parallel to the image frame (perpendicular to the constrained plane  $E1$ ). Therefore, we only need to compute the first two columns in  $R$  for the motion direction  $V_o$ . The real motion is de-

tected in continuous images until it arrive  $l_g$ . Because we simplify the relations a great deal, the estimation of  $V_o$  is not precise and we rely on visual feedback more in locating robot.

## References

- [1] L.E.Weiss, A.C. Anderson and C.P. Neuman, "Dynamic sensor based control of robots with visual feedback", IEEE Journal of Robotics and Automation, 5(3): pp.404-417, 1987.
- [2] J. Aloimonos and A. Bandyopadhyay, "Active vision", Proc. First Int. Conf. Computer Vision, 1987, pp.35-55.
- [3] D.H. Ballard, "Reference frame for animate vision", Proc. IJCAI-89, pp.1635-1641 (1989).
- [4] D.H. Ballard, "Parameter nets: A theory of low level vision: Kinetic depth", Proc. 2nd Int. Conf. on Computer Vision, pp.524-531, 1988.
- [5] R. Bajcsy, "Active perception", Proc. IEEE special issue on Computer Vision, August 1988, pp.996-1005.
- [6] S.Tsuji and J.Y.Zheng, "Visual path planning" Proc. IJCAI-87, Vol.2, pp.1127-1130 (1987).
- [7] K. Ikeuchi and T. Kanade, "Automatic generation of object recognition programs", Proc. IEEE special issue on Computer Vision, August 1988, pp.1016-1035.
- [8] A. Califano, R. Kjeldsen and R.M. Bolle, "Data and model driven foveation", Proc. 10th ICPR, Vol.1, pp.1-7, 1990.