Determining Grasp Configurations using Photometric Stereo and the PRISM Binocular Stereo System

Katsushi Ikeuchi
Electrotechnical Laboratory,
Sakura-mura, Niihari-gun, Ibaraki, 305, Japan

H. Keith Nishihara
Schlumberger Palo Alto Research,
3340 Hillview Avenue, Palo Alto, California 94304

Berthold K. P. Horn
Patrick Sobalvarro
MIT Artificial Intelligence Laboratory,
545 Technology Square,
Cambridge, Massachusetts 02139

Shigemi Nagata
Fujitsu Laboratory,
1015 Kamiodanaka, Nakahara-ku,
Kawasaki, 211, Japan

Abstract

This paper describes a system which locates and grasps parts from a pile. The system uses photometric stereo and binocular stereo as vision input tools. Photometric stereo is used to make surface orientation measurements. With this information the camera field is segmented into isolated regions of a continuous smooth surface. One of these regions is then selected as the target region. The attitude of the physical object associated with the target region is determined by histograming surface orientations over that region and comparing them with stored histograms obtained from prototypical objects. Range information, not available from photometric stereo, is obtained by the PRISM binocular stereo system. A collision-free grasp configuration is computed and executed using the attitude and range data.

1. Introduction

1.1. Overview

Image understanding research has produced various techniques for extracting information about visible surfaces from a scene. Two lines of research that have been investigated extensively are shape from shading (Horn 1975) and binocular stereo (Marr and Poggio 1979). This paper demonstrates how to use these methods in solving practical problems in robot manipulation. We explore the complementary use of photometric stereo and binocular stereo to solve problems in locating good grasp points on a part in a bin. The task requires the following steps:

1. Identify the *location* of the part in a complex scene,
2. Measure the *attitude* of the part,
3. Measure the *elevation* of the part above some reference plane, and
4. Compute a collision-free *grasp configuration.*

An earlier paper (Ikeuchi, Horn, Nagata, Callahan, and Feingold 1983) presented techniques for using photometric stereo to accomplish Nos. 1 and 2; and, in addition, to determining the class to which an object belongs from a set of known shape classes. In this paper we combine that system with a binocular stereo system PRISM designed for use in robotics (Nishihara and Poggio 1984). The purpose of this extension is not only to support the planning process with the range data from the PRISM stereo but also to demonstrate the importance of the hybrid use of complementary sensing mechanisms.

The International Journal of Robotics Research,
Vol. 5, No. 1, Spring 1986

Photometric stereo determines the orientation at a point on an object's surface from the image brightnesses obtained at the corresponding point in the image under three different illumination conditions. Distortions in brightness values due to mutual illumination or shadowing between neighboring objects are detected by the method as *impossible* brightness triples. The locations of these triples was used to segment the visual scene into isolated regions corresponding to different objects. The distribution of surface orientation, an orientation histogram, measured over one of these isolated regions was used to identify the shape from a catalogue of known shapes. The object's attitude in space was also obtained as a by-product of the matching process.

The part's elevation, however, was not known and had to be measured by moving the manipulator hand down the camera line of sight towards the part until a light beam between the fingers was broken. With the elevation known, the manipulator was retracted and a second approach made along a trajectory appropriate to the part's attitude.

There were two problems with this approach:

1. The pickup motion required two separate arm motions: the first, to measure elevation and the second, to grasp the object.
2. Collisions of the gripper with neighboring objects could not be predicted since their distances from the target were not available to the system.

In the hybrid approach presented here, a binocular stereo system is used to produce a coarse elevation map for determining a collision-free configuration for the gripper and for measuring the absolute height at the selected pickup point.

1.2. Related Works

Bin-picking tasks by detecting brightness changes have been attacked previously (Tsuji and Nakamura 1975; Baird 1977; Perkins 1977; Bolles and Cain 1982). Detecting brightness changes gives boundaries between regions that correspond to the objects. The boundaries obtained are compared with internal models to determine the attitude of the object. These edge-based approaches work particularly well with isolated objects, which lie on a uniform background, provided the objects rotate only in the plane of support. In other words, these algorithms work well on binary images; but such methods cannot extract the contour of an object from the image of a set of overlapping objects, which is typical in bin picking.

Kelley and others (Birk, Kelley and Martins 1981) highlight scenes to segment and to determine the position and the orientation of an object in a bin. Their system is limited to cylindrical workpieces with a metallic surface. Their vision system determines only two degrees out of three degrees of freedom in attitude.

2. Basic Vision Modules

There are three basic vision modules in our system: photometric stereo, binocular stereo using the PRISM algorithm, and extended Gaussian image matching.

2.1. Reflectance Map and Photometric Stereo

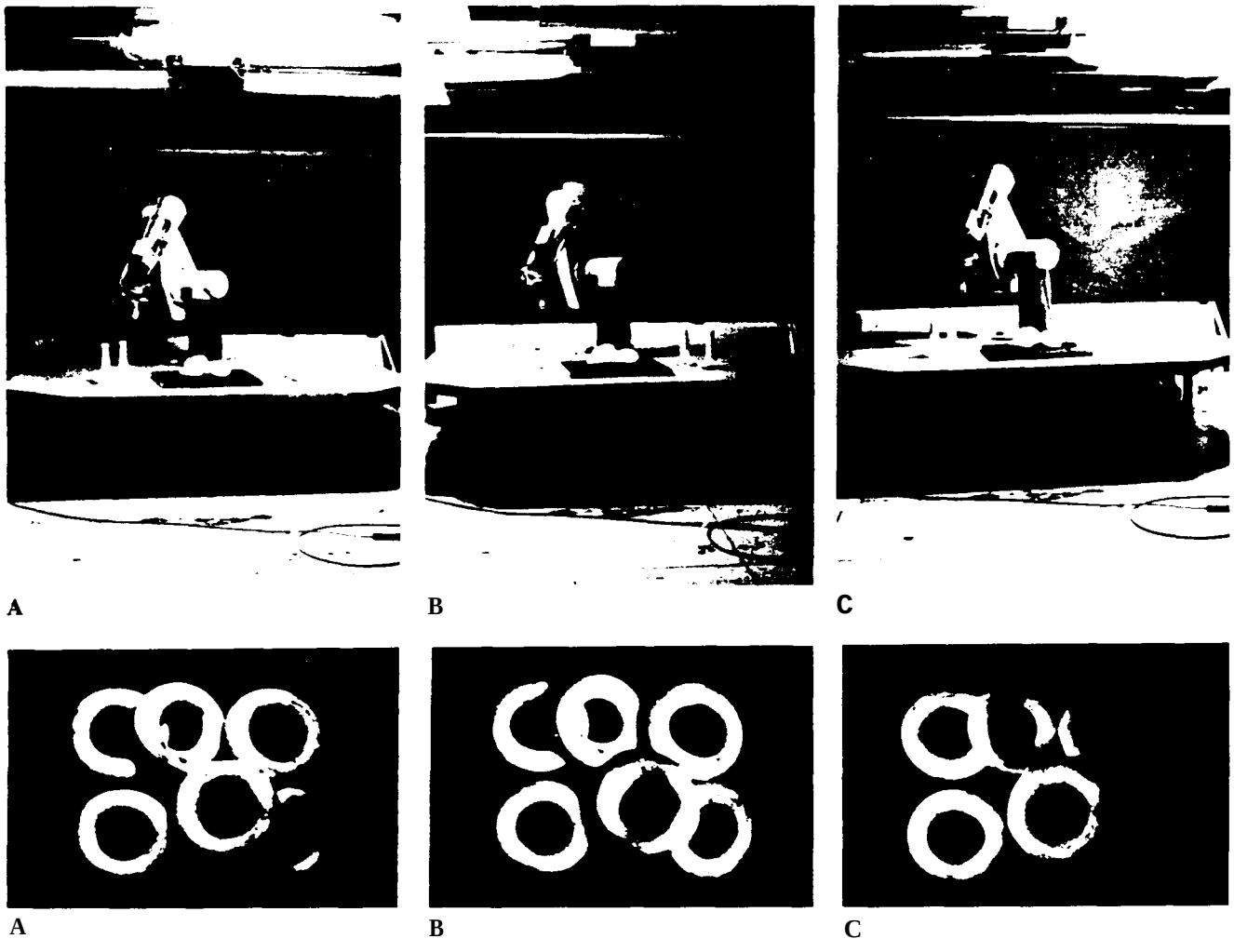
The reflectance map (Horn 1977) represents the relationship between surface orientation and image brightness. Since the direction of a surface normal has two degrees of freedom, we can represent surface orientation by points on a sphere or in a two-dimensional plane. The brightness value associated with each surface orientation — assuming a fixed light source and viewing configuration — can be obtained either empirically (Woodham 1979) or analytically from models of the surface microstructure and the surrounding light source arrangement (Horn and Sjoberg 1979).

The photometric stereo method takes multiple images of the same scene from the same camera position with various illumination directions in order to determine surface orientation (Horn, Woodham, and Silver 1978; Woodham 1978; Silver 1980; Woodham 1980; Ikeuchi 1981b; Coleman and Jain 1981). This setup gives multiple brightness values at each picture cell. Since different images are taken from the same point, there is no disparity between the images as there is with binocular stereo, so no correspondence problem has to be solved.

Each illumination configuration has a unique reflec-

Fig. 1. Light source for photometric stereo.

Fig. 2. Three brightness arrays.



tance map associated with it; so each of the three brightness measurements is consistent with **a** different set of surface orientations. Each of these sets corresponds to an iso-brightness contour on the reflectance map associated with that lighting configuration. The intersection of the three contours obtained will yield typically a unique surface orientation.

This method is implemented by using a lookup table. If we assume both the viewer and the light source are far from the object, then both the light source direction and the viewer direction are essentially constant over the image. Thus, for a particular light source, the same reflectance map applies everywhere in the image. In practice, a calibration object of known shape is used to determine the relationship between brightness and surface orientation. The points where iso-brightness lines cross can be precalculated and stored **as** a table of surface orientations that is indexed by triples of brightness values. Thus, the main operation of the algorithm is the lookup table which makes it possible to determine surface orientations very rapidly.

This lookup-calibration method also extends the scope of the photometric stereo applications. Since the lookup table is obtained from a calibration object, the object's albedo need not be known. **A** useful lookup table can be obtained for any albedo even when the albedo contains **a** strong specular component. **A** lookup

A

table can give an orientation surface with an arbitrary albedo provided that the surface has the same kind of albedo as the calibration object.

The result of the application of the photometric stereo method is called a needle diagram since it can be shown as a picture of the surface covered with short needles. Each needle is parallel to the local normal. The length of a line, which is the image of one of the needles, depends on how steeply inclined the surface is. The orientation of the line indicates the direction of steepest descent.

Three images are obtained with three light sources (banks of ordinary fluorescent lamps) and using a single CCD TV camera as shown in Fig. 1. Three images obtained under different illumination conditions are shown in Fig. 2. The photometric stereo generates a needle diagram as shown in Fig. 3.

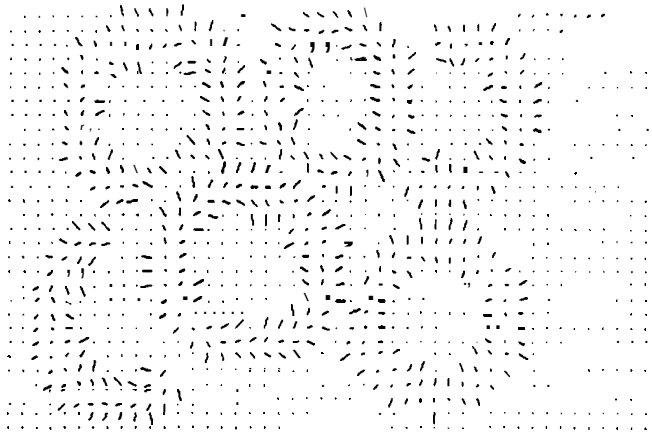


Fig. 3. A needle diagram generated using photometric stereo.

2.2. The PRISM System

The PRISM stereo-matching algorithm was designed to produce range measurements rapidly in the presence of noise. The algorithm is built on the zero-crossing stereo theory of Marr and Poggio (1979). Their approach uses scale specific image structure in a coarse-guides-fine matching strategy. Their matching primitive was defined in terms of local extrema in the image brightness gradient after approximate lowpass filtering with a two-dimensional Gaussian convolution operator. The lowpass filtering serves to attenuate high spatial frequency information in the image so local maxima in the gradient will correspond to coarse scale properties of the image. The locations are approximated by zero crossings in the Laplacian of the Gaussian filtered image, or equivalently, zeros in the image convolved with a Laplacian of a Gaussian $\nabla^2 G$ (Marr and Hildreth 1980). The PRISM algorithm, however, does not explicitly match zero-crossing contours.

The zero-crossing contours are generally stably tied to fixed surface locations, *but* their geometric structure carries more information, some components of which are closely coupled to system noise. Consequently, algorithms which explicitly match zero-crossing contours tend to be more noise sensitive than is necessary (Nishihara 1984). Matching the dual representation — regions of constant sign in the $\nabla^2 G$ convolution — produces useful results over a broader range of noise levels and does it more rapidly than algorithms that explicitly match the shape of the contours bounding regions of constant sign.

An additional consideration that has influenced the design of this system is the specific nature of most sensory tasks in robotics (Nishihara and Poggio 1984). Our view in this design has been that by avoiding the computation of details not necessary for accomplishing the task at hand, a simpler, faster, and possibly more robust performance can be obtained. The PRISM system (Nishihara 1984) was designed to test this notion.

The initial design task of the implementation was to detect rapidly obstacles in a robotics workspace and determine their rough extents and heights. In this case, speed and reliability are important while spatial precision is less critical.

Four components make up the system. The first is an *unstructured* light source used to illuminate the workspace. A simple slide projector covers the viewed surfaces with a random texture pattern to provide a high density of surface markings to drive the binocular matching. The specific geometry of the markings is not important to the matching. Markings already present in the physical surface do not interfere with, and in fact, assist the matching process. This is not the case with single camera *structured* light systems which depend on the measurement of the fine geometric structure of a known projected pattern. See Fig. 4.

The second component is a high speed convolution device (Nishihara and Larson 1981) which applies a

Fig. 4. Light source for binocular stereo.

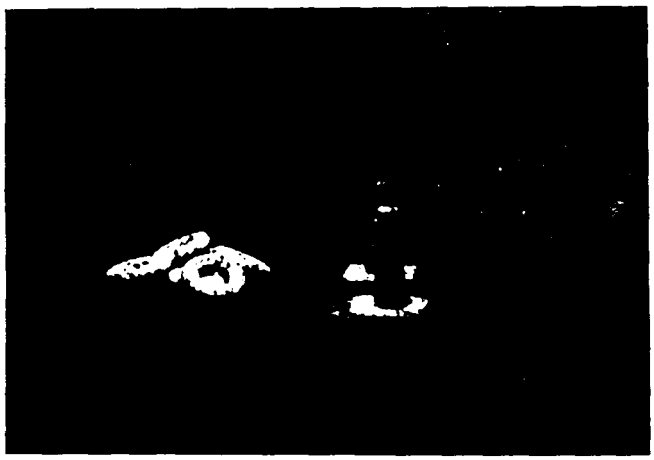
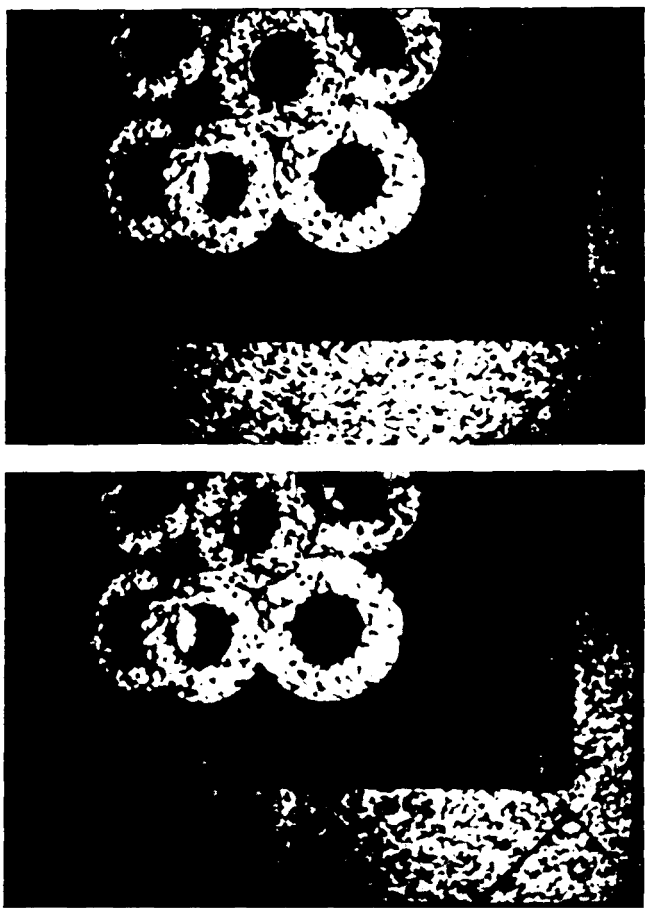


Fig. 5. Stereo pair of brightness arrays with unstructured light illumination.



32×32 approximation of the $\nabla^2 G$ operator to the left and right camera images.

The third component uses a binary correlation technique to determine the relative alignments between patches of the left and right filtered images that produce the best agreement between the convolution signs. This operation is accomplished at three scales of resolution using a coarse-guides-fine control strategy. The result is a disparity measurement indicating the best alignment, as well as a measure of the quality of the match between left and right images at that alignment.

The final component handles the conversion of image position disparity to physical height. Two conversion tables are used. One gives absolute elevation as a function of horizontal disparity. The other table gives vertical disparity as a function of horizontal disparity. Together they allow cameras with large, but stable, geometric distortion to be used. Both mappings depend on position in the image.

The test system uses a pair of inexpensive vidicon cameras. In the first implementation, vidicons were selected over solid state cameras to allow an assessment of the approach with particularly bad geometric distortion and limited brightness resolution. The cameras are mounted above the workspace of a commercial manipulator, the Unimation **PUMA.** The digitized video signals are fed to the high speed digital convolver which applies a 32×32 approximation of the $\nabla^2 G$ operator to the images at a 10^6 picture cell per second rate.

Matching is accomplished in software on a Lisp machine. The basic module of the program performs a test on a single patch in the image at a single disparity and determines whether or not a correlation peak occurs nearby. If one does, the approximate distance and direction in disparity to that peak is estimated. The detection range of this module is determined by the size of the convolution operator used. With the largest operator, a single application of the module covers a range of about 12 picture cells in disparity. Repeated applications of this module are used to produce a 36×26 matrix of absolute height measurements, accurate to approximately 10 mm with a repeatability about 5 times better. The matching covers a 100 picture cell disparity range and takes 30 seconds from image acquisition to final output.

Fig. 6. Output from the PRISM stereo module shown as a perspective plot.

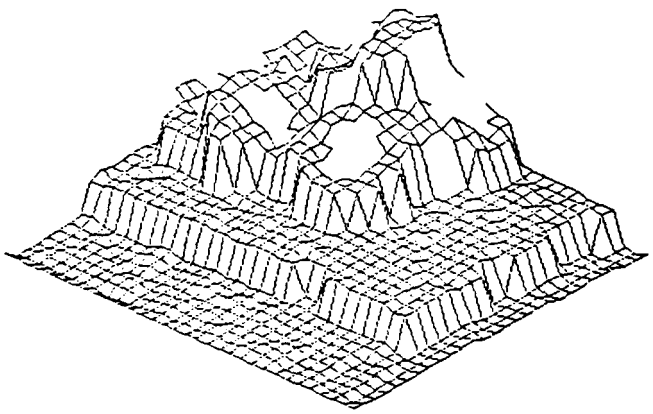
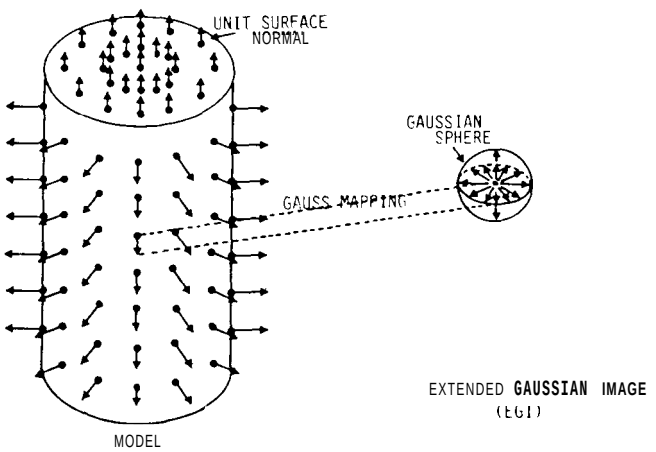


Fig. 7. Extended Gaussian Image.



A pair of images are obtained under the random texture illumination using a pair of vidicon TV cameras for the **PRISM** stereo system. A pair of brightness arrays for the binocular **PRISM** stero are shown in Fig. **5**. The output of the **PRISM** stereo as a perspective plot are shown in Fig. 6.

2.3. Extended Gaussian Image Matching

The extended Gaussian image (EGI) of an object can be approximated by the histogram of its surface orientations. Assume there is a fixed number of patches per unit surface area and that a unit normal is erected on each patch. These vectors can be moved without changing the direction they point in, so their *rails* are at a common point and their *heads* lie on the surface of a unit sphere. Each point on the sphere corresponds to a particular surface orientation. This mapping of points from the surface of the object onto the surface of a unit sphere is called the Gaussian image; the unit sphere used for this purpose is called the Gaussian sphere (Do Carmo 1976).

Assume that a mass, equal to the area of the patch it corresponds to, is now attached to each end point. The resulting distribution of masses **is** called the *EGI* of the object (Smith 1979, Bajcsy 1980, Ballard and Sabbah 1981, Ikeuchi 1981a, Horn 1983, Brou 1984, Little 1985) in the limit as the density of surface patches becomes infinite. (See Fig. 7). It has several interesting properties: the total mass is equal to the surface area of the object, the center of mass is at the center of the sphere, and there is only one convex object corresponding to any valid EGI.

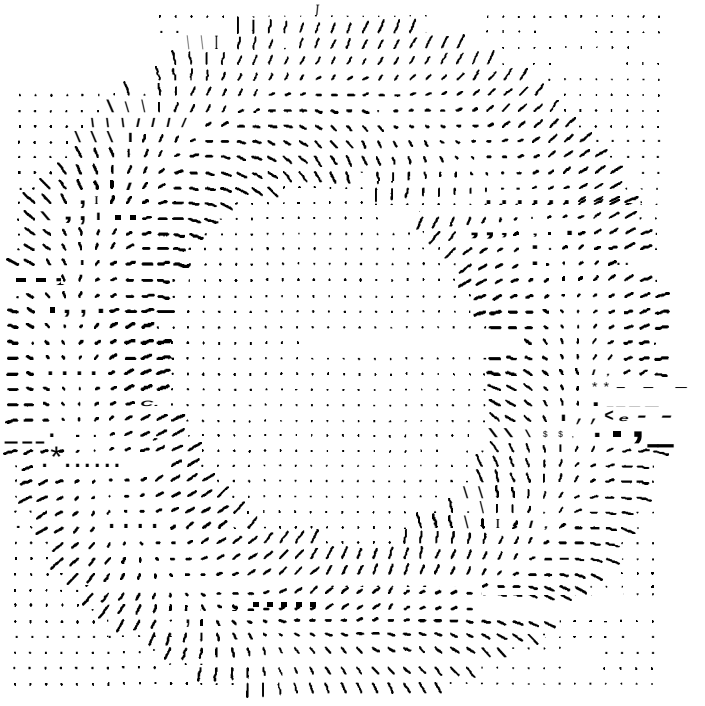
The EGI is invariant with respect to translation of the object. If it is normalized by dividing by the total mass, then it is also invariant with respect to scaling. When the object rotates, the EGI is changed in a particularly simple way; it rotates in the same manner as the object. These properties make it attractive for determining the attitude of an object.

A surface patch is not visible from a particular viewing direction if the normal to the surface makes an angle of more than 90° with respect to the direction towards the viewer. The orientations, which correspond to those patches that *are* visible, lie on a hemisphere that is obtained by cutting the Gaussian sphere with a plane perpendicular to the direction towards the viewer. This hemisphere will be referred to as the *visible hemisphere* (Ikeuchi 1983). It should be clear that we can estimate only one half of the EGI from data obtained using photometric stereo or depth ranging.

We will call the point where the direction towards the viewer intersects the surface of the visible hemisphere the *visible navel.* Surface patches that are visible have orientations which correspond to points on the Gaussian sphere whose distance from the navel, measured on the surface of the sphere, is no more than $\pi/2$.

There are two problems in matching the EGI estimated from experimental data with those obtained from object models and stored in the computer. They are the number of degrees of freedom of the attitude

Fig. 8. The detailed needle diagram over the target region.



of an object and the effects of self-obscuration on the observed EGIs for objects that are not convex.

The attitude in space of an object has three degrees of freedom. Correspondingly, there are three degrees of freedom in matching the observed EGI and a prototypical EGI. Two degrees of freedom correspond to the position on the prototypical Gaussian sphere of the visible navel of the observed EGI (that is, the direction towards the viewer). The remaining degree of freedom comes from rotation of the observed EGI, relative to the prototypical EGI, about its visible navel (that is, the rotation of the object about the direction towards the viewer). One approach is to evenly sample the space of rotations and perform a match for every trial attitude. This brute force method can be somewhat expensive if reasonable precision in determining the attitude is required, since the space of rotations is three-dimensional.

We use two notions to constrain orientation. First, note that the apparent (cross-sectional) area of an object depends on where it is viewed from. It can be shown that the height of the center of mass of the visible hemisphere of the EGI, above the plane through the edge of the hemisphere, is equal to the ratio of the apparent to the actual area. *So* the location of the center of mass of the observed EGI constrains the possible positions of the visible navel on the prototypical EGI. Note that the center of mass of the *whole* EGI is at the center of the sphere and so is of no use. Second, the direction of the axis of least inertia of the observed EGI can be used to determine the relative rotation between the two EGIs for a particular position of the navel on the prototypical EGI (Ikeuchi 1983).

In the case of a convex object, the EGI obtained from a needle diagram taken from a particular direction, is equal to the full EGI of the object restricted to the corresponding visible hemisphere. This is not the case, in general, when dealing with a non-convex object. Some surface patch may be obscured by another part of the object and invisible even if the normal makes an angle of less than 90" with the direction towards the viewer. *So* the contributions of surface patches to the EGI will vary with viewing direction.

One can deal with a non-convex object by defining a viewer-direction dependent EGI, which takes into account the effects of obscuration. We can store these EGIs in a table whose columns correspond to different viewer directions; whose rows correspond to different positions on a visible hemisphere; and whose contents correspond to EGI masses of different surface orientations under different viewer directions (Ikeuchi 1983). We will use this EGI lookup table for EGI matching.

A needle diagram of an object is shown in Fig. 8. The EGI obtained from the needle diagram is shown in Fig. 9. Note that the EGI is normalized so that the EGI minimal inertia direction agrees with the x axis of the coordinate.

2.4. Information Flow in Vision Module

The photometric stereo method and the matching of orientation histograms is implemented on a Lisp machine. This Lisp machine also controls the flow of execution. The PRISM stereo system is implemented on another Lisp machine running in parallel. Both Lisp machines and the PUMA arm controller are connected via a local area network, the Chaos net.

Information flow in the vision part is shown in Fig. 10.

Fig. 9. The EGI obtained from the needle diagram over the target region. Note that the EGI is normalized so that the EGI minimum inertia direction agrees with the x axis of the coordinate.

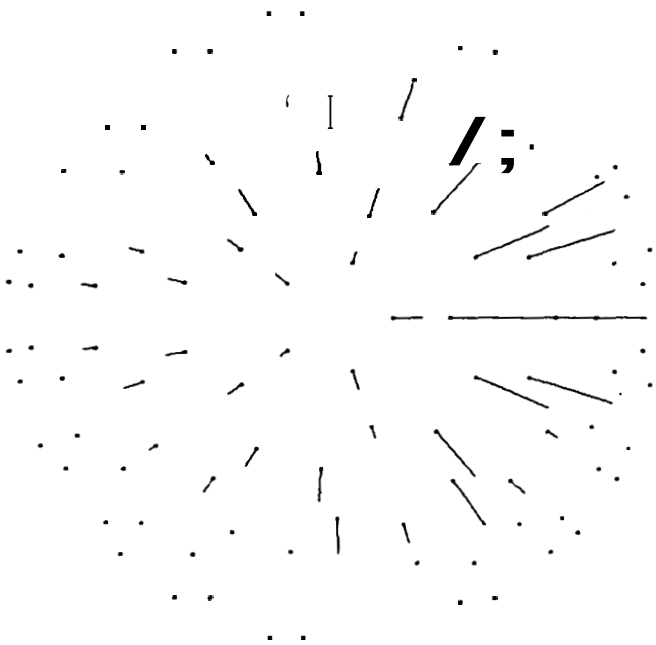
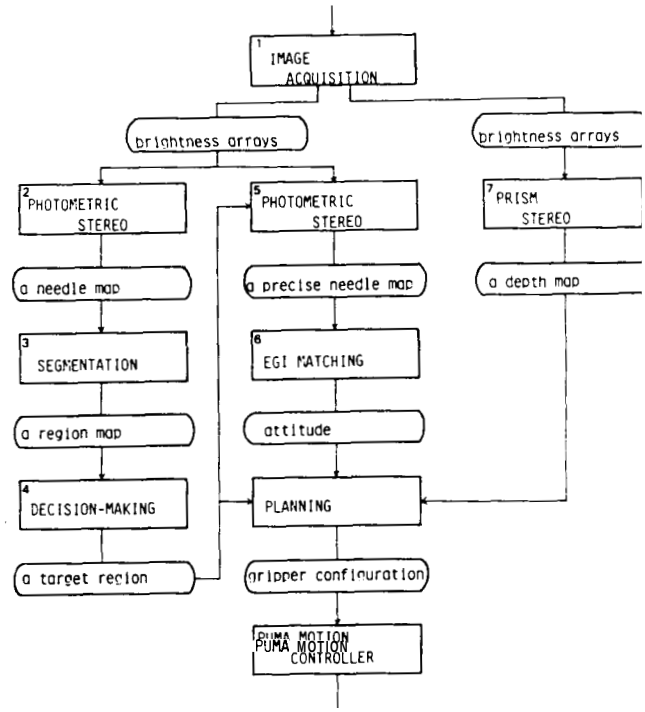


Fig. 10. Information flow in the vision part.



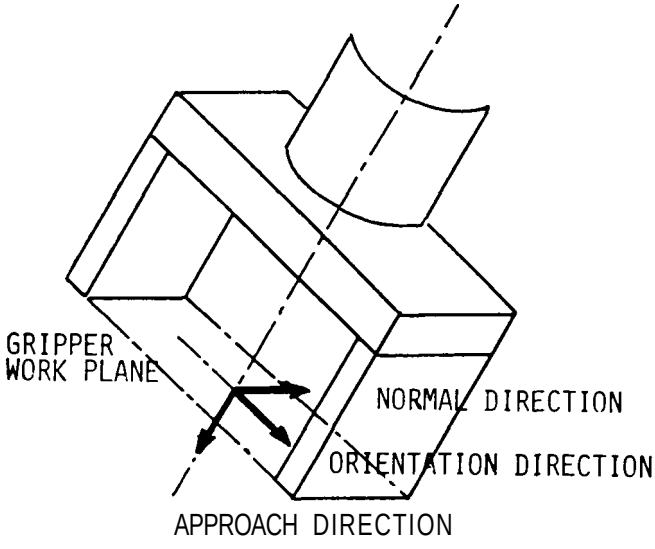
1. Three images of the scene are obtained using a single CCD TV camera and three different light sources. A pair of vidicon TV cameras for the PRISM stereo system obtain a pair of images under the random texture illumination.
2. The photometric stereo module generates a needle diagram of the scene by means of the lookup table that is developed by using a calibration object.
3. The segmentation process divides the input scene into isolated regions based on the needle diagram. Segmentation is based on:
   a. areas where the surface normal vanes discontinuously with position,
   b. areas where the system cannot determine surface orientation due either to shadowing or mutual illumination.
4. One target region is selected among the isolated regions based on the Euler number and the area of the region.
5. The photometric stereo module is run again on the original image data, using a different lookup table, to obtain more detail in the regions near the edge of the target object. New images, taken with different lighting conditions, could actually be used here. The result is used to produce an orientation histogram that is the discrete approximation of the EGI.
6. The EGI matching process compares the EGI obtained from the needle diagram with stored EGIs and determines the attitude of the object.
7. In parallel with steps (2–6), the PRISM system produces an elevation map over the image.

3. Grasp Configuration

The grasp configuration should satisfy the following two conditions (friction is assumed):

1. It should produce a mechanically stable grasp, given the gripper's shape and the object's shape. Such configurations will be called *legal grasp* configurations.
2. The configuration must be achievable without collisions with other objects. Grasp configurations are limited by the relationship between the shape of the gripper and the shapes of

Fig. 11. The parallel jaw gripper and three directions



neighboring obstacles. Such configurations will be called *collision-free* configurations.

These configurations depend on the type of gripper. We assume that the gripper has a pair of parallel rectangular jaws as is commonly the case in current industrial robots. (See Fig. 11).

3.1. Legal Grasp Configuration

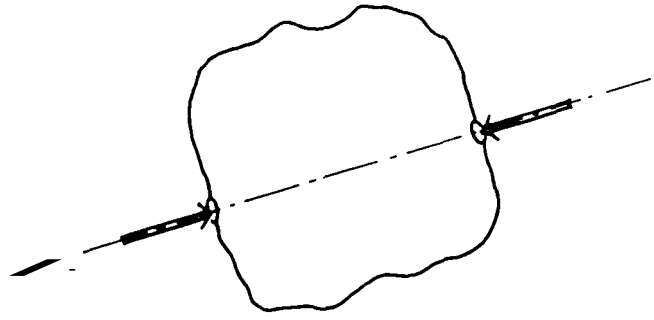
There are several definitions of legal grasping (Hanafusa and Asada 1977; Brady 1982; Boissonnat 1982). We define the legal grasping configuration as the one in which the object satisfies the following two conditions:

1. The object is not *translated* while the gripper is grasping the object.
2. The object is not *rotated* while the gripper is grasping the object.

A parallel jaw gripper applies forces at two points. In order to guarantee conditions of (1) and (2), the two applied forces should be identical in magnitude, opposite in direction, and lying along the line connecting the contact points, as indicated in Fig. 12.

Consider the force at one of these points of contact. Let the friction angle be the arc tangent of the coefficient of friction. If the angle between the surface normal direction and the line connecting the two grasping points is less than the friction angle, then the direction of the force applied by the gripper can agree with the line connecting the two points of contact (See Fig. 13a). If the angle is larger, the force does not lie along that line (See Fig. 13b), because friction can only contribute Nv_0 in the direction parallel to the surface where N is the normal force and v_0 is the coefficient of friction. In cases where we cannot predict the magnitude of the friction angle, the most conservative solution is one in which the surface normals at the two contact points lie on the same line. This is a necessary and sufficient condition for satisfying conditions (1) and (2) in the absence of friction information.

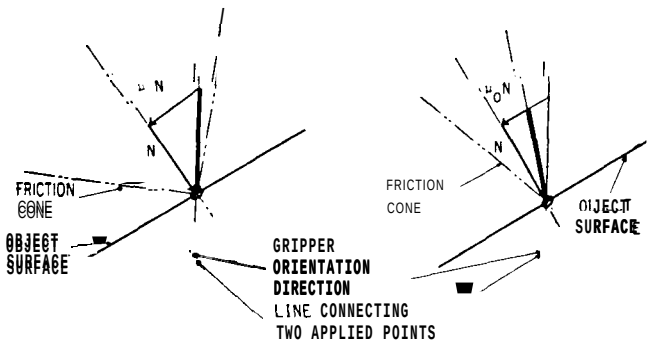
Fig. 12. The two applied forces: The appliedforces should be the same in magnitude, of opposite direction, and be along the line between the two contact points.



3.2. Determining Legal Grasp Configurations from Object Shape

The next task is to extract legal grasp points by using the previous rule. This task can be done by exploring the surface of the object. We assume that the surface normal direction at some point P can be determined. We will construct a line in a direction opposite to that of the surface normal and extend the line until it reaches the other side of the object. The symbol Q will represent the point reached. If the surface normal at the point Q agrees with the direction of the line, then the pair of points (P, Q) is added to the list of possible legal positions. It is possible that no such pairs will be found. In that case this simple algorithm decides that the object is not graspable. Usually, however, there is an infinite number of point pairs satisfying this condition.

Fig. 13. Friction cone and applied force. If the angle between the surface normal direction and the line connecting the two grasp points is less than half of the zenith angle of the friction cone. the direction of the force applied by the gripper coincides with the line connecting the two grasp points. Otherwise, the forces do not lie along the line, because the friction can only contribute Nv_0 in the direction parallel to the surface, where N is the applied force perpendicular to the surface and v_0 is the coefficient of friction.



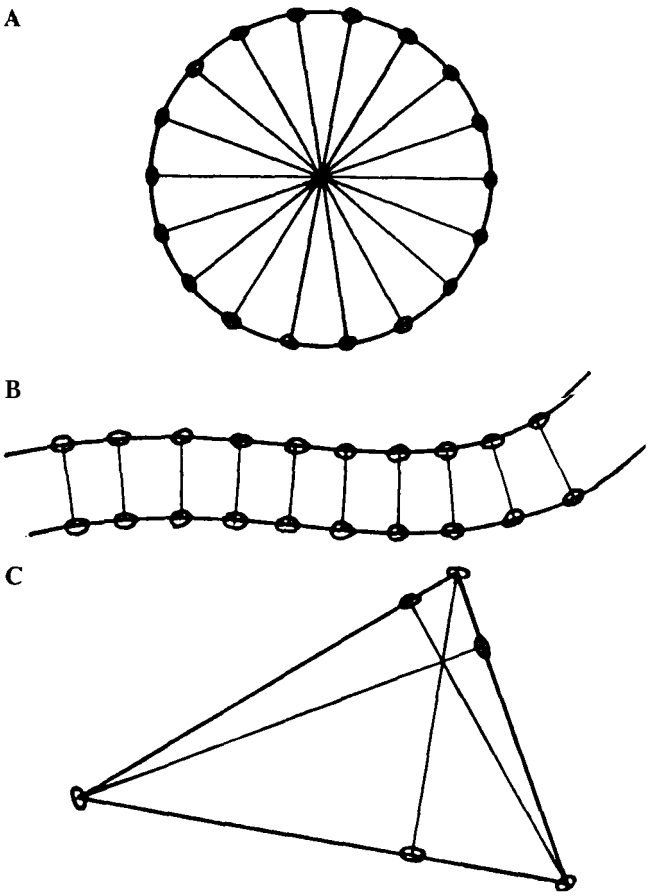
The silhouette is of particular interest for a smoothly curved object, since it can be determined from the image. There, the surface normal is parallel to the image plane and perpendicular to the silhouette in the image.

At some points—for example, at a crease in an object—the surface orientation may vary discontinuously with the position on the surface. We cannot use such points as the first point *P* in the above algorithm, because we cannot determine the surface orientation there. Such a point may be used for grasping, if it happens to be found as the second point Q' in the above algorithm, when starting from some other initial point *P'*.

Examples of legal grasp points on various objects are shown in Fig. 14. At this stage, the gripper's shape is treated simply as a pair of points. The attitude in space of the gripper is not fully defined at this point; only the direction of the line between the two grasping points, the normal direction of the gripper, is known.

The gripper has another degree of freedom; it can rotate about the line connecting the two grasping points. The range of rotation about this axis is constrained by the shape of the gripper and the shape of the object. We will call this degree of freedom the legal *rotation* of the gripper. The legal grasp configuration is a general name for the legal grasp points *and* the legal grasp rotation. If we use the point halfway between the grasping points to represent the position of the gripper and the direction halfway between the boundaries of the object to represent the direction of the gripper, then our legal grasp configuration becomes identical to Lozano-Perez's Legal Grasp Position (GSETS)(Lozano-Perez 1976; 1981).

Fig. 14. Examples of legal grasp points.

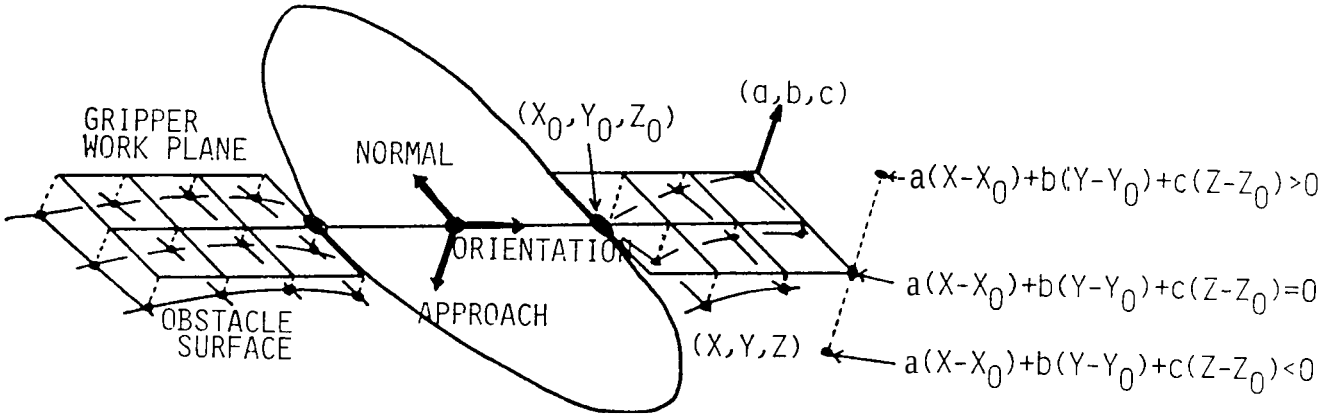


3.3. Collision Free Configurations

Legal grasp configurations only describe the relationship between the gripper and the object grasped. Among these legal grasp configurations, we have to choose a grasp configuration that can be achieved without hitting other objects,

One approach to doing this is using the method of configuration space obstacles (CSO)(Lozano-Perez 1981; 1983). CSO uses an equivalent representation in which the obstacles are enlarged and the gripper is reduced to a point. We do not follow this approach, since the number of neighboring obstacles in bin-

Fig. 15. Gripper work space and obstacle surface. The grasp motion sweeps out a pair of rectangular volumes which will be occupied by the fingers. The bottom faces of these volumes pass through the legal grasp points. One of these faces lies in a plane $a(x-x_0)+b(y-y_0)+c(z-z_0)=0$ *where* (x_0, y_0, z_0) *is one of the legal grasp points. and* $(-a,-b,-c)$ *is the gripper approach direction at the legal grasp configuration. We check z values (elevation supplied by binocular stereo) within the two rectangular footprints to see that they are below this plane. If any point is over the plane* $(a(x-x_0)-(y-y_0)-(z-z_0)>0)$*, the gripper will collide in that configuration. If all points over the rectangular area are below the plane* $(a(x-x_0)-(y-y_0)-(z-z_0)<0)$*. then the configuration is a collision-free configuration. One may measure the finger clearance by the distance from the plane to the highest point of the obstacles.*



picking tasks can be quite large and the computation of the CSOs correspondingly expensive. Also, the obstacles typically overlap so that individual CSOs must be combined to make composite CSOs.

Instead, we use a direct method. The central idea is to check every candidate grasp configuration among the legal grasp configurations, one after another, to see whether or not the gripper will hit an obstacle in that configuration.

The grasp motion sweeps out a pair of rectangular volumes that will be occupied by the fingers. The inner faces of these volumes pass through the legal grasp points: their orientation is determined by the legal grasp rotation; their width and thickness correspond to the dimensions of the fingers. We will check whether these rectangular areas intersect the other objects or not.

Each of the rectangular areas lies in a plane,

$$a(x-x_0)+b(y-y_0)+c(z-z_0)=0,$$

where (x_0, y_0, z_0) is one of the legal grasp points and $(-a, -b, -c)$ is the gripper approach direction at the legal grasp configuration. We check z values, elevation supplied by binocular stereo, within the two rectangular footprints to see whether they are below this plane. If any point is over the plane, the gripper will collide in that configuration. Conversely, if the left hand side of the above equation is less than zero for all points in the footprint, then the configuration is a collision free configuration (See Fig. 15). We may even choose the best grasping configuration in the sense of the one in which the highest point of the obstacles has the lowest height relative to the rectangular areas representing the gripper jaws.

4. Planning and Grasping

This chapter shows examples of how to apply the basic theories described above to picking up an object. The previous vision modules provide the position and the attitude of the target object and depth information around the object. From this information a grasp configuration will be determined by using the theory discussed in Chapter 3.

4.1. EXAMPLE 1 (Simple minded strategy)

Choosing the highest point of the target object as the grasp point minimizes the likelihood of collision with neighboring objects. We will follow this strategy at this time. The position of the highest point is determined analytically from the attitude of the object obtained by using photometric stereo and matching orientation histograms. The pickup point is selected as the legal grasp point at the highest evaluation.

The execution of the pickup operation is shown in Fig. 16. Note that the manipulator approaches the doughnut shaped object directly from the initial configuration. The system described earlier required an additional arm motion (Ikeuchi, Horn, Nagata, Callahan, and Feingold 1983).

Fig. 16 Pickup motion by the PUMA arm.

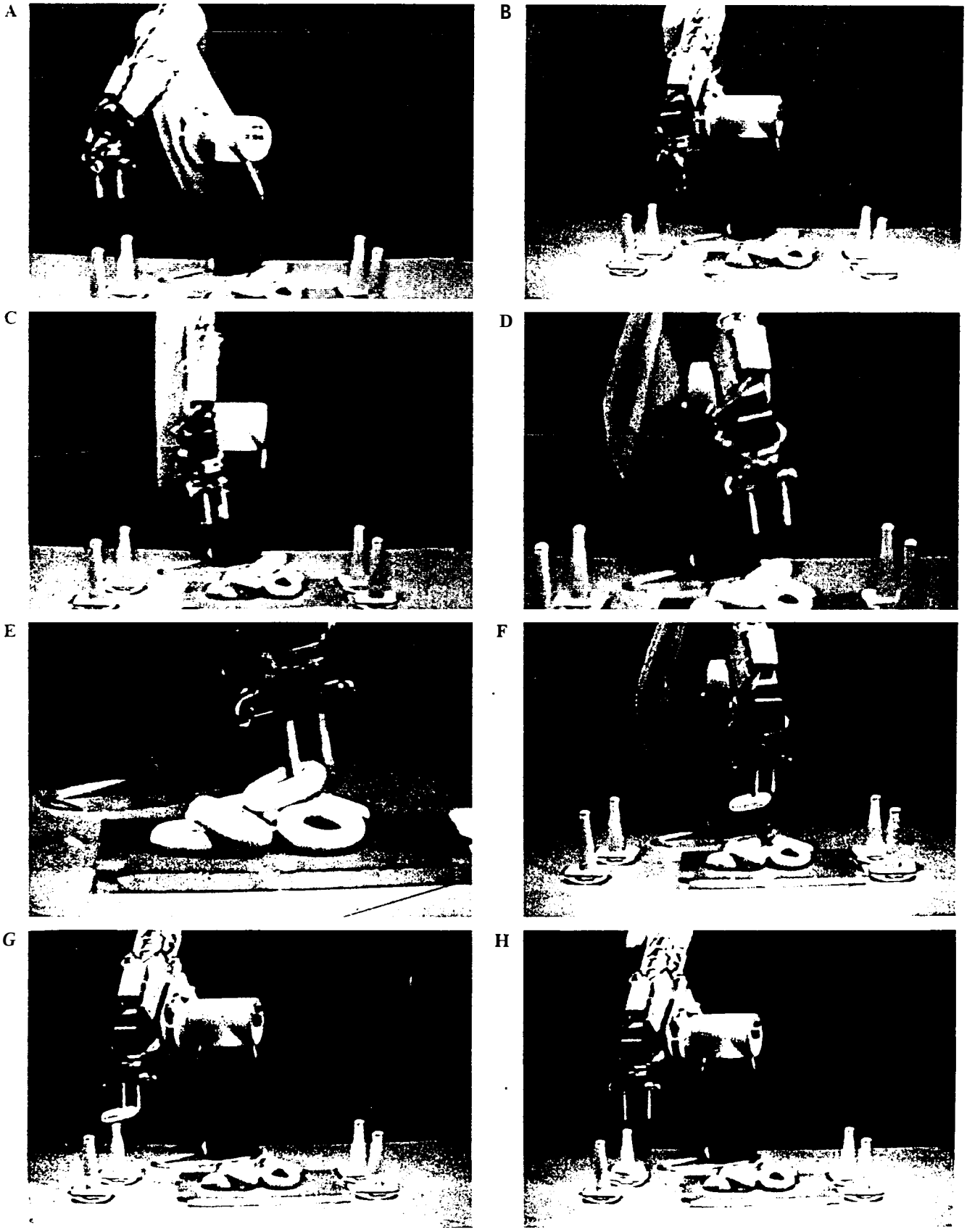


Fig. 17. The legal grasp points of a doughnut.

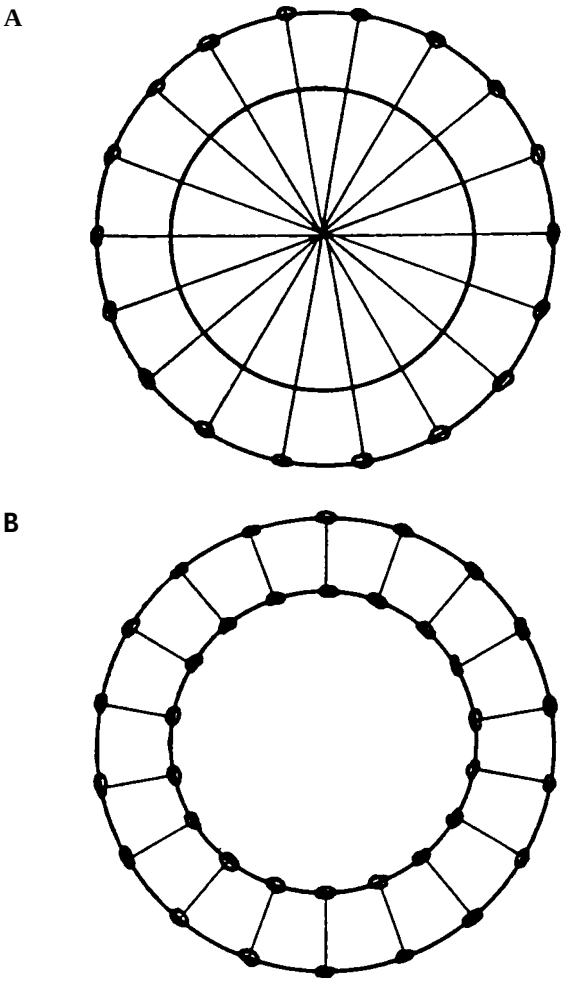


Fig. 18. The legal grasp configurations and the world coordinate. The gripper approach direction, $(-a, -b, -c)$ is $(-D_x, -D_y, -D_z)$ where (D_x, D_y, D_z) is a normal vector perpendicular to the plane, the gripper normal direction is the direction from the grasp point to the center of the 2D ring specified by α.

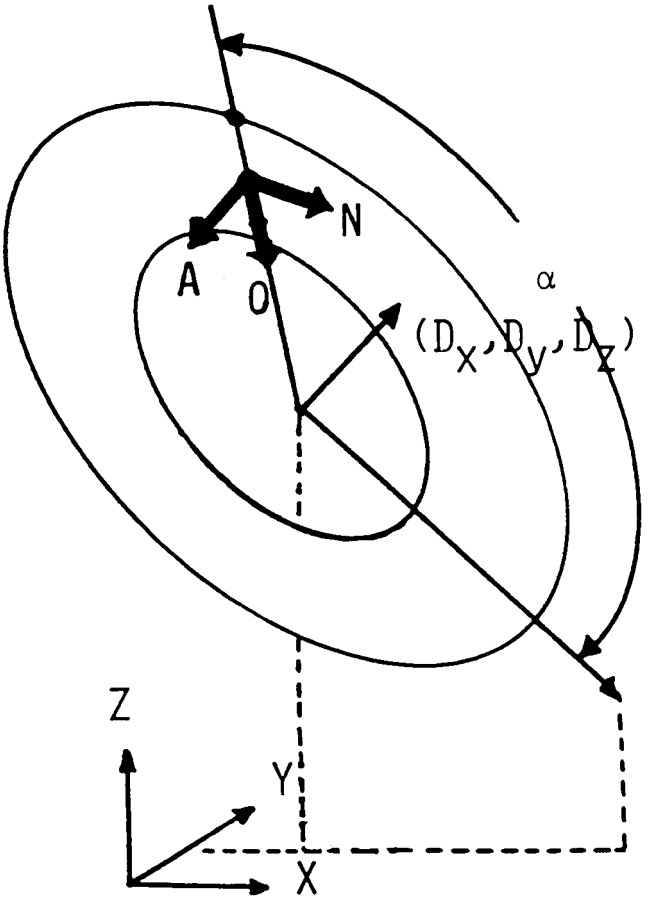
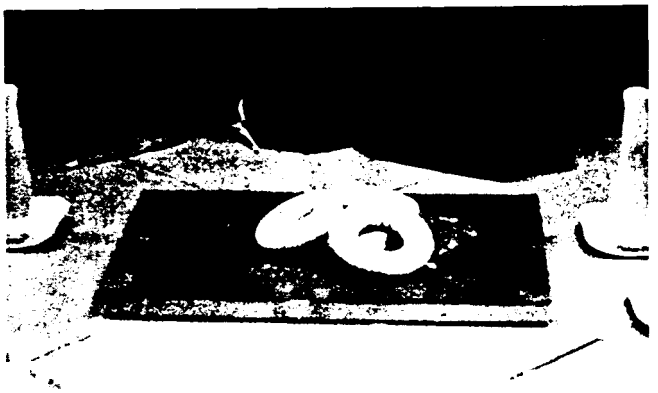


Fig. 19. A more difficult case where simple minded strategy would fail.



4.2. EXAMPLE 2 (Two-dimensional model)

Determining a grasp configuration requires: (1) extracting legal grasp configurations from a model; **(2)** determining legal grasp points using the observed data; **(3)** finding a collision-free configuration from an observed depth map. The first task is to extract legal grasp configurations from an object model. ***This*** example models the doughnut shape as a two-dimensional ring. We will refer to the plane on which the two-dimensional ring exists as the ***object plane.*** Two classes of legal grasp points are extracted from the discussion in Section **3.2** and shown in Fig. 17. Legal

Fig. 20. Profile of highest points over the rectangular area of the gripper work space along legal grasp configurations specified by the rotation angle α. Note that highest here means the largest value of $a(x-x_0)+b(y-y_0)+c(z-z_0)$, and the gripper workplane corresponds to the horizontal line. If the highest point at a configuration is below the gripper workplane, the configuration is a collision-free configuration and the doughnut can be picked up using the configuration. The box mark shows the optimal grasp configuration which gives the greatest finger clearance among the collision-free configurations.

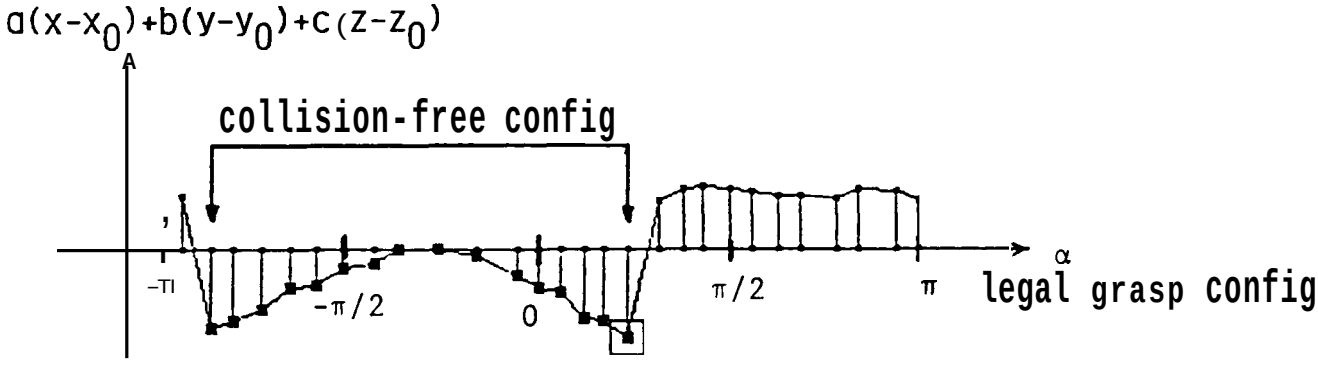


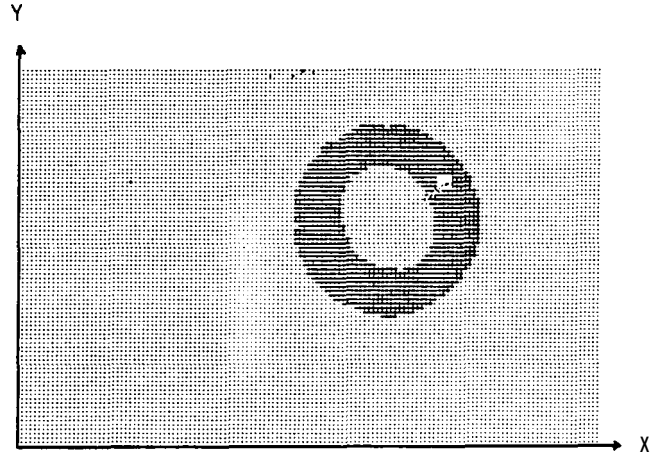
Fig. 21. The center position of the grasp points determined.

grasp points of Class 1 (Fig. 17a) require too large a gripper opening so they are discarded. Legal grasp points of Class 2 (Fig. 17b), on the other hand, can be used.

A unique legal grasp configuration is determined at each legal grasp point. Since the direction perpendicular to the plane of the doughnut is the only possible legal rotation, due to the doughnut shape, a legal grasp configuration is determined at each legal grasp point. Namely, the gripper approach direction, $(-a, -b, -c)$ is $(-D_x, -D_y, -D_z)$, where (D_x, D_y, D_z) is a normal vector perpendicular to the plane; the gripper normal direction is the direction from the grasp point to the center of the two-dimensional ring (See Fig. 18).

The next task is to establish the relationship between the legal grasp points in the model and the observed data. In the two-dimensional model, legal grasp points occur only along the silhouette of the object. Fortunately, the silhouette of the object has already been extracted by the segmentation process. Each silhouette point, which corresponds to a legal grasp point, can be specified by the rotation angle α around the normal to the two-dimensional ring as shown in Fig. 18. In other words, this rotation angle α can denote each legal grasp point and, thus, each legal grasp configuration in the observed data.

The final task is to find collision-free configurations among the legal grasp configurations. For each legal grasp configuration, we check the corresponding rectangular regions for the distance to which the fingers can be moved past the plane of the doughnut before a collision occurs. The equation requires (x, y, z), (x_0, y_0, z_0), and the gripper approach direction. Each



legal grasp configuration specified by α gives the legal grasp point (x_0, y_0, z_0), the approach direction $(-a, -b, -c)$, and the rectangular footprints. The depth map from the PRISM stereo gives (x, y, z).

While the simple mind strategy in Example 1 often identifies a collision-free configuration, it can easily fail as illustrated by the example in Fig, 19. The strategy based on the two-dimensional model will be applied to the scene. A profile of the highest points over the rectangular area of the gripper footprint with respect to the gripper workplane, and along legal grasp configurations specified by the rotation angle α is shown in Fig. 20. Note that highest here means the largest value of $a(x-x_0)+b(y-y_0)+c(z-z_0)$, and the gripper workplane corresponds to the horizontal line in Fig. 20. If the highest point at a configuration is below the gripper workplane, the configuration is a collision-free configuration and the doughnut can be

Fig. 22. The pickup motion determined by the second strategy. (b) shows the original configuration selected with the simple mind strategy which would have resulted in a collision. In (c), the gripper is rotated around the axis of the doughnut, so that the gripper configuration agrees with the optimal configuration selected in Fig. 20.

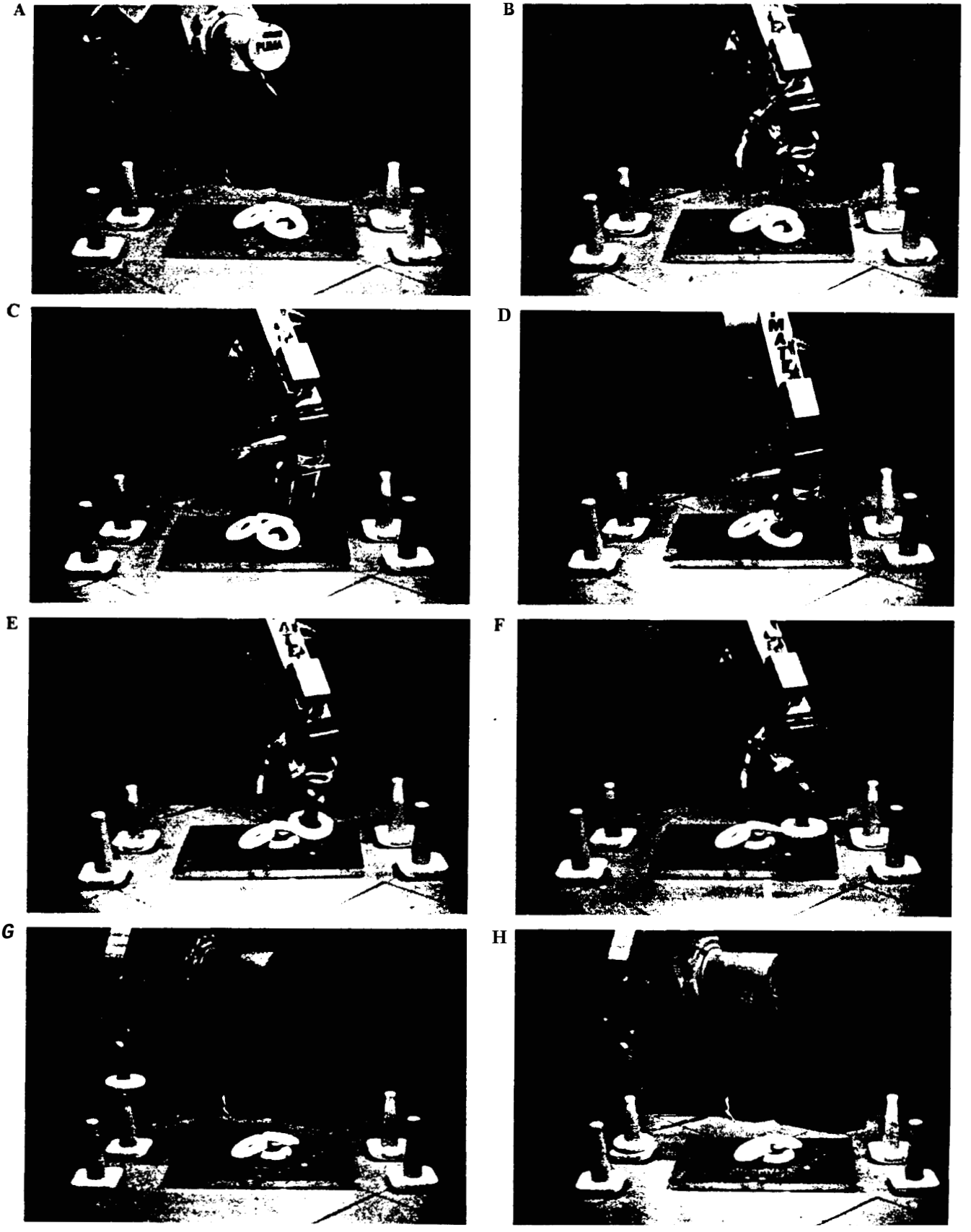
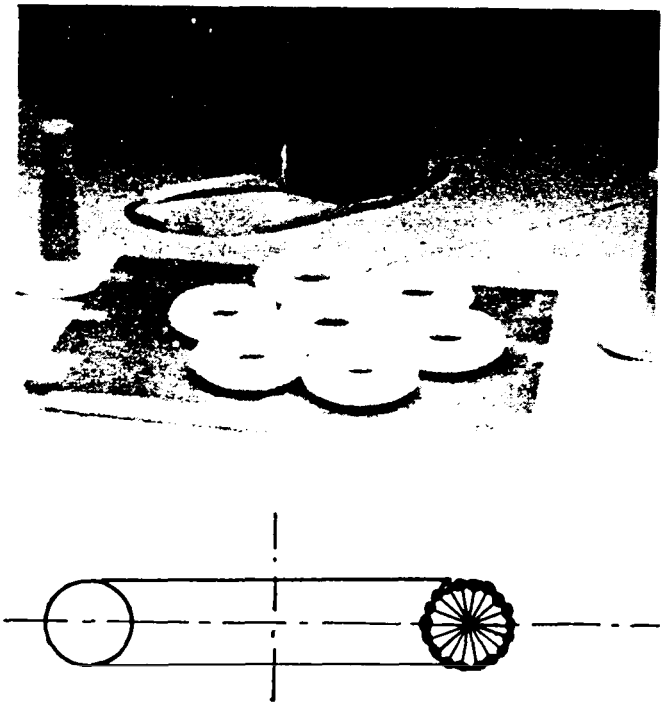
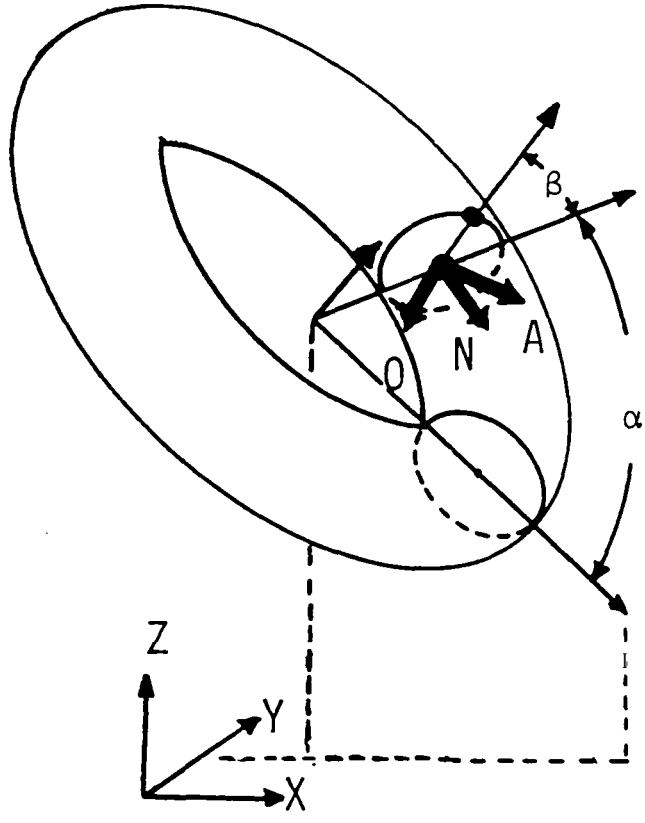


Fig. 23. A situation where 2D model strategy would fail.

Fig. 24. Additional legal grasp points for a 3D doughnut (cross-sectional view).

Fig. 25. The legal grasp configurations and the world coordinate. The legal grasp Configuration can be characterized using two parameters, α and β. The first parameter, α, denotes the rotation around the axis of the doughnut to indicate an axial plane on which both the approach direction and the orientation direction exist. The symbol β indicates the angle between the orientation direction and the doughnut plane on the axial plane or, equivalently, the angle between the approach direction and the doughnut attitude on the axial plane. Legal grasp configurations in the 2D model correspond to the case where β is zero.





picked up using the configuration. The box mark in Fig. 20 shows the optimal grasp configuration which gives the greatest finger clearance among the collision-free configurations. The center position of the gripper selected is shown in Fig. 21, and a pickup sequence using our second strategy is shown in Fig. 22. The original configuration selected with the simple mind strategy which would have resulted in a collision is shown in Fig. **22b.** In Fig. 22c, the gripper is rotated around the axis of the doughnut so that the gripper configuration agrees with the optimal configuration selected in Fig. 20.

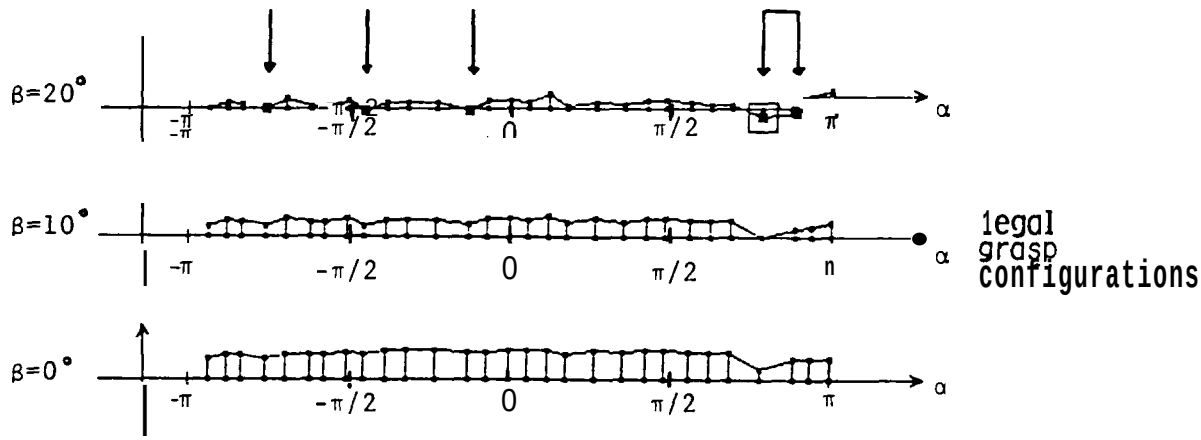
4.3. EXAMPLE 3 (Three-dimensional model)

The doughnut in the middle, in the example shown in **Fig. 23,** could not be picked up using two-dimensional model strategy. That doughnut is surrounded by obstacles and there is no position around its circumference with sufficient clearance for the fingers to get below the plane of the doughnut. In cases like this, it is still possible to find a collision-free grasp configuration, but it is necessary now to model the doughnut as a three-dimensional object.

Three classes of legal grasp points are extracted from the three-dimensional model of a doughnut, namely, those shown in Fig. 17 and the additional one shown in Fig. **24.** The only possible legal rotation is that the gripper approach direction exists on the axial plane. The legal grasp configuration can be characterized as using two parameters, α and β. The first parameter α denotes the rotation around the axis of the doughnut to indicate an axial plane on which both the approach direction and the orientation direction exist. The second parameter β indicates the angle between the orientation direction and the doughnut plane on the axial plane or, equivalently, the angle between the approach direction and the doughnut attitude on the axial plane (See Fig. 25.). Legal grasp configurations in

Fig. 26. The profile of the highest points with respect to the work space plane over legal grasp configurations. If a configuration has the value below the horizontal line, then the configuration is a collisionfree configuration. The optimal configuration, marked as a box, has the greatestfinger clearance among the collision-free configurations.

Fig. 27. The center position of the grasp points selected.



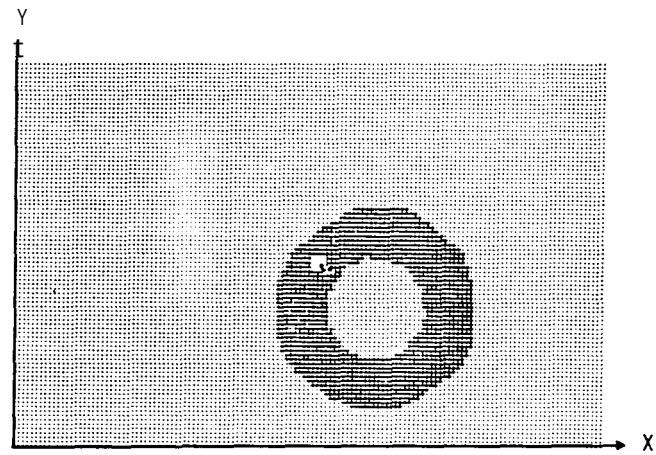
the two-dimensional model correspond to the case where β is zero.

The next task is to determine legal grasp points using the observed data. In the three-dimensional model, legal grasp points occur over the doughnut surface. Each legal point can be determined by using α, β, and the silhouette of the object extracted.

The final task is to find collision-free configurations among legal grasp configurations. The profile of the highest points with respect to the work space plane over legal grasp configurations is shown in Fig. 26. If a configuration has the value below the horizontal line in Fig. 26, then the configuration is a collision-free configuration. The optimal configuration, marked as a box in Fig. 26, has the greatest finger clearance among the collision-free configurations. The center position of the gripper selected is shown in Fig. **27.** The process of picking up a doughnut without collision is shown in Fig. 28. The original grasp configuration selected using the simple mind strategy is depicted in Fig. 28b. In Fig. 28c the gripper rotates around the axis of the doughnut; then in Fig. 28d the gripper rotates around the normal direction of the gripper so that the gripper configuration agrees with the optimal configuration in Fig. 26.

5. Concluding Remarks

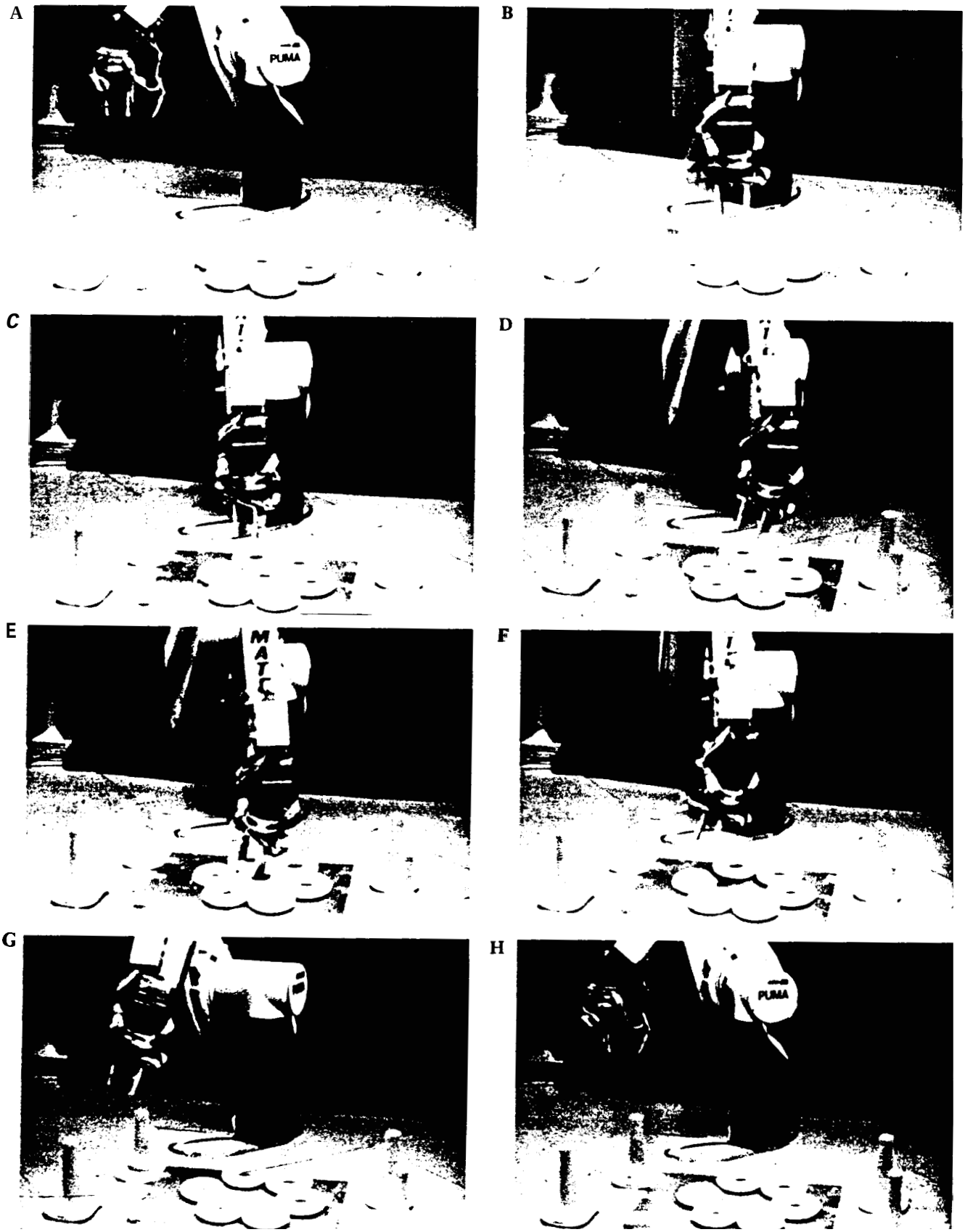
We have described a hand-eye system which performs bin-picking **tasks.** Four basic modules are used: photo-



metric stereo, binocular stereo using the PRISM algorithm, extended Gaussian image matching, and collision-free configuration planning for the gripper.

Photometric stereo determines the orientation of surface patches corresponding to each picture cell, based on the brightness values in three images and obtained using different light sources. Segmentation is based on a needle diagram, the smoothness constraints, shadow areas, and mutual illumination. The attitude in space of the object is determined by comparing the orientation histogram of the object's surface with stored orientation histograms of prototypes. The orientation histogram is a discrete approximation of the extended Gaussian image. An elevation map produced by the PRISM stereo algorithm is used to deter-

Fig. 28. The pickup motion. Position (b) is the original grasp position selected using the simple minded strategy. In (c) the gripper is rotated around the axis of the doughnut; in (d) the gripper rotates around the normal direction of the gripper so that the gripper configuration agrees with the optimal configuration in Fig. 26.



mine object elevation and to check finger clearance at the proposed grasp configurations.

The two low-level vision modules produce reliable but restricted information about the visible surfaces imaged. In one case, high resolution local surface orientation measurements; in the other, absolute height measurements at a lower spatial resolution. We have combined these two systems to produce one that takes advantage of both and to solve a problem that neither system could solve well alone.

Our visual system also possesses various modules to recover three-dimensional information such as shape from shading, shape from texture, shape from motion, and shape from binocular stereo. The questions are whether each module produces a unified representation or independent representations to be interpreted by a higher module; and the kind of control structure used to establish harmony among these modules. We have to explore the hybrid mechanism to give answers to these questions.

The present form extracts EGI from a mathematical model. For more complicated objects, this method could be difficult though still possible. Usual machine parts can be modeled by a CAD model and its CAD data can often be available. It is necessary to explore how to use CAD models for extracting EGI and other matching features efficiently (Brooks 1981; Bolles, Horand, and Hannah 1984; Ikeuchi 1985). Both the photometric and PRISM stereo modules have simple kernels that can be easily adapted for use in other problems and that lend themselves to high-speed implementation on special purpose hardware. This special purpose hardware design should be explored for real applications.

Acknowledgment

The following people also helped in part on this project: Noble Larson constructed the CCD-TV camera interface; John Purbrick built the fingers of the gripper; Oded Feingold and John Cox made the LED sensor and collision sensor; Tom Callahan prepared the lighting stage and general setup. Without their effort, we could not have completed this project.

Tomas Lozano-Perez and the referees provided many useful comments which have improved the readability of this paper. Finally, we thank Ikko Ikeuchi for preparing this manuscript and some of the drawings.

This paper describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by; the Office of Naval Research under Office of Naval Research contract N000 14-77-C-0389, the System Development Foundation, and the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N000 14-80-C-0505.

References

Bajcsy, R. 1980, Miami Beach. Three-dimensional scene analysis, *Proc. 5th ICPR.* 1064–1074.

Ballard, D. H., and Sabbah, D. 1981, Vancouver. On shapes. *Proc. 7th IJCAI,* 607–612.

Birk, J. R., Kelley, R. B.. and Martins, H. A. S. 1981. An orienting robot for feeding workpieces stored in bins. *Trans. IEEE. SMC,* Vol. SMC-11; No. 2, 151–180.

Boissonnat, J. D. 1982. Stable matching between a hand structure and an object silhouette. *Trans IEEE. PAMI,* Vol. PAMI-4; No. 6, 603–612.

Bolles, R. C., Horand, P., and Hannah, M. J. 1984. 3DPO: a three-dimensional *parts* orientation system. *Proc. Int. Symp. Robotic Res. 1.* Brady, M. and Paul, P., eds. 413–424. MIT Press.

Brady, M. 1982. Parts description and acquisition using vision. *Proc. SPIE. Robot Vision:* Vol. 336, 20–28,

Brooks, R. 1981. Symbolic reasoning among 3D models and 2D images. *Artificial Intelligence,* Vol. 17, No. 1–3: 285–348.

Brou, P. 1984. Using the Gaussian image to find the orientation of objects. *Int. J. Robotics Res.* Vol. 3, No. 4: 89–125.

Coleman, E. N., and Jain, R. 1981, Vancouver. Shape from shading for surfaces with texture and specularity. *Proc. 7th IJCAI.* 652–657.

Do Carmo, 1976. *Differential geometry of curves and surfaces.* Englewood Cliffs, N.J.: Prentice-Hall.

Hanafusa, H., and Asada, H. 1977. Stable prehension by a robot hand with elastic fingers. *Proc. 7th ISIR.* 361–368.

Horn, B. K. P. 1975. Obtaining shape from shading information. *The Psychology of Computer Vision.* Winston, P. H., ed. 115–155. New York: McGraw-Hill.

Horn, B. K. P. 1977. Image intensity understanding. *Artificial Intelligence,* Vol. *8,* No. 2: 201–231.

Horn, B. K. P., Woodham, R. J., and Silver, W. M. 1978. Determining shape and reflectance using multiple images. *AI Memo No. 490*. Cambridge; Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Horn, B. K. P., and Sjoberg, R. W. 1979. Calculating the reflectance map. *Applied Optics*. Vol. 18: 1770–1779,

Horn, B. K. P. 1983. Extended Gaussian images. *AI Memo 740*. Cambridge: Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Ikeuchi, K. 1981a, Vancouver. Recognition of 3D object using extended Gaussian image. *Proc. 7th IJCAI*. 595–600.

Ikeuchi, K. 1981b. Determining surface orientations of specular surfaces by using the photometric stereo method. *Trans. IEEE on PAMI*. Vol. PAMI-?, No. 6: 661–669.

Ikeuchi, K. 1983. Determining attitude of object from needle map using extended Gaussian image. *AI Memo 714*. Cambridge: Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Ikeuchi, K., Horn, B., Nagata, S., Callahan, T., and Feingold, O. 1983. Picking up an object from a pile of objects. *AI Memo 726*. Cambridge: Massachusetts Institute of Technology, Artificial Intelligence Laboratory. Also available as *Proc. Int. Symp. Robotics Res. 1*, Brady M. and Paul, R. P., eds: 139–162. Cambridge: MIT Press.

Ikeuchi, K. 1985. A vision system for bin-picking tasks guided by an interpretation tree from a CAD model. *Inf. Processing Soc. Japan*, CVWG-38-6. (Japanese).

Little, J. 1985. *Recovering shapes and determining attitude from extended Gaussian image*. Ph. D. Thesis. University of British Columbia, Department of Computer Science.

Lozano-Perez, T. 1976. The design of a mechanical assembly system. *AI-TR-397*. Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Lozano-Perez, T. 1981. Automatic planning of manipulator transfer movements. *Trans. Sys. Man, Cyber. IEEE* SMC-I 1,681–698.

Lozano-Perez, T. 1983. Spatial planning: a configuration space approach. *Trans. Computers* C-32 *IEEE:* 108–120.

Marr, D., and Hildreth, E. 1980. Theory of edge detection. *Proc. R. Soc. Lond.* B. Vol. 207: 187–217.

Marr, D., and Poggio, T. 1979. A computational theory of human stereo vision. *Proc. R. Soc. Lond, B.* Vol. 204: 301–328. Also available as *AI Memo 451*. Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Nishihara, H. K. 1984. PRISM: a practical realtime imaging stereo matcher. *Optical Eng.* Vol. 23: 536–545. Also available as *AI Memo* 780. Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Nishihara, H. K., and Larson, N. G. 1981 April. Towards a real time implementation of the Marr-Poggio stereo matcher. *Proc. A.R.P.A Image Understanding Workshop*, L. S. Baumann, ed. Science Applications, Inc.

Nishihara, H. K., and Poggio, T. 1984. Stereo vision for robotics. *Proc. Int. Symp. Robotics Res. 1:* 439–505. Brady, M. and Paul, R. P., eds. Cambridge: MIT. Press.

Silver, W. A. 1980. *Determining shape and reflectance using multiple images*. MS Thesis. Cambridge: MIT, EECS.

Smith, D. A. 1979. Using enhanced spherical images for object representation. *AI Memo* 530, Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Tsuji, S., and Nakamura, A. 1975, Georgia, USSR. Recognition of an object in a stack of industrial parts. *Proc. 4th IJCAI:* 811–818.

Woodham, R. J. 1978 August. Photometric stereo: a reflectance map technique for determining surface orientation from a single view. *SPIE 22nd Annual Tech. Symp., Image Understanding Systems and Industrial Applications*. Vol. 155: 136–143.

Woodham, R. J. 1979. Reflectance map techniques for analyzing surface defects in metal casting. *AI-TR-457*. Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

Woodham, R. J. 1980. Photometric method for determining surface orientation from multiple images. *Optical Eng.* Vol. 19, No. 1: Jan/Feb., 139–144.