Massachusetts Institute of Technology

Artificial Intelligence Laboratory

A.I. Memo No. 726 May, 1983

Picking up an object from a pile of objects

Katsushi Ikeuchi, Berthold K.P. Horn, Shigemi Nagata

Tom Callahan, and Oded Feingold

Abstract

This paper describes a hand-eye system we developed to perform the bin-picking task. Two basic tools are employed: the photometric stereo method and the extended Gaussian image. The photometric stereo method generates the surface normal distribution of a scene. The extended Gaussian image allows us to determine the attitude of the object based on the normal distribution.

Visual analysis of an image consists of two stages. The first stage segments the image into regions and determines the target region. The photometric stereo system provides the surface normal distribution of the scene. The system segments the scene into isolated regions using the surface normal distribution rather than the brightness distribution. The second stage determines the object attitude and position by comparing the surface normal distribution with the extended-Gaussian-image.

Fingers, with LED sensor, mounted on the PUMA arm can successfully pick an object from a pile based on the information from the vision part.

Keywords: Photometric stereo, Extended Gaussian image, Puma arm, Bin-picking, Hand-eye system, Visual guidance

Acknowledgment: 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 Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-80-C-0505 and in part by the Office of Naval Research under Office of Naval Research contract N00014-77-C-0389.

1. Introduction

Sensory capabilities will extend the functional range of robots. Without sensing the outer world, robots can only repeat pre-programmed tasks. Thus, the task is very rigid; such a system cannot overcome any small disturbance. Therefore, sensory capability is an essential component of a flexible robot.

Vision could be the most important type of robotic sensor. Since a vision sensor is a non-contact sensor, we can get the necessary input information without disturbing the environment. Also vision can acquire global information about a scene. This is not the case for the tactile sensor.

There are basically three tasks where the vision feedback could play a essential role:

(i) finding the target object and determining the grasping point,

(ii) bringing an object from an initial point to a destination point while avoiding collision with other objects,

(iii) assembling something using the object brought.

This report explores the method of manipulator visual guidance in domain (i). A manipulator without vision can only pick up an object whose position and attitude is pre-determined. Such a system needs the help of another machine or human for feeding objects at a pre-determined place in a pre-determined attitude. Since this feeding job is tedious, the job is quite unsuitable for a human being. Some research has aimed at solving this feeding problem by introducing mechanical vibration methods. These methods may cause defects in objects due to collisions. This paper proposes a method to solve this problem by visual guidance of manipulators instead.

There are two problems in this task. The first one is how to isolate an object from the background. The second one is how to determine the attitude of the object relative to the camera. Historically, these two problems have been attacked by detecting brightness changes [Tsuji and Nakamura, 1975, Yachida and Tsuji, 1975, Baird, 1977, Perkins, 1977, Shirai, 1978, Bolles and Cain, 1982]. Detecting brightness changes gives boundaries between regions corresponding 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 lying on a uniform background provided the objects only rotate in the plane of support. In other words, these algorithms work well on binary images. However, such methods cannot extract the contour of an object from the image of a set of overlapping objects, which is the typical case in bin-picking.

The extracted contours should also be compared with some internal model to determine the attitude. A contour is usually approximated as a collection of straight lines. The internal model consists of the lengths of each segment and the angles between segments [Price and Reddy, 1979, Davis, 1979, Sakane, 1981]. This approach cannot treat 3D curved objects.

Kelley and others [Kelley, 1982] highlight scenes to segment and determine the position and orientation of an object in a bin. This system is limited to cylindrical

workpieces with metallic surface. Also, their vision system only determines two degrees out of three degrees of freedom in attitude.

Recent work in image understanding [Barrow and Tenebaum, 1978, Marr, 1982, Brady, 1982, Ballard, 1982, Grimson, 1981] has led to techniques for computing local surface gradient. Such methods include shape from shading [Horn, 1975, Ikeuchi and Horn, 1981, Terzopoulos, 1982], shape from photometric stereo [Woodham, 1978, Silver, 1980, Ikeuchi, 1981b, Coleman et al., 1981], shape from texture [Kender, 1980, Witkin, 1981, Kanade, 1981, Kender and Kanade, 1980, Ikeuchi, 1980, Ohta et al., 1981], and shape from polarization [Koshikawa, 1979]. The local gradient representation is called a needle map [Horn, 1979] or $2\frac{1}{2}D$ sketch [Marr, 1982, Nishihara, 1981]. Since this local information is obtained over a region, it is more robust than silhouette information which comes only from the boundaries. The use of this robust local information should be explored.

This paper attacks the bin-picking problem using a needle map obtained by photometric stereo for both segmentation and attitude-determination. The segmentation is done with a needle map rather than a brightness map. An extended Gaussian image (EGI) [Smith, 1979, Bajcsy, 1980, Ikeuchi, 1981, Ballard and Sabbah, 1981] obtained from the surface normal distribution of an isolated region determines the object attitude.

2. Overview of the system

2.1. Problem

The basic assumptions are as follows:

(i) *The objects stored in a pile are all of the same kind.*
In order to simplify the problem, we introduce this assumption. We do not worry about "recognition" in this paper. The system, however, can be easily extended to treat multiple kinds of objects [Ikeuchi, 1981a].

(ii) *The surface material of the object is known.*
The photometric stereo method determines the surface orientation using the relationship between apparent brightness and the surface orientations. Thus, it is necessary to know the surface material.

(iii) *The object shape is known.*
There are two reasons for this assumption. One is that we segment an image using rapid changes in surface orientation. Also the EGI system determines the object attitude by comparing the observed EGI with prototypical EGI's. The prototypical EGI's are obtained from the object shape.

2.2. Strategy

Basically, the problem consists of two tasks: (1) segmenting a scene into isolated regions and finding the target region and (2) determining the position and attitude of the object. The basic strategy for these tasks is as follows:

(i) Obtain a needle map using photometric stereo.

(ii) Segment a scene into isolated regions using the needle map.

(iii) Use the EGI to determine the object attitude.

(iv) Determine x-y position of the grasping point using the target region's position in the image.

(v) Determine the absolute depth of the grasping point using a proximity sensor.

3. Basic Tools

3.1. Photometric Stereo

3.1.1. Reflectance Map

The reflectance map represents the relationship between surface orientation and image brightness [Horn, 1977, Woodham, 1978, Horn and Sjoberg, 1979]. Generally speaking, apparent brightness of a surface patch depends on four components: (1) surface material, (2) light source direction, (3) viewer direction, and (4) surface normal. Let us assume that every direction is expressed with respect to the direction towards the viewer. This coordinate system is called a viewer-centered system. If we represent both light source direction and surface material in this coordinate system, the apparent brightness is a function of surface normal direction. This function is called a reflectance map.

The reflectance map can be expressed on a stereographic plane. Since surface normal directions have two degrees of freedom, we can express the relationship between the apparent brightness and surface normal directions on a 2-D plane. Here points on the 2-D plane represent surface normals. At each points, the apparent brightness value, which can be obtained either empirically[Woodham, 1979] or theoretically [Horn and Sjoberg, 1979], is plotted. We use the stereographic plane for expressing surface orientations [Ikeuchi and Horn, 1981]. Thus, the brightness distribution for known light source direction and surface material can be shown as height over the stereographic plane.

3.1.2. Photometric Stereo Method

The photometric stereo method takes multiple images of the same scene from the same position with different illumination [Woodham, 1978, Silver, 1980, Ikeuchi, 1981b, Coleman and Jain, 1981]. This setup gives multiple brightness values at

each image point. Since different images are taken from the same point, there is no disparity between images as with binocular stereo; so there is no need to match multiple images.

The photometric stereo method determines the surface orientation from the brightness values at each point. Each illumination situation gives a unique reflectance map, because the reflectance map depends on illumination direction. We can track the iso-brightness contour corresponding to the obtained brightness value on each reflectance map. The contour represents possible surface orientations. Since there is no change in the relationship between the viewer and the surface orientation, the orientation should correspond to the same point in the different reflectance maps. Thus, the intersection point of the iso-brightness contours gives the surface orientation. Typically, a pair of contours intersect at two points due to non-linearity. The third brightness value can resolve this ambiguity.

This method is usually implemented using a lookup table. If we assume both the viewer and the light source are far from the object, then both the light source directions and the viewer direction are assumed constant over the image. Thus, for each light source we can use a single reflectance map everywhere in the image. The points where iso-brightness lines cross can be pre-calculated. The crossing points can be registered in a table. The table can convert triples of brightness values into surface orientation. The indices to this lookup table correspond to the brightness values for each light source direction. The contents of the table are surface orientations corresponding to the brightness values. Since the calculation is performed by table lookup, the surface orientation at each point can be determined very rapidly.

3.2. EGI Matching

3.2.1. Extended Gaussian image

Roughly speaking, the extended Gaussian image of an object is a spatial histogram of its surface normal distribution. Let us assume that there is a fixed number of surface patches per unit surface area and that a unit normal is erected on each patch. The collection of these normals are like porcupine quills [Horn, 1979]. These vectors can be moved so that their "tails" are at a common point and their "heads" lie on the surface of a unit sphere. This mapping is called the Gauss map; the unit sphere is called the Gaussian sphere [Do Carmo, 1976]. If we attach a unit mass to each end point, we will observe a distribution of mass over the Gaussian sphere. The resulting distribution of mass is called the extended Gaussian image (EGI) of the object [Smith, 1979].

Let us define a visible hemisphere. Commonly, one observes an object from one direction. So we always obtain only one half of the EGI over a Gaussian hemisphere. This hemisphere will be referred to as the visible hemisphere [Ikeuchi, 1983]. The axis of symmetry of the visible hemisphere corresponds to the line of sight. Each point on the visible hemisphere corresponds to a surface orientation whose angle with the line of sight is no more than $\frac{\pi}{2}$. In the following discussion we will work

with this EGI over the visible hemisphere. Also we will normalize the distribution of EGI mass to have unit mass over the visible hemisphere.

3.2.2. Freedom of matching

An apparent image of an object varies with the following factors: (a)translation of the object, (b)expansion of the object, (c)rotation of the object. The normalized EGI is independent of (a) and (b). The EGI rotates in the same way as (c) as will be shown next:

(a) Neither the surface normals nor the Gauss map depend on the position of the origin. Thus, the resulting EGI is not affected by translation of the object.

(b) If the object expands, the total mass over the Gaussian hemisphere increases. Yet, the EGI mass is normalized to have unit mass over the hemisphere. Thus, the normalized EGI does not change with object expansion. This characteristic is very convenient in object recognition. In general, the distance between the TV camera and the object changes in each situation. Thus, the apparent size of an object will also vary, but the normalized EGI derived from the image is independent of the apparent size.

(c) When an object rotates, its EGI also rotates. Fortunately, the EGI rotates in the same manner as the object. In other words, this rotation does not effect the relative EGI mass distribution over the sphere. This is analogous to the fact that the relative distribution of continents on the earth does not change as the earth rotates. If an observed EGI is identical to one part of the prototypical EGI, we can find which part of the object is observed at that time, and we can find the object's relative attitude. In other words, matching an observed EGI with a model EGI involves three degrees of freedom. There are two degrees of freedom corresponding to which point on the Gaussian sphere is perpendicular to the line of sight. The remaining degree of freedom comes from rotation about the line of sight.

We will use two constraints to reduce these degrees of freedom. Although a brute force technique, such as search through the space of possible attitudes [Ballard, 1981, Brou, 1983] can be used, we will reduce this search space using constraints before EGI comparison [Ikeuchi, 1983]. The EGI mass center position constrains the line of sight. Furthermore, the EGI inertia direction constrains the rotation around the line of sight.

3.2.3. EGI mass center

Elevation of the EGI mass center from the hemisphere base plane gives a constraint on the line of sight [Ikeuchi, 1983]. Even though the EGI mass center over the whole sphere is at the center of the sphere [Smith, 1979], the EGI distribution over a visible hemisphere always has some bias. Since this mass center is different for different visible hemispheres, correspondence of the EGI mass centers becomes a necessary condition for correspondence of the EGI distribution. Thus, comparing the observed EGI mass center with that of the prototypes reduces the freedom of the line of sight.

This elevation $A(v)$ at the direction, v is obtained as

$$A(v) = \frac{\int\int_{V.H.} Z_o(s,t) EGIM(s,t)\sqrt{EG - F^2}dsdt}{\int\int_{V.H.} EGIM(s,t)\sqrt{EG - F^2}dsdt},$$

where (s,t) is a parameterization over the Gaussian hemisphere, $EGIM(s,t)$ is EGI mass density there, V is the line of sight direction, and $V.H.$ stands for a visible hemisphere defined by v. $(X_o(s,t), Y_o(s,t), Z_o(s,t))$ denotes the coordinate value of the point (s,t) on the Gaussian hemisphere in the viewer-centered coordinate system. E, F, G are the coefficients in the first fundamental form of the Gaussian hemisphere of the parameterization (s,t) [Do Carmo, 1976, pg.268]. $\sqrt{EG - F^2}$ may be regarded as a Jacobian of the transformation from (s,t) to the hemisphere surface. Note that the line of sight, v is equivalent to the Z_o axis from the definition.

We will call this elevation, $A(v)$ the projection area ratio, because this value equals the ratio of projected area to surface area [Ikeuchi, 1983]. Comparing an observed ratio with the registered ratios prunes the search for the correct line of sight.

3.2.4. EGI inertia direction

The Gaussian hemisphere can also be rotated about the candidate line of sight. This degree of freedom is determined using the 2D EGI inertia axis [Ikeuchi, 1983]. This inertia axis is defined on the tangential plane (image plane) to the visible hemisphere at the north pole. The axis direction is going by:

$$I_{xx}(v) = \int\int_{V.H.} EGIM(s,t)X_o(s,t)X_o(s,t)\sqrt{EG - F^2}dsdt$$
$$I_{xy}(v) = \int\int_{V.H.} EGIM(s,t)X_o(s,t)Y_o(s,t)\sqrt{EG - F^2}dsdt$$
$$I_{yy}(v) = \int\int_{V.H.} EGIM(s,t)Y_o(s,t)Y_o(s,t)\sqrt{EG - F^2}dsdt$$

$I_{xx}(v), I_{xy}(v), I_{yy}(v)$ gives the principal inertia direction for the line of sight v,

$$\alpha(v) = (\frac{1}{2})\tan^{-1}\frac{2I_{xy}(v)}{I_{xx}(v) - I_{yy}(v)}$$

Thus, $\alpha(v)$ gives the direction of the minimum inertia axis on the image plane.

4. Implementation of the system

4.1. Implementation of the Photometric Stereo Method

4.1.1. Orthographic Projection: Image modeling

If the size of the object is small compared to the viewing distance, then the image-formation can be approximated by an orthographic projection. To standardize the image geometry, the viewer direction is aligned with the z-axis. The object point (x,y,z) maps into image point (u,v) where:

$$u = x$$
$$v = y.$$

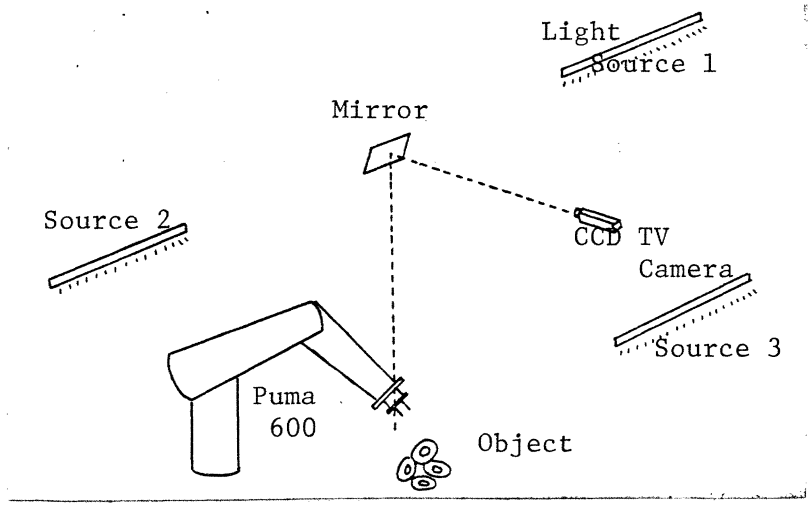


Figure 1. The experimental set up. A CCD camera is mounted in the ceiling and observes the scene through a mirror. Three light sources are controlled by the lispmachine.

4.1.2. Light Source Calibration

Fluorescent lamps are employed for lighting. This kind of extended source can increase the range of normals which the photometric stereo method can handle. With point sources, the normal directions illuminated by all three sources are limited.

These light sources are controlled by a Lisp Machine. Fig. 1 shows the experimental setup. A CCD camera is mounted in the ceiling and observes the scene through a mirror. Thus, the optical axis of the TV camera is horizontal between the TV camera and the mirror, and is vertical between the mirror and the objects.

The photometric stereo method assumes that the illumination due to a given light-source is constant over the points of the field of view. This assumption works if the field of view is narrow or the light source is distant. In our case, it does not work, due to the finite distance between the object and the light source.

This distance effect can be modeled as

$$E_i^{real}(x,y) = d_i(x,y)E_i^{standard}.$$

where i denotes light source i, $E_i^{real}(x,y)$ denotes brightness of the source i at the point (x,y), of the image plane. And $d_i(x,y)$ is the effect due to distance at the point (x,y). We assume that the distance between the point and the light source is much larger than the width of the field of view. Then, the formula can be expanded by Taylor series:

$$d_i^{real}(x,y) = d_i(x_o,y_o) + a_i(x-x_o) + b_i(y-y_o)$$

We can determine coefficients $d_i(x_o,y_o), a_i, b_i$, from the brightness distribution of a white board using a least squares fit.

Conversion factors, $\frac{1}{d_i(x,y)}$ from E_i^{real} to $E_i^{standard}$ are calculated at each point under each light source, and are stored in the computer. Fig. 2a shows the

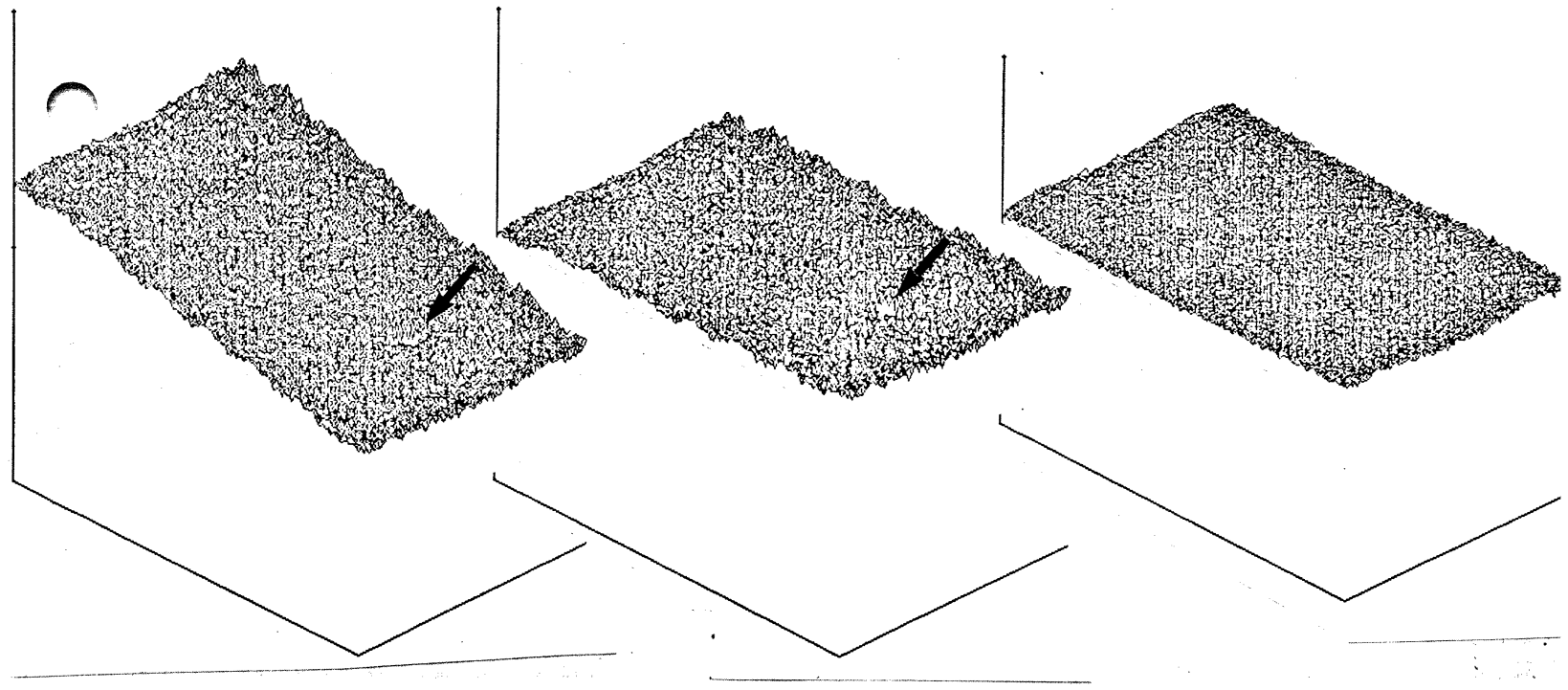


Figure 2. Calibration of brightness distribution.

original brightness distribution of a test board whose orientation is different from the calibration board used to make the coefficients. Fig. 2b shows the converted brightness distribution. Note that it is almost constant after the correction.

4.1.3. Albedo Effect Calibration

Sometimes the object surface has a slightly variable color. Also the TV camera has uneven sensitivity across its image plane. The collection of these effects may be called *the albedo effect.*

Let the real brightness values be $E_i^{real}(x,y)$ and the measured values be $E_i^{measure}(x,y)$.

$$E_i^{measure}(x,y) = \gamma(x,y)E_i^{real}(x,y),$$

where $\gamma(x,y)$ is an albedo function, and i ranges over the light sources.

We can cancel γ by normalizing the measured brightness values.

$$\begin{aligned} E_i(x,y) &= \frac{E_i^{measure}(x,y)}{(E_1^{measure}(x,y) + E_2^{measure}(x,y) + E_3^{measure}(x,y))} \\ &= \frac{E_i^{real}(x,y)}{(E_1^{real}(x,y) + E_2^{real}(x,y) + E_3^{real}(x,y))}, \end{aligned}$$

Fig. 2c shows the result of this operation. Note that the TV camera has a dark spot indicated by an arrow. This defect is canceled by this operation.

4.1.4. Constructing Lookup Table

The photometric stereo method is usually implemented using a lookup table which converts triples of image values into surface orientations. The lookup table is

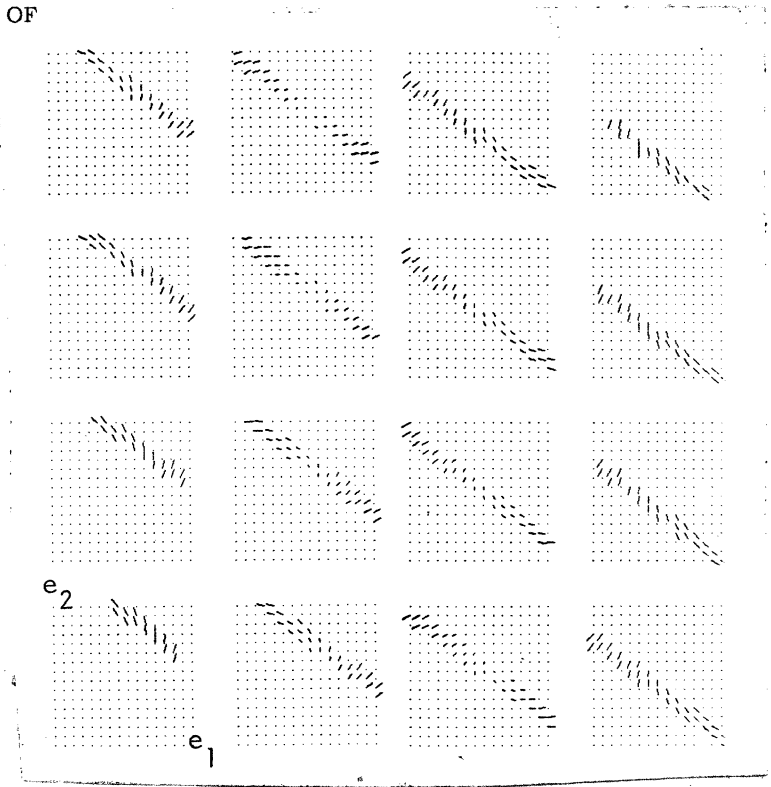


Figure 3. The lookup table for photometric stereo system.

indexed by the brightness values and contains the surface normal. Thus, the lookup table should be constructed either from a reflectance map given as an explicit formula [Ikeuchi, 1981b, Coleman, 1981] or experimentally [Silver, 1980].

We will follow the experimental method to construct the lookup table. An object of known shape, a sphere in our case, is imaged. At first, all three light sources are turned on to extract the contour lines. No self shadow lines are observed from the TV camera, when three light sources are on. Thus, we can easily extract the sphere's contour line from the dark background by thresholding the brightness array.

The grey-levels obtained at a particular point under each light source are used to locate an entry in the table. At that point the surface orientation can be computed using the known shape of the calibration object. This orientation value is entered in the lookup table. The resulting lookup table is a 3-D lookup table; each dimension corresponds to one of the three values. Fig. 3 shows the 3-D lookup table which converts a triple of brightness values into the surface orientation.

4.2. Segmentation

4.2.1. Smoothness Constraints and Occluding Boundary

A segmentation process is necessary for finding a target object. In order to find a target object from the jumbled background, it is necessary to cut the scene into pieces of connected regions. This segmentation process requires two operations: (1) making a binary image to indicate objects, (2) isolating the binary image into connecting regions corresponding to individual objects.

We use a needle map for this segmentation. Previously, this constructing-binary-image has been attempted using thresholding of brightness arrays. In this paper, however, a binary image is produced by the photometric stereo system. A point is one in a binary image if the photometric stereo can determine surface orientation; a point is zero if the system cannot determine it.

Previously, the isolation process also has been done by detecting brightness changes over the scene or thresholding brightness values. However, these edges are strongly disturbed by noise. Sometimes even "higher level knowledge" is employed to track edges [Shirai, 1978]

On the other hand, overlapping is a physical fact. There are many useful consequences which are caused by overlapping. The brightness change is only one among them. Two consequences of the physical situation are used in this paper. The first is that smoothly curved surfaces turn away from the viewer at occluding boundaries. Here, occluding boundaries are boundaries where an object overlaps another object [Ikeuchi and Horn, 1981]. Regions near occluding boundaries have surface normals nearly perpendicular to the viewer direction. The photometric stereo system cannot determine surface orientation if the surface is too steep. Thus, we cannot calculate orientation just inside occluding boundaries.

The second consequence is that the surface-smoothness assumption [Ikeuchi and Horn, 1981] is violated at the occluding boundary. It is necessary to use this consequence because in some cases, mutual illumination from one object to another makes an area near the edge appear as if it was less steep than it actually is. This bridging effect can be avoided using the smoothness constraints, provided that we can assume that the target object has a smooth surface. The smoothness constraints tells us that neighboring points should have similar surface orientations. It is unlikely that the orientation of the occluded object matches that of the occluding object. Thus, these regions violate the conditions. By using the first and second consequences we can reliably segment the binary image into isolated regions.

4.2.2. Labeling

Each connected region may be found by labeling [Winston and Horn, 1981]. A different number is assigned to every connected component of the image. The labeling operation can be performed by moving a 2 by 2 window across the image:

a
b

We scan along each row from left to right starting at the top.

If the smoothness measure is large (not smooth) enough
between A and B or between A and D,
A is 0, otherwise

If A is 0,
it remains 0, otherwise

If B and D are 0,
A is assigned a new label, otherwise

If D is 0 and B is labeled n,
so is A, otherwise

If B is 0 and D is labeled m,
so is A, otherwise

If B and D are labeled n,
so is A, otherwise

If B is labeled n and D is labeled m,
then label A as n,
and note that the m is equivalent to n.

After the first pass, some objects may have multiple labels. A second pass will solve this problem. First a lookup table for labels is constructed. The lookup table is then used to reassign each picture cell to the smallest label in an equivalence class.

4.3. Decision-Making

4.3.1. Area and Ruler Number:Feature

In order to pick a target region, we have to use heuristics particular to the object. We will use donuts for object. Thus, area and Euler number may be used as features for objects.

The easiest donut to pick up is the least tilted one on top of the pile. If the donut is nearly flat, we can observe a hole in the region. We cannot observe a hole for a highly tilted donut. If a donut is occluded by another donut, the donut may appear as a horseshoe shape, and again, the region has no hole. Thus, the region having one hole is more favorable than the region having no hole. The largest region is the most favorable region because the region most likely is at the top of the pile; if an object is occluded by other objects, the observable area of the region is reduced. Also, a less tilted donut has greater apparent area. Thus, the largest region whose Euler number is zero is the best region for grasping.

4.3.2. Area

Each region's area can be obtained by counting the number of picture cells in each region. This operation is done at the same time as the labeling operation. The five largest regions are selected as candidate regions.

4.3.3. Euler number

Each region is swelled in 3 picture cells to erase noise spots inside of the region. This process assumes that the hole of the donut is too large to be closed by this operation. Three-pixel swelling is done by applying eight-neighbor swelling, four-neighbor swelling, and then eight-neighbor swelling once more. Four-neighbor swelling propagates on-center into four-neighbor on-cells: two horizontal neighboring cells and two vertical neighboring cells. Eight-neighbor swelling propagates on-center into eight-neighbor on-cells: two horizontal neighboring cells, two vertical neighboring cells, and four diagonal neighboring cells. Thus, one on-cell is swelled into forty-five neighboring on-cells as shown in Fig. 4.

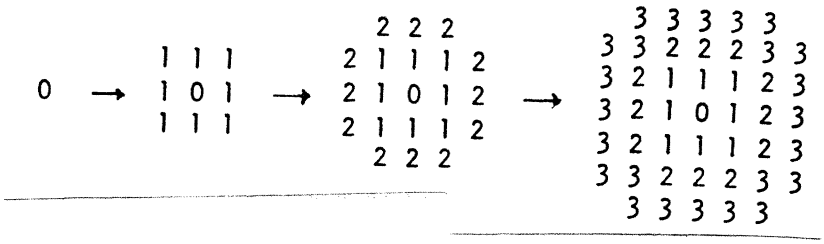


Figure 4. Swelling operation.

The Euler number is found by looking for distinctive patterns [Winston and Horn, 1981]:

```
0 0   0 1
0 1   1 1
```

The Euler number is obtained as the difference of the number of occurrences of the first pattern and the number of occurrences of the second pattern.

4.3.4. Decision Procedure

The target region is selected using the following procedures.

(step 1)

Five largest regions among the isolated regions are selected.

(step 2)

If **more than two regions** *have Euler number 0,*
then the largest region is selected.

If **only one region** *has Euler number 0,*
then that region is selected.

If **no region** *has Euler number 0,*
then the largest region is selected
and also notice the the region has no holes for the EGI matcher.

4.4. Implementation of EGI

4.4.1. Tessellation of the Gaussian sphere

In order to represent and manipulate an EGI on a digital computer, one must tessellate the Gaussian sphere uniformly. A continuous surface such as an elliptic surface is mapped to a continuous EGI mass distribution. A tessellated sphere is needed to represent this image in a computer. The tessellation method must provide a uniform division of the Gaussian sphere. Since we cannot predict the line of sight, a tessellation method should have the same angular resolution in every direction. Thus, each cell on the tessellated sphere is required to have the same area and the same distance from its adjacent cells.

The projection of a regular polyhedron onto a sphere has this property [Wenninger, 1979]. A regular polyhedron has faces of equal area, which are evenly

distributed in direction with respect to the center of gravity. Projecting edges of a polyhedron onto the circumscribed sphere with respect to the sphere center tessellates the sphere uniformly.

Since the highest order regular polyhedron, the icosahedron only has twenty faces, we actually have to use a geodesic dome [Ikeuchi, 1981a]. A geodesic dome is obtained by division of each triangle of the tessellated sphere into smaller triangles. We use a geodesic dome made from a two frequency dodecahedron, because this geodesic dome has a more uniform facet area distribution than other domes of the same tessellation order [Wenninger, 1979].

The tessellated dome is used for two purposes. One is to accumulate an EGI image. A particular object surface patch corresponds to a cell with a given surface orientation on the dome. Measured surface area will be added to the corresponding cell. The cumulative image on the dome is the distributed version of the object's EGI. The other purpose is to sample the possible line of sight. Since the cells are distributed uniformly over the dome, the center position of each cell can define the spatial direction. Therefore, the line of sight space is sampled uniformly by this dome.

4.4.2. Normalized EGI: Non-Convex Object

A non-convex object may have hidden surfaces with orientations in the visible hemisphere. This occlusion problem requires us to redefine the EGI for each line of sight. This can be done using either a geometrical modeler [Hosaka et al., 1974] or a mathematical expression of the object.

The EGI of a non-convex object can be expressed using four parameters. The line of sight can be expressed using two parameters (s_l, t_l). A EGI mass distribution over the visible hemisphere at a line of sight is expressed using another two parameters (s, t). Namely,

$$EGI = EGIM(s, t, s_l, t_l),$$

Note that (s_l, t_l) is similar to the light source direction and (s, t) is similar to the surface orientation of the reflectance map [Horn, 1977].

We can store this four dimensional EGI distribution in a two dimensional table [Ikeuchi, 1983]. Since tessellation cells on the dome can be ordered along a one dimensional row, an EGI mass distribution for the line of sight (s_l, t_l), can be represented as a one dimensional vector. The possible lines of sight are also ordered in one dimension. Therefore, an EGI can be stored in a two dimensional table, with each row corresponding one possible line of sight. Each element contains an EGI mass (surface area) corresponding to the surface orientation for that line of sight.

Storing constraint information adds two additional columns to each row. The first column keeps the projection ratio. The second column stores the original inertia direction relative to X, Y axis of the image plane. The EGI mass distribution over the remaining elements is rotated so as for the EGI inertia direction to agree with the X axis of the image plane. We will refer to this recalculated EGI as the normalized EGI (NEGI). Comparing a NEGI from an observed needle map with

NEGI's on the table can be done without reference to rotation around the line of sight.

4.4.3. Matching Function

A matching function determines a similarity measure between an observed NEGI and the NEGIs on the table [Ikeuchi, 1983]. The matching function checks whether each column has a similar amount of EGI mass to the corresponding table column. More precisely, the following operation is done at each column. A cumulative sum represents the similarity of row v. Note that this operation is done only when row v's projection area ratio is similar to the observed ratio.

if $EGIM^{observe}(i) = 0.0$
do nothing

if $EGIM^{observe}(i) \neq 0.0$
$\Delta = \frac{|EGIM^{observe}(i) - EGIM^{model}(v, i+e)|}{EGIM^{observe}(i)}$
$d = distance(i, i+e)$
if $d < d^{standard}$, *and* $\Delta < \Delta^{standard}$
add $EGIM^{observe}(i) \cdot (1 - \Delta) \cdot \cos d$ *to total point, S(v).*

$EGIM^{observe}(i)$ is observed NEGI mass at i cell, $EGIM^{model}(v, i+e)$ is EGI mass at $(v, i+e)$ cell of the table, and $distance(i, i+e)$ is the inner product between the cell directions. Δ is the relative error of $EGIM^{model}(v, i+e)$ assumed to correspond to $EGIM^{observe}(i)$. Thus, the first term represents how important the $EGIM^{model}(i)$ is. The second term represents how different the two mass is. The third term represents how far the two cell are. If $EGIM^{model}(v, i)$ has the exactly same mass as $EGIM^{observe}(i)$, $(1 - \Delta) = 1$ and $d = 1$, then, $EGIM^{observe}(i)$ is added to the total. If this correspondence is established at each column i, the total becomes 1, because total EGI mass is 1, one is the highest score of the matching function.

The direction having the highest score is the estimate of the observed line of sight. The NEGI table also registers how many degrees the prototype is rotated to align the least inertia axis with the X axis. We already know how many degrees the observed image is rotated to bring the least EGI-inertia axis into coincidence with the X axis. These two angles give the rotation angle of the observed image about the line of sight.

4.5. Eye and Hand Interface

4.5.1. Determining the Gripper Configuration

We employ a parallel jaw gripper for grasping. The grasping position of the gripper is the highest point inside the target region, because if we grasp an object at the highest point, it is less likely that the hand will hit other objects. The approach direction of the gripper is the direction of the donut's axis. The orientation direction of the gripper about the approach vector is aligned with the steepest descent direction of the plane of the donut. See fig. 5.

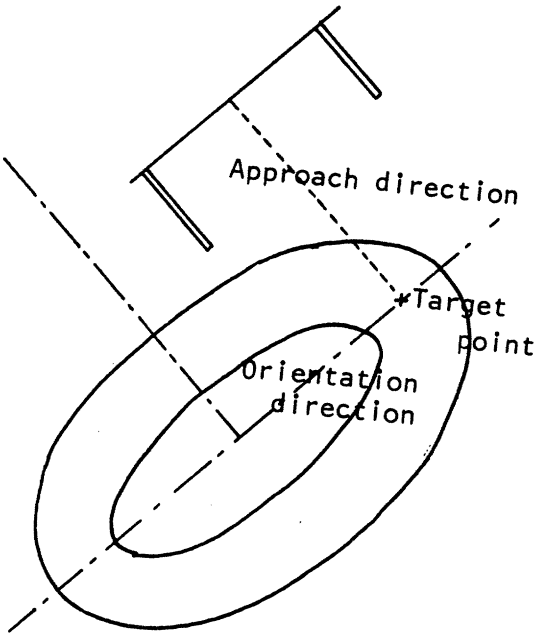


Figure 5. The gripper configuration. The approach direction of the gripper is the direction of the donut's axis. The orientation direction of the gripper about the approach vector is aligned with the steepest descent direction of the plane of the donut.

Since we know the attitude of the object, we can determine the direction of steepest of decent of the object. The mass center of the binary image can be determined on the target region. We can determine a line along the steepest decent direction from the mass center. The highest point should lie on this line. The (x, y) position of the grasping point is the mid point of the line segment cut by the two boundaries of the donut.

4.5.2. Camera Model

The relationship between the camera coordinates and the Puma's world coordinates is determined empirically. Although it would be possible to solve the equations from the camera coordinates to Puma's world coordinates analytically using careful measurement of camera position and attitude [Paul, 1981], we did not follow this method. Assume that (u, v) is the camera coordinate system and (x, y, z) is the Puma's world coordinate system. Then, (x, y) can be approximated using the affine transformation as

$$x = au + bv + e$$
$$y = cu + dv + f.$$

These coefficients can be determined using a least squares fitting. The board shown in Fig. 6 is displayed under the TV camera at 9 points in the field of view. At each point both the TV coordinate of the cross section and the Puma's world coordinate are registered. Least squares gives the matrix from these measurement.

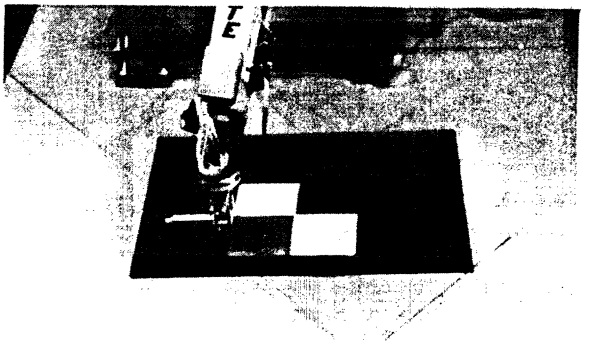


Figure 6. Camera model calibration. At nine points both the TV coordinate of the cross section and the Puma's world coordinate are registered. Then, least squares gives the coefficients from these measurement.

This matrix is a function of the Z axis of world coordinate. The matrix is determined at two levels ($Zaxis$). A linear interpolation method gives the desired matrix at any other level.

5. Experiment

5.1. Control Structure

The vision part consists of two stages. The first part segments the image into regions and determines the target region. The second part determines the position and the attitude of the object. The (x, y) position of the grasping point, the approach direction, and the orientation direction are computed.

The segmentation part consists of four subparts: image acquisition module, photometric stereo 1 module, labeling module, and decision module. The labeling module isolates the image into connected regions using the needle map provided by photometric stereo 1. The decision module determines the most favorable region among the isolated regions.

The attitude calculation part consist of four subparts: photometric stereo-2 module, EGI-generating module, EGI-matching module, and planning module. The photometric stereo-2 module recalculates surface orientation using a more precise lookup table. EGI-generating module produces an EGI image from the output of the photometric stereo-2 module. The EGI-matching module determines the attitude of the object by comparing the generated EGI with the EGI lookup table. The planning Module determines the grasping point from the binary image of the target region and the attitude from the result of EGI matching.

The manipulator motion controller picks up an object by two stages. The first step determines the absolute depth of the grasping point. The second step aligns the gripper to the object's attitude and moves to the grasping point. Then, the gripper grasps the object and moves to the destination point.

Figure 7. Image acquisition. Each picture shows a scene when one of the three light source is turned on.

Figure 8. Obtained brightness value.

5.2. Eye Part

5.2.1. Image Acquisition

The three light sources are controlled by the Lisp Machine. A Hitachi solid state CCD camera is connected directly into the Lisp Machine, and the input data is written directly to the main memory of the Lisp Machine. Thus, a lisp program can access the data using ordinary array access functions. Fig. 7 shows pictures when one of the three light sources is turned on.

The three brightness arrays are shown in Fig. 8. Note that a simple edge-based method would not work well in this case.

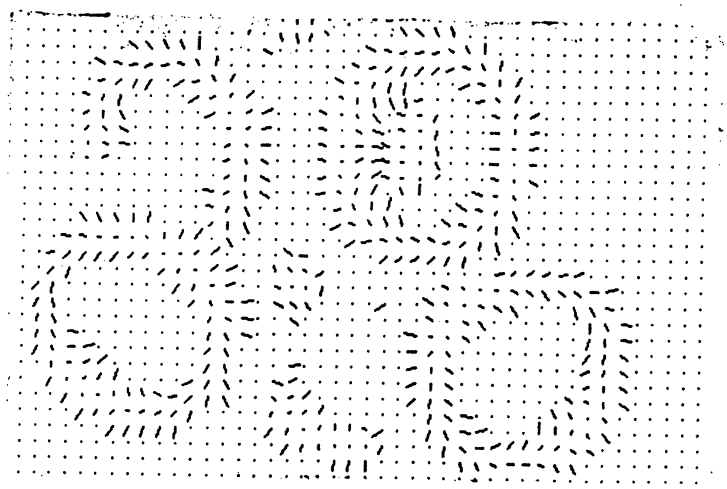


Figure 9. The needle diagram obtained by the photometric stereo1

5.2.2. Photometric Stereo-1 Module

This experiment uses two modules of the photometric stereo system. Photometric stereo-1 module is used to produce input for the segmentation. The segmentation is done using a needle map made by the photometric stereo-1 module, while a needle map by Photometric stereo-2 module is used to the input for the EGI-matching.

We can adjust the characteristics of the photometric stereo system by varying the accuracy of the lookup table. We can vary detectable areas in photometric stereo by pruning the needles in the lookup table. We can speed up the system by using a coarse table. Since the purpose of the Photometric Stereo-1 module is to separate each region using the needle map, the Photometric stereo-1 module uses a coarse table which contains surface orientations whose zenith angle are less than 40 degrees.

If we reduce this zenith angle, we can separate each region more easily. However, this may cause an object to be divided into more than region. It is difficult to reconnect two regions into one region which corresponds to an object. On the other hand, the more steeply an object tilts, the less likely the puma can pickup the object due to joint limit 5. Thus, 40 degree is a trade off angle. Fig.9 shows the needle map obtained by photometric stereo-1 module.

5.2.3. Segmenting Module

A segmentation program based on the surface smoothness is applied. The program labels the image with a different number for each isolated region. Fig.10 shows the isolated regions.

5.3. Decision-Making Module

The target region is selected based on the Euler number and the area of each region. Fig.11 shows the target region selected by a decision process.

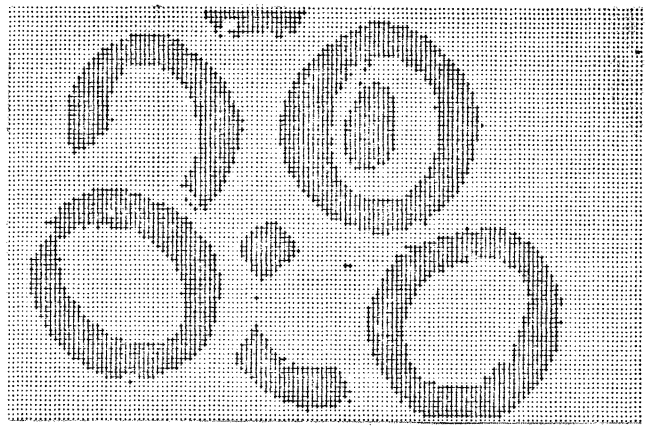


Figure 10. The segmented regions.

5.3.1. Photometric Stereo-2 Module

Photometric Stereo-2 module determines surface orientation up to 60 degrees of zenith angle at the target region. In order to expand the reduced area by the segmenting module, the target region is swelled. Over the expanded region, photometric stereo-2 module determines the surface orientation by using the precise lookup table. This module also computes EGI mass by look up in a table which converts triples of brightness values into EGI mass. Fig.12 shows the needle map generated by Photometric stereo-2 module.

5.3.2. EGI-Generating Module

The projection ratio, the least inertia axis, and the NEGI are calculated from the surface normals at the target region. Fig. 13 shows the obtained NEGI on the visible Gaussian hemisphere observed from the north pole. Each needle represents the EGI mass there. Since Fig. 13 is the NEGI rather than the EGI, the Gaussian hemisphere is rotated around the Z axis (perpendicular to the paper surface) so that the least EGI inertia axis agrees with the x axis.

5.3.3. EGI-Matching Module

Since a donut has an axis of rotational symmetry, the necessary sampling directions are points along a 90 degree section of a great circle on the Gaussian sphere containing the symmetry axis of a donut. At 10 degree increments along the great circle, the NEGI, projection ratio, and the least inertia axis direction are calculated using a mathematical model of the donut shape. Since directions near the axis may have relatively large error in the least inertia axis direction, NEGI's rotated by a small amount around the exact alignment are also registered in the lookup table.

5.3.4. Planning Module

The highest point among the target region is selected as the grasping point. Fig.14 shows the output of the algorithm which indicates both the (x, y) coordinate of the grasping point and the gripper approach direction.

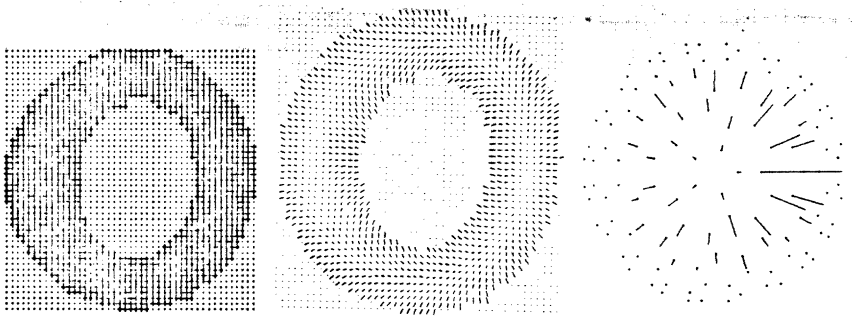


Fig. 11
The Target region

Fig. 12
Recalculated needle map

Fig. 13
Obtained NEGI

Figure 13. Obtained NEGI

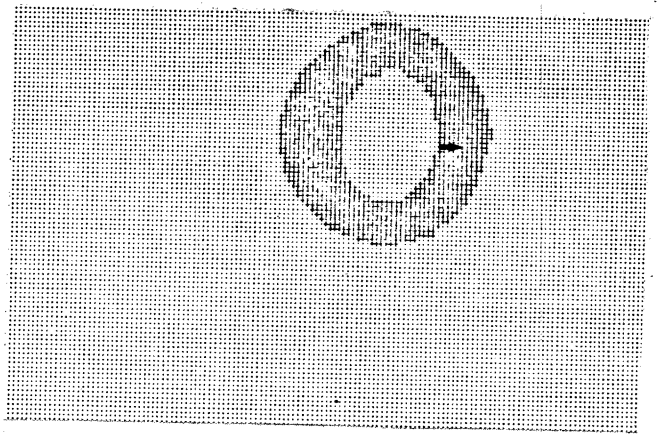


Figure 14. The output of the vision part. The algorithm indicates both the (x, y) coordinate of the grasping point and the gripper approach direction.

5.4. Hand Part

5.4.1. Determining Absolute Depth

The photometric stereo can only determine relative depth. Thus, the hand system has to determine the absolute depth. Since the grasping points exists somewhere along the line of sight corresponding to the TV x-y coordinate, the gripper approaches the point along the line of sight in order to determine the absolute depth. See Fig. 15(a).

The gripper is a parallel jaw gripper developed at MIT. A pair of LED sensors is mounted and can detect an object between the finger. If the gripper senses the object between the fingers, the position gives the actual target point. See Fig. 15(b).

5.4.2. Grasping an Object

The gripper moves to a point on the approach line defined by the target point and the approach direction. We will call this point as a entry point. Then, it begins to approach the donut along the approach line until the LED sensor senses an object between the fingers as shown in Fig.15(c).

After the LED sensor finds an object between the fingers, the gripper moves 1 cm more along the direction for a firm grasp. Then, the gripper grasps the object by closing the fingers as shown in Fig.15(d)

5.4.3. Bringing an Object

The gripper retracts along the approach line until it reaches the entry point as shown in Fig. 15(e) The gripper changes its attitude at that point as in Fig. 15(f). If the gripper fails to grasp a donut, the system orders the eye module to repeat the analysis after moving the gripper to the original point. If the system succeeds in grasping the donut, the gripper brings the donut to the pre-determined point as shown in Fig.15(g),15(h),15(i).

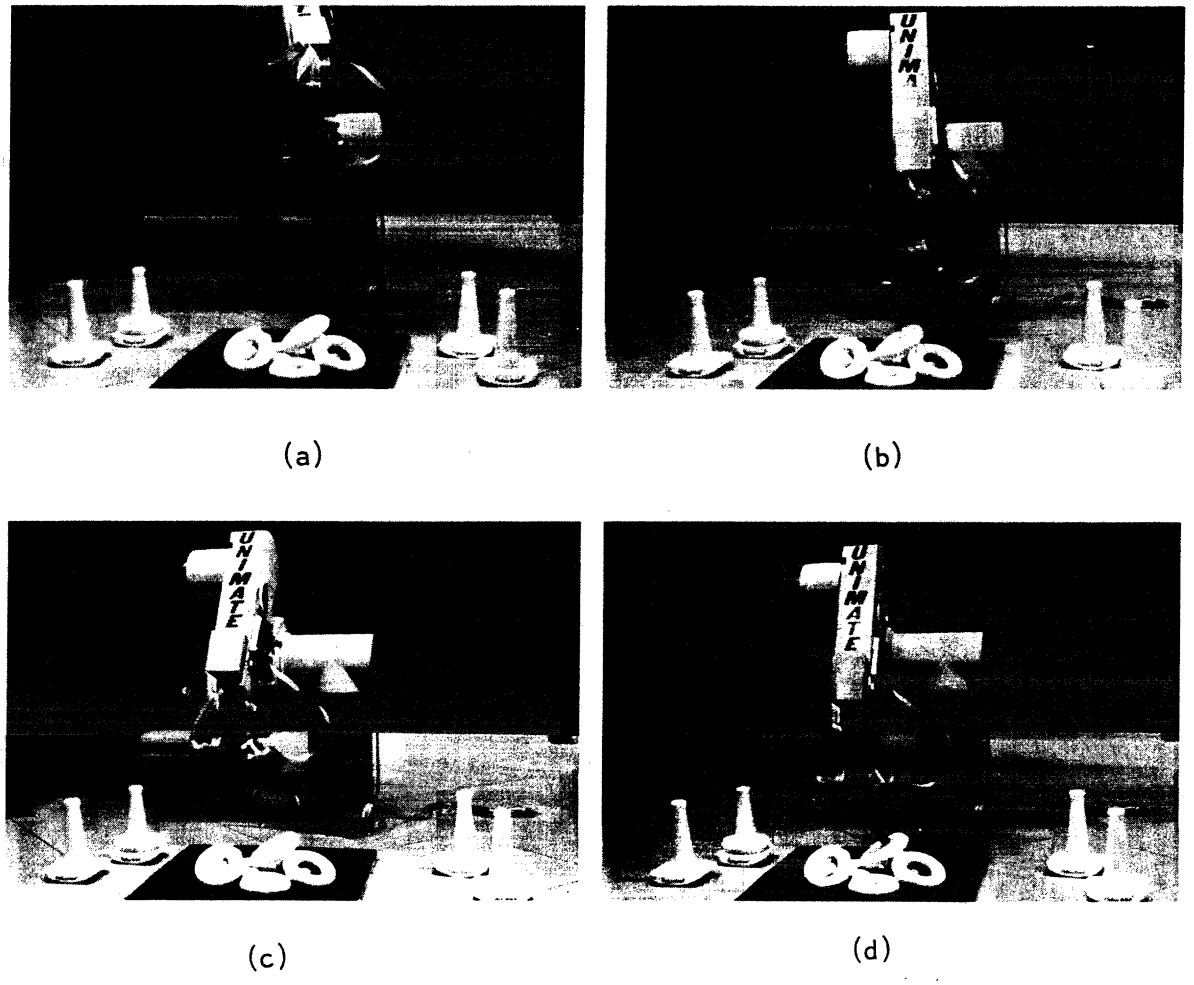
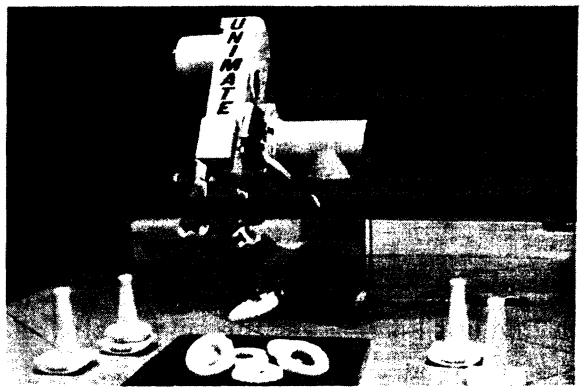
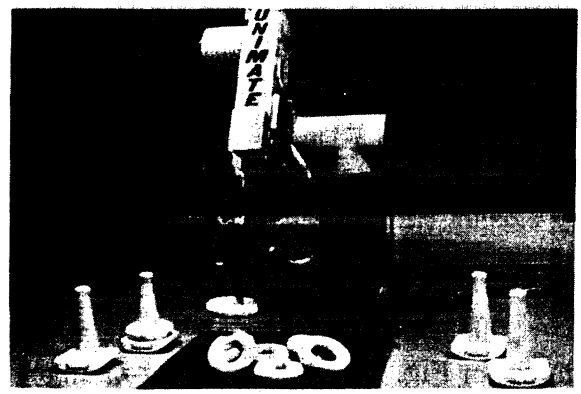


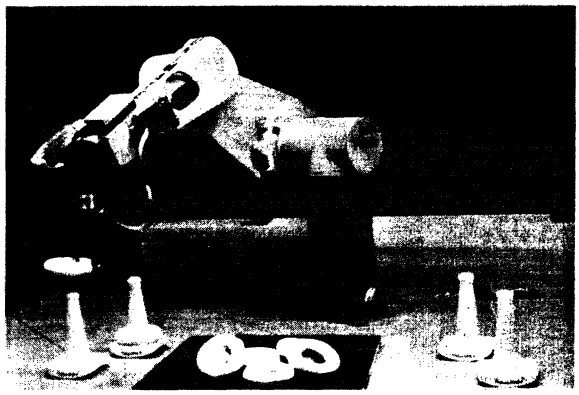
Figure 15. Picking up an object.



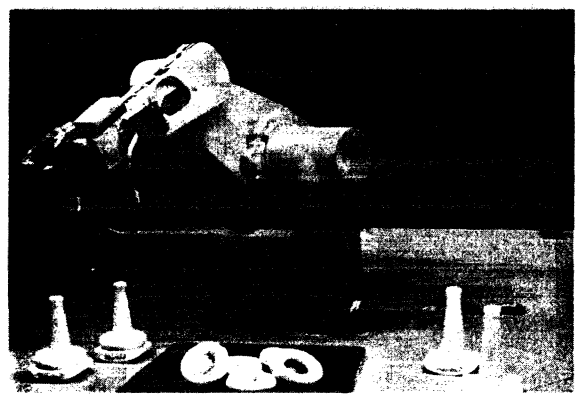
(e)



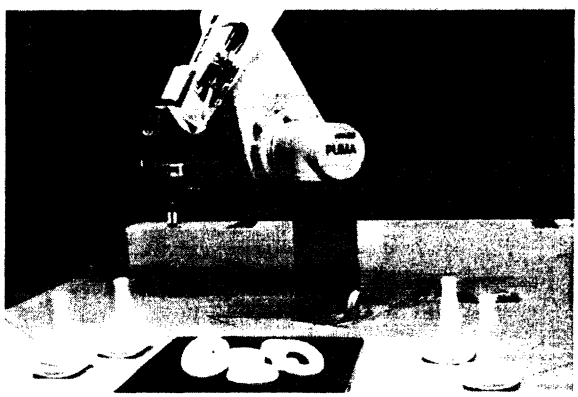
(f)



(g)



(h)



(i)

Figure 15. Picking up an object.
(continued)

6. Summary

A hand-eye system is presented which can perform the bin-picking task. Two basic tools are used: photometric stereo and extended Gaussian images. The vision subsystem consists of two parts: segmentation based on the needle map and attitude-determination using the extended Gaussian image. The system can pick up an object out of a pile of objects successfully.

The execution time of the vision part is roughly 40 - 50 seconds on a lisp-machine. This time includes frame-grabbing time. The entire system is coded using zeta-lisp (lispmachine lisp). It is compiled into "machine code" using the zeta-lisp compiler.

Vision is a very tough problem. Thus, we have to establish a well-defined domain. In the early days, the blocks world was attacked, because the world only consisted of straight lines. This paper explores the so-called bin-picking domain. We believe that the bin-picking domain is a substantial extension of the blocks world. This world involves curved objects whose surface shape may be much more complicated than an object in the blocks world. However, we have prototypes of the object which the system may encounter. We also know the surface material of the objects. We can control the environment where the objects are observed. This is not the case in a natural scene. Before starting to worry about a "general purpose vision machine" which works in the natural world, we should establish a small but well-defined subset domain and explore the solution in this domain.

It may appear that this paper does not address the problem of object recognition. However, in order to define the term recognition, we have to define the domain of recognition. In other words, recognition means to select the most likely one among the candidates based on the observed data. Since the 3D object varies its shape depending on the viewer direction, to determine the attitude of an object is the most primitive and well-define domain for the recognition. It is also not difficult to extend this system to multiple objects. [Ikeuchi, 1980].

Previously, segmentation has been attacked using edge detection, because edge detection is believed to be the starting point for vision [Marr, 1982]. In some cases, however, higher levels of representation help to segment the scene into regions. In our case, we used needle maps for this purpose. The depth map obtained from a range-finder is also a promising tool for segmentation [Oshima and Shirai, 1981, Sugihara, 1979]. As these higher level representations are becoming more easily obtainable, techniques which treat these representations should be explored extensively.

7. Acknowledgment

The following people also joined the project in part: Noble Larson constructed the CCD-TV camera interface. Patrick Sobalvarro provided the interface between VAL and zeta-lisp. John Purbrick built the fingers of the gripper. John Scott helps

to make the LED sensor. Without their effort, this project could not have been completed.

Draft of this article were proofread by R. Brooks and B. Schunk. Their comments are appreciated.

Finally, *domo arigato* Mr. PUMA for excellent performance, and Ikko for preparing this manuscript and some of the drawings.

8. References

Baird, M.L. 1977. Image Segmentation Technique for Locating Automotive Parts on Belt Conveyers, *Proc. 5th-IJCAI*, Cambridge, pp. 694-695.

Birk, J.R., Kelley, R.B., and Martins, H.A.S. 1981. An Orienting Robot for Feeding Workpieces Stored in Bins, *IEEE trans. SMC*, Vol SMC-11, No. 2, pp. 151-160.

Brooks, M.J. 1982. Shape from Shading Discretely, *Ph.D. Thesis*, Essex University, CS Dept.

Brooks, R. 1981. Symbolic Reasoning among 3-D Models and 2-D Images, *Computer Vision*, ed. M. Brady, Amsterdam, North-Holland, August, pp. 285-349.

Bajcsy, R. 1980. Three-dimensional Scene Analysis, *Proc. 5th-ICPR*, Miami Beach, pp. 1064-1074.

Ballard, D.H. and Sabbah, D. 1981. On Shapes, *Proc. 7th-IJCAI*, Vancouver, pp. 607-612.

Ballard, D.H. and Brown, C.M. 1982. *Computer Vision*, New Jersey, Prentice-Hall,

Barrow, H.G. and Tenenbaum, J.M. 1978. Recovering Intrinsic Scene Characteristics from Images, ed. A.R. Hanson, A.R. and E.M. Riseman, *Computer Vision Systems,* New York, Academic Press.

Binford, T.O. 1971. Visual Perception by Computer, *IEEE Conf. on Systems and Control* , Miami.

Bolles, R. 1977. Verification Vision for Programmable Assembly, *5th-IJCAI*, Cambridge, pp. 569-575.

Brady, M. 1982. Computational Approaches to Image Understanding, *ACM Computing Surveys*, Vol 14, pp. 3-71.

Coleman, E.N. and Jain, R. 1981. Shape from Shading for Surfaces with Texture and Specularity, *Proc. 7th-IJCAI*, Vancouver, pp. 652-657.

Davis, L.S. 1979. Shape Matching Using Relaxation Techniques, *IEEE Trans. on PAMI*, Vol. PAMI-1, No. 1, pp. 60-72.

Do Carmo, 1976. Differential Geometry of Curves and Surfaces, Englewoods Cliffs, Prentice-Hall.

Grimson, W.E.L. 1981. *From Images to Surface: A Computational Study of the Human Early Visual System*, Cambridge, MIT Press.

Horn, B.K.P. 1975. Obtaining Shape from Shading Information, *The Psychology of Computer Vision*, ed. P.H. Winston, New York, McGraw-Hill, pp. 115-155.

Horn, B.K.P. 1977. Image Intensity Understanding, *Artificial Intelligence*, Vol. 8, No. 2, pp. 201-231.

Horn, B.K.P. 1979. SEQUINS and QUILLS – representation for surface tomography. *AI Memo No. 536*, Cambridge, MIT, AI Lab.

Horn, B.K.P. and Sjoberg, R.W. 1979. Calculating the Reflectance Map, *Applied Optics,* Vol. 18, pp. 1770-1779.

Hosaka,M., Kimura, F., and Kakishita, N. 1974. A Unified Method for Processing Polyhedra, *Proc. IFIP-74*, pp. 768-772.

Ikeuchi, K. 1980. Shape from Regular-patterns An Example of Constraint Propagation in Vision. in *Proc. 5th-ICPR*, Miami Beach, pp. 1032-1039.

Ikeuchi, K 1981a. Recognition of 3D object using Extended Gaussian Image, *Proc. 7th-IJCAI*, Vancouver, pp. 595-600.

Ikeuchi, K. 1981b. Determining Surface Orientations of Specular Surfaces by Using the Photometric Stereo Method, *IEEE Trans. on PAMI* , Vol. PAMI-2, No. 6, pp. 66l-669.

Ikeuchi, K. 1983. Determining Attitude of Object from Needle map using Extended Gaussian Image, *AI Memo No. 714*, Cambridge, MIT, AI Lab.

Ikeuchi, K and Horn, B.K.P. 1981. Numerical Shape from Shading and Occluding Boundaries. *Computer Vision* , ed. M. Brady, Amsterdam, North-Holland, pp. 141-184.

Kanade, T. 1981. Recovery of the Three-Dimensional Shape of an Object from a Single View, *Computer Vision* , ed. M. Brady, Amsterdam, North-Holland, pp. 409-460.

Kelley, R.B., Birk, J.R., Martines, H.A.S., and Tella, R. 1982. A Robot System Which Acquires Cylindrical Workpieces from Bins, *IEEE Trans. on SMC*, Vol. SMC-12, pp. 204-213.

Kender, J.R. 1980. Shape from Texture, *Ph.D. Thesis* , Pittsburg, CMU, Computer Science Dept.

Kender, J.R., and Kanade, T. 1980. Mapping Image Properties into Shape Constraints, *Proc. AAAI-80*, Stanford, pp. 4-6.

Koshikawa, K. 1979. A Polarimetric Approach to Shape Understanding of Grossy Object, *Proc. 6th-IJCAI*, Tokyo, pp. 493-495.

Marr, D. 1982. *Vision* San Francisco, W.H. Freeman.

Nishihara, K.H. 1981. Intensity, Visible-Surface, and Volumetric Representation, *Computer Vision*, ed. M. Brady, (ed.), Amsterdam, North-Holland, pp. 265-284.

Ohta, Y., Maenobu, K., and Sakai, T. 1981. Obtaining Surface Orientation from Texels under Perspective Projection, *Proc 7th-IJCAI,* Vancouver, pp. 746-751.

Oshima, M. and Shirai, Y. 1981. Object Recognition Using Three-dimensional Information, *Proc 7th-IJCAI* Vancouver, pp. 601-606.

Paul, R.P. 1981. *Robot Manipulators,* Cambridge, MIT Press.

Perkins, W.A. 1977. Model-based Vision System for Scene Containing Multiple Parts, *Proc. 5th-IJCAI,* Cambridge, pp. 678-684.

Price, K. and Reddy, R. 1979. Matching Segments of Image, *IEEE Trans. on PAMI,* Vol. PAMI-1, No. 1, pp. 110-116.

Smith, D.A. 1979. Using Enhanced Spherical Images for Object Representation, *AI Memo No. 530,* Cambridge, MIT, AI Lab.

Sakane, S. 1981. An Algorithm for Scale Invariant Segment-Matching, *Proc. PRIP-81,* Dallas, pp. 565-571.

Shirai, Y. 1978. Recognition of man-made objects using edge cues, *Computer Vision System,* ed. A. Hanson and E. Riseman, New York, Academic Press.

Silver, W. A. 1980. Determining Shape and Reflectance using Multiple Images,*MS Thesis,* Cambridge, MIT, EECS.

Sugihara, K. 1979. Range-data analysis guided by a junction dictionary, *Artificial Intelligence,* Vol. 12, 1979, PP. 41-69.

Terzopoulos, D., Multi-level Reconstruction of Visual Surfaces, *AI Memo. 671,* Cambridge, MIT, AI Lab.

Tsuji, S, and Matsumoto, F. 1977. Detection of Elliptic and Linear Edges by Searching two Parameter Space, *Proc. 5th-IJCAI,* Cambridge, pp. 700-705.

Tsuji, S. and Nakamura, A, 1975. Recognition of an Object in a Stack of Industrial Parts, *Proc. 4th-IJCAI,* Georgia, USSR, pp. 811-818.

Woodham, R.J. 1978. Reflectance Map Techniques for Analyzing Surface Defects in Metal Casting, *AI-TR-457,* Cambridge, MIT, AI Lab.

Wenninger, M.J. 1979. *Spherical Models* , Cambridge, Cambridge University Press.

Winston, P.H. and Horn, B.K.P. 1981. *LISP,* Reading, Addison-Wesley.

Witkin, A.P. 1981. Recovering Surface Shape and Orientation from texture, *Computer Vision,* ed. M. Brady, Amsterdam, North-Holland, pp. 17-47.

Yachida, M. and Tsuji, S. 1975. A Machine Vision for Complex Industrial Parts with Learning Capability, *Proc. 4th-IJCAI,* Georgia, USSR, pp. 819-826.