NOTICE WARNING CONCERNING COPYRIGHT RESTRICTIONS:
The copyright law of the United States (title 17, U.S. Code) governs the making of photocopies or other reproductions of copyrighted material. Any copying of this document without permission of its author may be prohibited by law.

CMU-CS-86-144

Generating an Interpretation Tree from a CAD Model to Represent Object Configurations For Bin-Picking Tasks

Katsushi Ikeuchi
Department of Computer Science
Carnegie-Mellon University
Pittsburgh, PA 15213

Abstract

This paper describes a model based vision system for bin picking tasks. The system contains three components: a dual photometric stereo systems, a CAD model SOLVER, and an interpretation tree. The system has two modes: a compile mode, and a run mode.

In the compile mode, SOLVER is used to generate apparent shapes of an object under various viewer directions. Representative attitudes are extracted from the shapes. Then, an interpretation tree is generated to classify an observed attitude into one of the representative attitudes, and to determine the attitude precisely by matching features from SOLVER with observed ones.

In the run mode, intensity maps, needle maps, and a depth map are obtained by the dual photometric stereo. Then, the interpretation tree determines maps, areas, and features to be examined for attitude determination. The final representation is expressed in the world of SOLVER.

This research was sponsored by the Defense Advanced Research Projects Agency, DOD, through ARPA Order No. 4976, and monitored by the Air Force Avionics Laboratory under contract F33615-84-K-1520. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or of the U.S. Government.

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, information can be obtained without disturbing the environment. Also, vision can acquire global information about a scene; this is not the case for a tactile sensor.

There are basically three tasks where the vision feedback can play an essential role:

1. finding the target object and determining the grasping points,
2. bringing the object from its initial point to a destination point while avoiding collision with other objects, and
3. assembling something using the object.

This report explores a method for visual guidance of a manipulator in the first task domain: finding an object. A manipulator without vision can only pick up an object whose position and attitude are pre-determined. Such a system needs the help of another machine or a human for feeding objects at a pre-determined place in a pre-determined attitude. Since this feeding job is tedious, it is quite unsuitable for a human being. Traditional mechanical feeding methods rely on a known part geometry to orient the part by forcing it through a sequence of gates rails and stops. Besides being inflexible and capable of dealing with a very limited number of part types, these methods, including vibration, may cause defects in the objects due to collision.

Historically, bin-picking tasks have been attacked by detecting brightness changes [Tsuji and Nakamura 75, Baird 77, Perkins 77, Bolles and Cain 82]. 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-

UNIVERSITY LIBRARIES
CARNEGIE-MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15213

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 typical in bin-picking.

We [Ikeuchi etal. 84, 86, Horn and Ikeuchi, 84] have presented techniques for using photometric stereo and an extended gaussian image to determine object attitude. The photometric stereo determines surface orientations from the images under three different illumination conditions. A brightness triple at each point determines the surface orientation there. 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 are used to segment the visual scene into isolated regions corresponding to different objects. The distribution of surface orientations---an orientation histogram---measured over one of these isolated regions is used to identify the shape from a catalogue of known shapes. The object's attitude in space is also obtained as a by-product of the matching process. This system can pick up such a simple object as a doughnut successfully. This method, however, has three problems:

1. It is often difficult to express a complicated object such as a machine part with a mathematical function from which the extended Gaussian image is derived.

2. The extended Gaussian image is sometimes not powerful enough to determine the attitude of a machine part due to self occlusion, narrowness of observable areas, or scatter of observable regions of the object due to self shadows.

3. The previous system lacks a general representation of the outer world from which a planner can easily make a grasp plan. The purpose of robot vision is to provide the outer world information to task-acheiving parts. The representation can serve as the starting point to the task-acheiving module. Thus, the representation should be somehow a copy of the outer world and be in a convenient form to operate with it.

This paper resolve these problems using a CAD model. This system has following characteristics:

1. The system uses a depth map, a needle map, and edge maps.

2. An interpretation tree controls the process of determining attitude by using the most appropriate features derived from these maps at each determining process.

3. The interpretation tree classifies a target region into a representative attitude, and then determines the attitude more precisely over the attitude range of the representative attitude.

4. The attitude and the position obtained is represented in the world in SOLVER.

DERIVING THE INTERPRETATION TREE (Compile Mode)

SOLVER

SOLVER is a Solid modeler developed by Koshikawa [Koshikawa 84] for object recognition at the Electrotechnical Laboratory (ETL) in Tsukuba. The origin of SOLVER comes from a solid modeler GEOMAP [Kimura and Hosaka 77]. Since GEOMAP is a modeler not for object recognition but for object display, GEOMAP lacks some of the essential recognition facilities. Thus, Koshikawa revised GEOMAP into SOLVER for object recognition purposes.

Both SOLVER and GEOMAP share common data structures. Since SOLVER comes from GEOMAP, the basic part of SOLVER is the same as that of GEOMAP. For example, both SOLVER and GEOMAP represent an object as links of vertices, edges and faces. The relationships between them are expressed using the winged-edge representations [Baumgart 72]. While this winged-edge representation expresses the topological relationship between faces, edges, and vertices, a vertex cell contains the positional information. Transformation of an object can be realized by multiplying a transformation matrix by the positional information of the vertex cells. Complex objects can be synthesized from simple primitive objects using unification or subtraction.

SOLVER has the following three new features for object recognition.

1. **Exposure information** An exposure means one visible part of one face under one particular viewer direction. GEOMAP has a property called a region to express a face projected under one viewer direction. Each region is enclosed with some visible arcs or some invisible arcs occluded by some other

faces. GEOMAP, however, has no direct information of visible parts of a region. SOLVER has the following information of visible parts as exposure information.

a. area size of an exposure

b. arc list which encloses an exposure, and

c. neighboring exposures of a visible arc

2. **Original faces of unified object** SOLVER constructs a complicated object unifying simpler objects. During this operation, SOLVER keeps the information to be able to identify the original face from a unified face.

3. **Original face, edge or vertex of a projected region, arc or point** SOLVER can generate an image of an object under one viewer direction. During this operation, SOLVER keeps the information to be able to identify an original face, edge, and vertex from a region, arc, and point of the image, respectively.

SOLVER is used for deriving representative attitudes of an object, extracting work models for the interpretation process, and representing the interpretation results in our system.

REPRESENTATIVE ATTITUDE

A three-dimensional object varies its apparent shape depending on the viewer direction and rotation. These apparent shapes of an object fall into groups such that each group consists of roughly the same shapes. Here, "roughly the same" means that the same faces can be observed in almost the same condition. Some researchers explore this characterization with visible lines; this paper explores the characterization with faces observable by photometric stereo.

The number of observable regions of a non-convex object depends on the viewer direction. The photometric stereo can determine the surface orientation at the place where the three light sources project their light directly. A non-convex object is often observed as a few detectable regions which are isolated from each other by shadowing or mutual illumination between neighboring faces. The number of the isolated regions and their corresponding faces depends on the viewer direction. Thus, we can characterize the viewer direction based on the observable faces. (Note that we use the terms "object

attitude" and "viewer configuration" interchangeably. The object attitude and the viewer configuration have three degrees of freedom; two degrees of freedom in the viewer direction, and one degree of freedom in the viewer rotation. While the viewer rotation does not affect the number of observable regions, the viewer direction does.

Each viewer direction can be characterized by those faces visible from that direction. Let us suppose that

$$X_i = \begin{cases} 1 & \text{face i is visible} \\ 0 & \text{face i is not visible} \end{cases}$$

$(X_1, X_2, \ldots, X_n)$ denotes one label of an apparent shape based on the detectable faces under the photometric stereo. We can characterize each viewer direction with this label, and

The set of viewer directions that have the same visible face label becomes an attitude group. There are two ways to generate attitude groups: an analytic method, and an exhaustive method. If the target object is a convex object, then the analytic method is easy. Because the face visibility is determined by the relationship between the viewer direction and the surface orientation. The viewer direction has two degrees of freedom and can be described as a point on the Gaussian sphere at whose center a target object is located. The visible viewer directions of a surface is described as a circle on the Gaussian shpere. The circle center corresponds to the surface orientation, and the radius of the circle is $\pi/2$. Inside area of the circle corresponds to the viewer directions visible to the surface. Drawing these visible circles on the Gaussian sphere, attitude groups can be determined from combination of the circle covers on the sphere.

If the target object is non-convex, then the visible circle is distorted due to the self occlusion and the analytic method becomes difficult. Thus, the exhausted method is applyed. Essentially, the exhausted method generates all possible viewer directions and its apparent shapes of the object, and then examining face labels of the generated shapes gives the attitude groups. First task is to sample the Gaussian sphere evenly; A geodesic dome is used to tesselate the Gaussian sphere evenly [Brown 79]. Each tessellated triangle corresponds to a particular viewer direction. Thus, at each sampled viewer direction a observable shape of the object is generated using SOLVER. The observable shape gives a

face label, $X_1, X_2, \ldots, X_n$. After obtaining all face labels of all generated shapes, attitude groups are generated based on the face labels.

One representative attitude will be selected from each attitude group and each attitude group is represented by its representative attitude. That is, the viewer directions over one particular range are represented by one representative attitude. Usually, the viewer direction which gives the largest sectional area within the group is selected as the viewer direction for the representative attitude. The viewer rotation for the representative attitude is determined so that the maximum inertia direction agrees with the x axis on the image plane.

Figure 1 shows an example of this process. Figure 1a is a picture of an object. Figure 1b is a model synthesized using SOLVER. Figure 1c shows apparent shapes of the object observed from sixty different viewer directions, where the faces enclosed with bold lines are observable by the photometric stereo. These shapes are fallen into 7 attitude groups as shown in Figure 1d. Through face group 1 to group 5, five representative attitudes are generated as shown in Figure 1e. Since group 6 corresponds to a hole region of the object and group 7 has too small a visible area, no representative attitudes are generated from the groups 6 and 7.

WORK MODELS

The work models consist of original face information such as the original face inertia, the original face shape, the original face relationship, the original edge relationship, the surface characteristic distribution, and the extended Gaussian image. These work models will be used to classify one target region into a representative attitude, and to determine the attitude of an object observed as the target region. These work models are derived from SOLVER in the modeling process, and are derived from needle maps and/or edge maps in the determining process.

The work models are generated at each representative attitude. Since the surface orientation is available at each region from the needle map, the original face information can be recovered from the observed region information using an affine transform. For example, when the surface orientation, the affine matrix, and the observed region shape

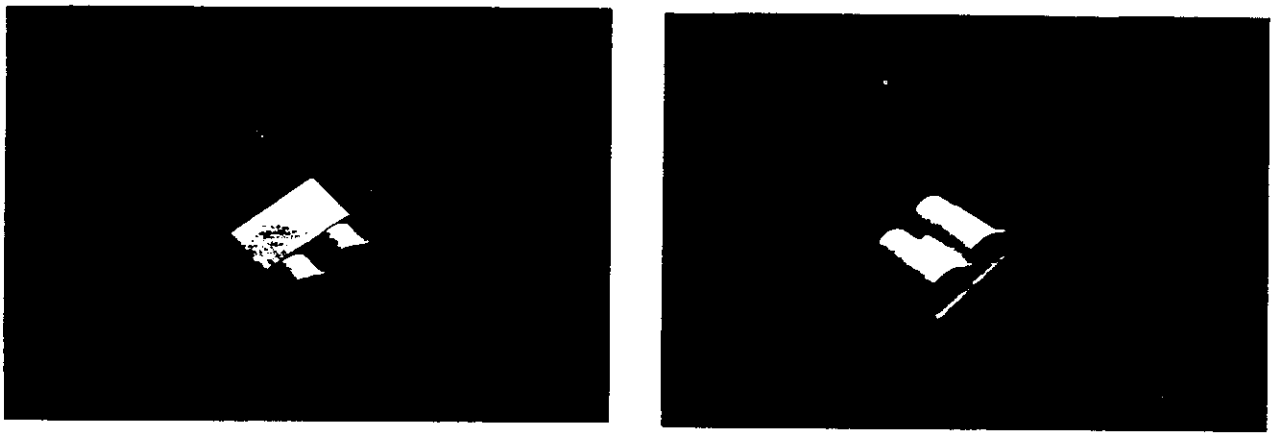


Figure 1a An object.

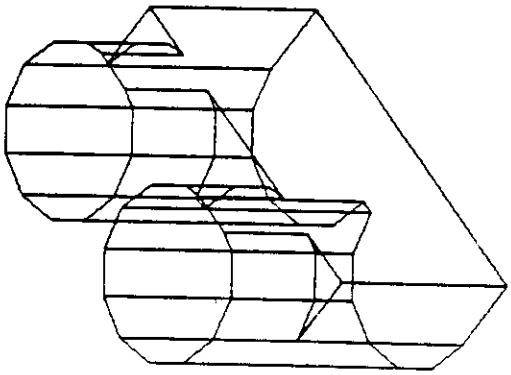


Figure 1b A synthesized model of the object in SOLVER.

Figure 1. An object and its representative attitude.

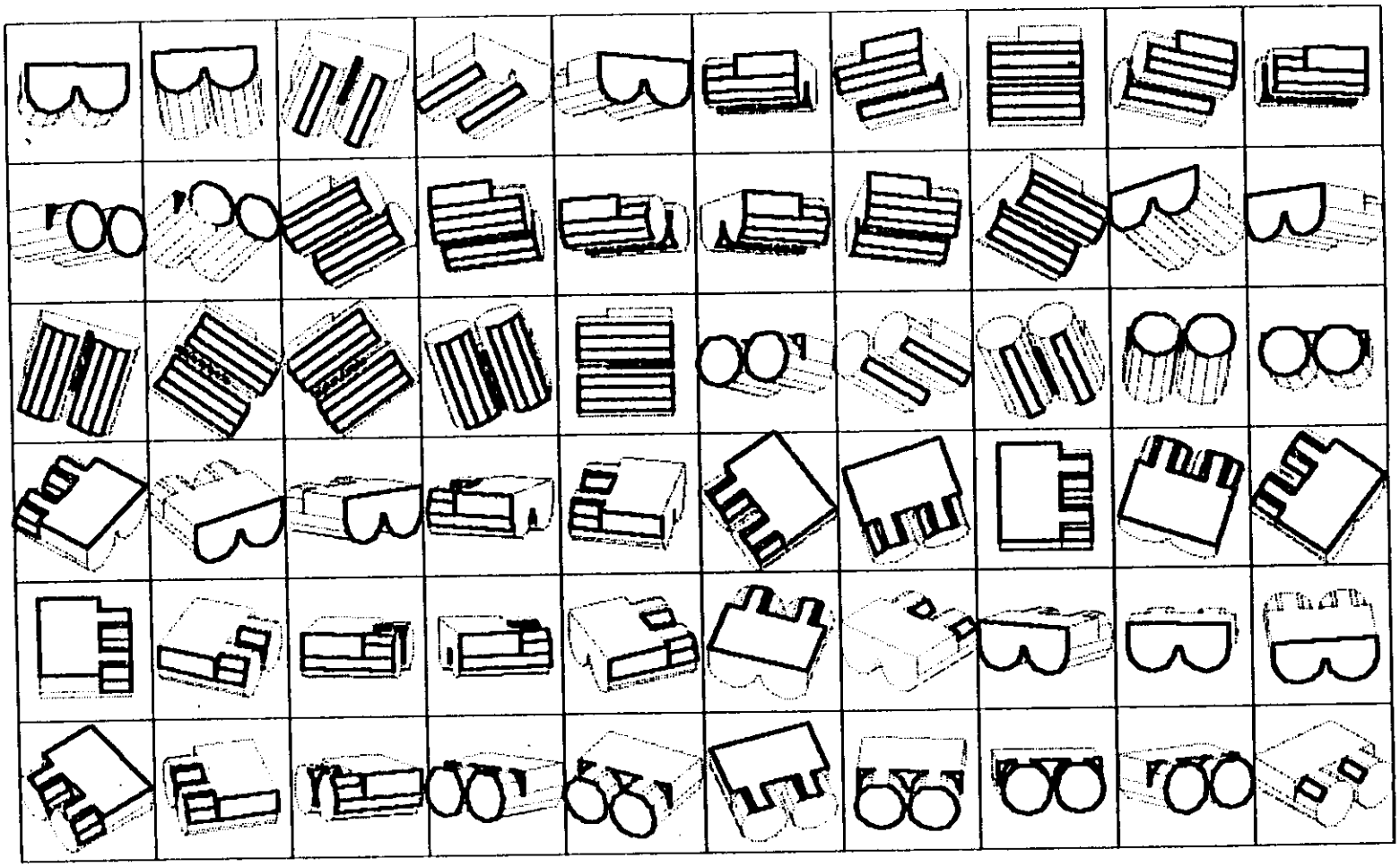


Figure 1c Sixty apparent attitudes of the object.

Figure 1 (continued).

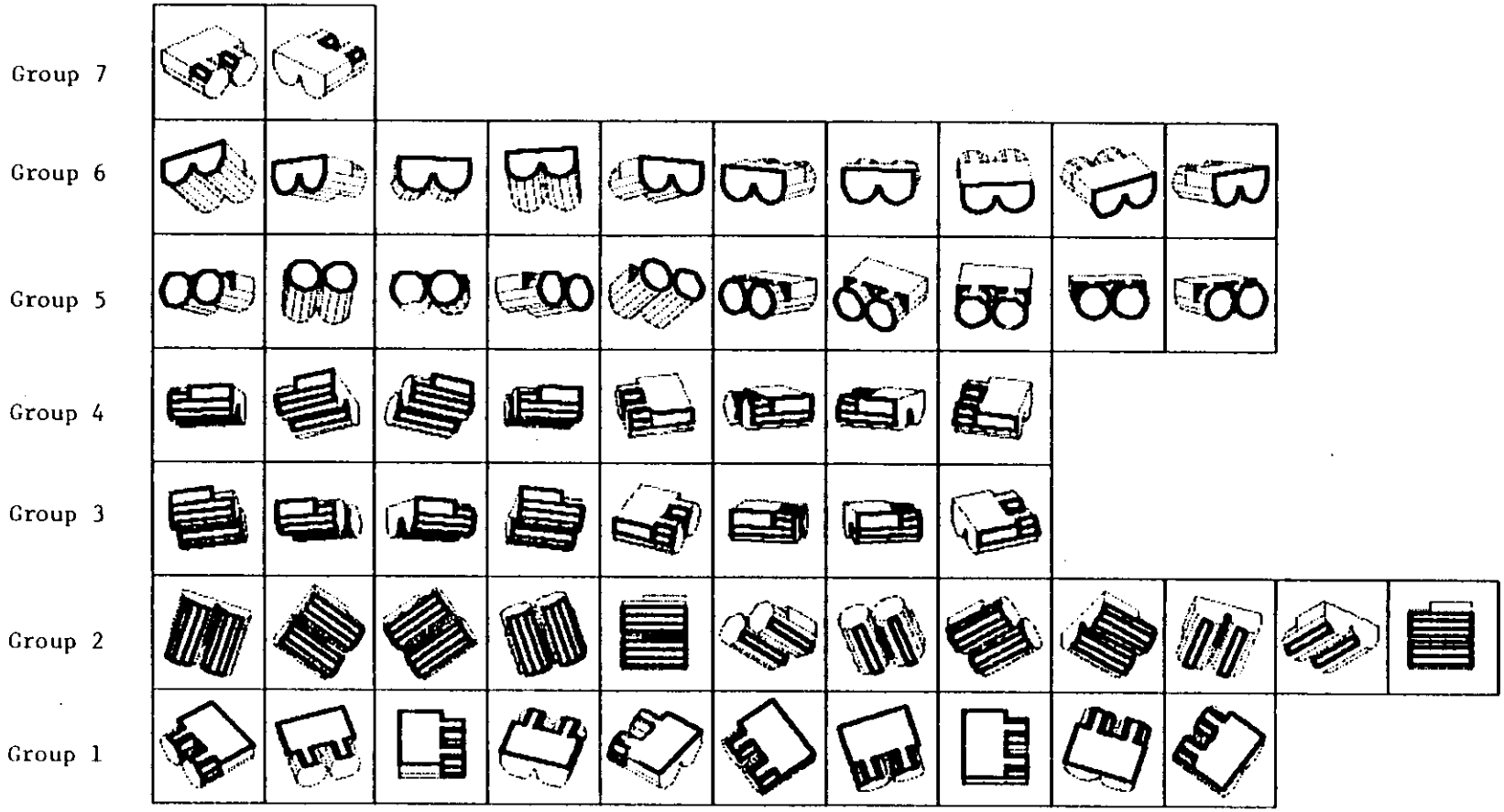
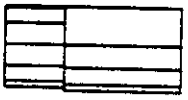


Figure 1d Seven attitude groups.

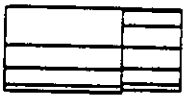
Figure 1 (continued).

Representative Attitude 5

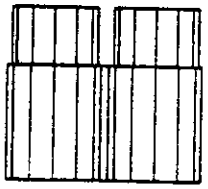
Representative Attitude 4



Representative Attitude 3



Representative Attitude 2



Representative Attitude 1

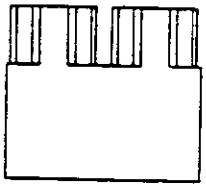


Figure 1e Five representative attitudes.

Figure 1 (continued).

are known, the original face shapes can be recovered from the skewed region shape with the affine transform. Information for only one attitude is necessary at each attitude group in which detectable faces are the same and they are reachable from each other by the affine transformation. The work models are, thus, generated at each representative attitude which represents one attitude group.

Let p,q be surface orientation of one face.

$$T=\begin{bmatrix} 1+p^2 & (pq)/(1+p^2) \\ 0 & (1+p^2+q^2)/(1+p^2) \end{bmatrix} \tag{1}$$

gives the affine matrix to recover the original face information from the observed face information.

Original Face Inertia

One work model is the original face inertia. The original face inertia gives the rough shape information of a face. In order to obtain the inertia, we have to convert a needle map into a binary map. Here, the binary map has 1 at each pixel where the surface orientation can be obtained, and 0 at each pixel where the surface orientation cannot be obtained. The obtained binary map is represented as $m(x,y)$. From this $m(x,y)$ and the affine matrix T,

$$\begin{aligned} I_{xx} &= \int m(x',y')dx'dx' \\ I_{xy} &= \int m(x',y')dx'dy' \\ I_{yy} &= \int m(x',y')dy'dy' \end{aligned} \tag{2}$$

where

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = T\begin{pmatrix} x-\bar{x} \\ y-\bar{y} \end{pmatrix}$$

and $(\bar{x},\bar{y})$ is the observed mass center of the face. From these I_{xx},I_{xy},I_{yy} we can determine the maximum inertia I_{max} and the direction α as follows:

$$I_{max}=(I_{xx}+I_{yy}+\sqrt{(I_{xx}+I_{yy})^2-4(I_{xx}I_{yy}-I_{xy}I_{xy})})/2 \tag{3}$$

$$\alpha=(\tan^{-1}\{(2\,I_{xy})/(I_{xx}-I_{yy})\})/2 \tag{4}$$

Original Face Relationship

A non-convex object often appears as multiple isolated regions under the photometric stereo. In this case, the relationships between regions are used as a work model.

For each region, the relative position of other regions are stored. The relative position is described by a vector whose length corresponds to the distance between the mass centers of the two regions and whose direction indicates the direction from the mass center of the region to the other mass center based on the maximum inertia direction and the surface orientation of the region. If the region has no unique inertia direction, for example, a circular region, only the distance is stored.

Original Face Shape

The original face shape is also used to characterize a region. The face shape is described as the distance from the mass center of the face to the boundary of the face as a function of the angle round the mass center, $d=d(\theta)$. The rotation angle θ is calculated with respect to the maximum inertia direction. This is a two dimensional well-tessellated surface representation of the shape [Brown 79].

Original Edge Relationship

Some of the prominent edge information is also used. In some cases the needle map from the photometric stereo cannot determine the object attitude uniquely. In this case some of the prominent edge information is used to reduce this ambiguity. Thus, some of the edge information is stored if necessary.

The edge information is described by the starting position and the ending position. These positions are denoted relative to the mass center of the face and the maximum inertia direction. In application, a position is converted into the position on the image plane using the affine matrix. Then, the connecting place between the converted starting position and the converted ending position will be searched on the edge map to determine whether there is an edge or not.

Extended Gaussian Image

Roughly speaking, the extended Gaussian image of an object is a spatial histogram of its surface orientation distribution [Smith 79, Horn 79, Ikeuchi 81, Horn 84, Brou 84, Little 85]. Let us assume that there is a fixed number of surface patches per unit surface area, and that a unit normal is elected on each patch. These normals 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. 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.

The EGI has the following properties:

1. Neither the surface normal nor the Gauss map depend on the position of the origin. Thus, the resulting EGI is not affected by translation of the object.

2. When an object rotates, its EGI also rotates. However, 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.

Surface Characteristic Distribution

The surface characteristic distribution is available from the surface orientation distribution. A surface patch has a characteristic such as planar, cylindrical, elliptic, or hyperbolic. The first and the second fundamental forms can be obtained from the surface orientation and its derivatives, and from these the Gaussian curvature and the mean curvature are obtained [do Carmo 76, Brady etal 85]. The characteristic, defined in terms of the Gaussian curvature and the mean curvature are independent of the viewer direction and the rotation.

Let us denote surface orientation as (p,q), where $p=z_x$ and $q=z_y$. Then, the first fundamental forms E,F,G are

$$\begin{aligned} E &= (1+p^2) \\ F &= pq \\ G &= (1+q^2). \end{aligned} \tag{5}$$

The second fundamental forms e,f,g are

$$e=p_x/\sqrt{1+p^2+q^2}$$
$$f=p_y/\sqrt{1+p^2+q^2} \quad (6)$$
$$g=q_y/\sqrt{1+p^2+q^2}$$

These coefficients give the Gaussian curvature K and the mean curvature H of the surface as follows:

$$K=(eg-f^2)/(EG-F^2)$$
$$H=(1/2)((eG-2fF+gE)/(EG-F^2)) \quad (7)$$

Gaussian curvature K and mean curvature H determine the surface characteristic as follows:

1. $K=0$ *and* $H=0$ *then planar surface*
2. $K=0$ *and* $H \neq 0$ *then cylindrical surface*
3. $K > 0$ *and* $H > 0$ *then convex elliptic surface*
4. $K > 0$ *and* $H < 0$ *then concave elliptic surface*
5. $K < 0$ *then hyperbolic surface*

The surface characteristic distribution is stored at each representative attitude. A subregion is generated based on a surface characteristic, and described by the surface characteristic and the rectangular existence area whose vertices are referenced to the coordinate of the mass center and the maximum inertia direction. In application, the vertex positions are converted to image plane coordinates using the affine matrix. Then, the corresponding area is examined to determine whether surface patches having the characteristic exist or not.

INTERPRETATION TREE

An interpretation tree determines the viewer direction and the rotation of an object observed as one target region. The interpretation tree reduces the freedom step by step by comparing the most appropriate feature in work models with the feature obtained from the observed data over one target region. The interpretation tree consists of three parts. The first part classifies an unknown region into one of the representative attitudes. This operation reduces some of the freedom in the viewer direction. At this point, non-linear

shape changes have been solved; after this point, visiblity of each face does not change depending on the viewer direction within each attitude group. The second part determines the viewer direction uniquely. The third part determines the viewer rotation around the viewer direction uniquely.

The interpretation tree is derived by the extraction rules before execution of the determining process. Each extraction rule is examined to determine whether the rule can constrain some of the freedom in the viewer direction and the rotation. If the rule can constrain some of the freedom, the rule is adopted into the interpretation tree. This adoption operation generates an interpretation tree to determine the viewer direction and the viewer rotation completely.

Classifying into Representative Attitude

The first step to derive the interpretation tree is to generate branches from the root to the attitude groups. Branches are generated using the face labeling. The leaves of the tree correspond to the attitude groups, while the root corresponds to the unclassified stage.

The attitude group depends on the face groups which generate the face labeling. At first we will put faces of an object in area order; $f_1, f_2, \ldots, f_n$. Then, we will consider the subsets of the face groups $g_1=\{f_1\}, g_2=\{f_1,f_2\}, \ldots, g_n=\{f_1,f_2,\ldots,f_n\}$.

g_1 is a subset which consists of only one face f_1, which is the largest among the faces of the object. This subset generates a face label x_1. Using this label the general attitude space is divided into two sub-attitude groups. Under any attitude in the attitude group which has $x_1=1$, the photometric stereo can observe the largest face 1; under any attitude in the attitude group which has $x_1=0$, the photometric stereo cannot observe the largest face 1.

Next we will consider the relationship between the attitude group from g_{i+1} and the attitude group from g_i. The attitude group of g_{i+1} is obtained by dividing the attitude group of g_i based on the visibility of f_{i+1}. This division sequence generates a tree structure which gradually reaches the final attitude groups. This tree structure will be used as the structure of the interpretation tree.

Let us consider the relationship between EGI and the face-labelling. Among three degrees of freedom in object attitude, the viewer rotation does not affect the visibility of the surface. The viewer direction has two degrees of freedom and can be expressed as a point on the Gaussian sphere. The surface orientation and its EGI mass can be also expressed as a point on the Gaussian sphere. We will express both in the same sphere.

The photometric stereo determines a visible disk on the Gaussian sphere. The center of the disk locates at the point corresponding to the surface orientation of the surface; the center of the disk is the point of EGI mass of the surface. The radius of the disk is determined depending on the characteristic of the photometric stereo. (in our system, $\pi/4$.) The surface is visible from any viewer direction within this visible disk. Thus, any viewer direction inside of this disk has the label x_i=1, where the face is f_i.

The viewer directions having x_1=1 are contained in the disk whose center is the heaviest EGI mass. Since the faces are sorted in area order, the labelling operation described previously sections the Gaussian sphere sequentially with the visible disks of the EGI masses in weight order.

Figure 2 shows the branch obtained from the object shown in Figure 1. In the application, it often occurs that two faces have the same area. In this case, at the first step, we will divide the attitude groups into sub-attitude groups; any one of the faces are observable (xx), and none of the faces are observable (00). Then, (xx) attitude groups are divided on the visibility of the faces. This is because we will divide the resembling attitudes at the later stage. The B0 branch corresponds to the two cylindrical surfaces, B1 corresponds to the wide planar surface, B2 corresponds to the hole region, B3 corresponds to the two circular surfaces, and B4 and B5 correspond to the side planar surfaces. These branches divide the attitude groups into seven attitude groups.

Classification rules

This section gives rules to generate the classification part of the interpretation tree. Each branch examines whether one of the rules can discriminate the attitude groups. If one of the rules can discriminate, the the rule is registered at the branch. The decision whether the rule can divide them or not is made by human at present.

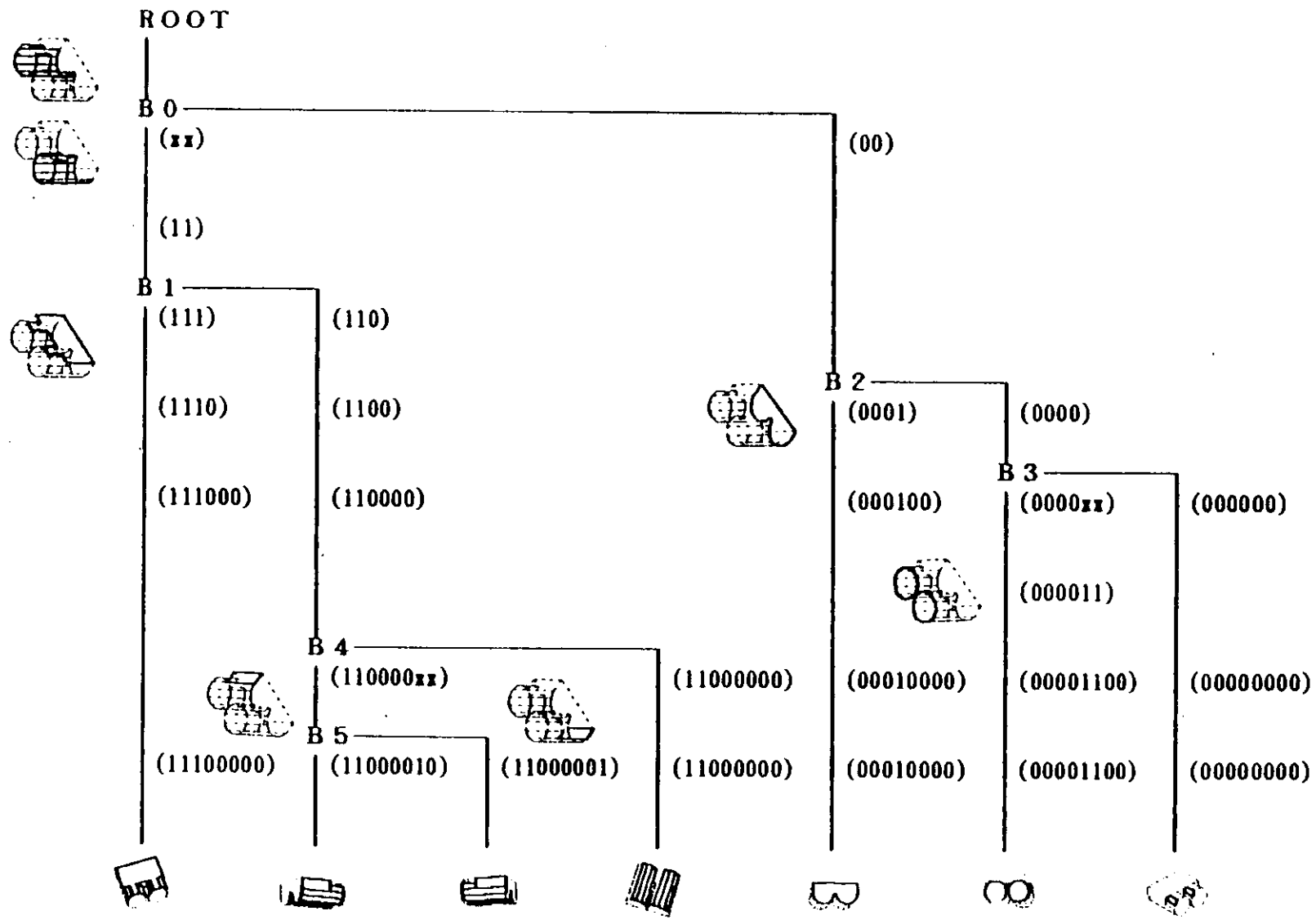


Figure 2. Branches based on face labelling.

L1: *Comparison based on the original face inertia.*

L2: *Comparison based on the original face shape.*

L3: *Comparison based on the extended Gaussian image.*

L4: *Comparison based on the surface characteristic distribution.*

L5: *Comparison based on the edge distribution.*

L6: *Comparison based on the region distribution.*

L7: *Comparison based on the relationship between a particular edge and a particular surface characteristic distribution.*

If the observed shape of an object cannot be classified into a representative attitude with these rules, it means that the object is observed with the same number of regions whose area sizes, inertia moments, edge distributions, and surface characteristic distributions are identical in two different attitudes. Such objects are beyond the scope of our technique.

Deriving the classification part of the tree

The classification part of the interpretation tree, Figure 3, is generated for the object shown in Figure 1a. At the B0 branch, the rule L1 (original face inertia) can divide the all attitude groups into two attitude groups. At the B1 branch, rule L1 can divide the attitude groups. Both B2 and B3 have branches at which the attitude groups are not visible. Thus, these branches are pruned.

At branch B4, none of L1 (inertia), L2 (shape),L3 (EGI), L4 (characteristic), L5 (edge) can divide the attitude groups. L6 (topology) can divide the branch. L7 (edge-region) can discriminate the attitude groups at the branch.

Thus, B0-L1, B1-L1, B2-pruned, B3-pruned, B4-L6, B5-L7 are adopted into the interpretation tree. Since B0 and B1 branches have the same rule and they are consecutive, they are joined into a three-branch node.

Determining Viewer Direction and Viewer Rotation

Determination rules

This section gives the rules to generate the part of the interpretation tree which determines the viewer direction and the rotation. If one rule can reduce some of the remaining freedom in the viewer direction and rotation, the rule will be adopted into the tree. The decision whether the rule can reduce the freedom or not is made by human at present.

A1: *Using the mass center of EGI mass distribution.*

A2: *Using the extended Gaussian image.*

A3: *Using the position of observable areas distribution.*

A4: *Using the inertia direction of original face.*

A5: *Using the rotation of original face shape.*

A6: *Using the position of the surface characteristics distribution.*

A7: *Using the position of the edges.*

A8: *Using the position of the edges with respect to the position of the surface characteristics distribution.*

If we cannot determine the viewer direction and the rotation with these rules, it means that the object is observed with the same number of regions whose area sizes, inertia moments, edge distributions, and the surface characteristic distributions are identical in two different attitudes. Such objects are beyond the scope of our technique.

The viewer direction and rotation are determined at each representative attitude using the most effective feature at each step. The most powerful rule for determining the viewer direction and rotation depends on the representative attitude and the stage of the determining process. Thus, we will discuss which rule will be used for generating the determination part of the interpretation tree at each representative attitude.

Representative Attitude S1

The main visible part of this representative attitude is a planar surface. A1 (EGI mass center) can determine the viewer direction, while viewer rotation can be constrained with neither A1 nor A2. More precisely, since the observable region of representative attitude S1 is a planar surface, both the EGI and the EGI mass center position [Ikeuchi 83] can determine the viewer direction uniquely. However, neither the EGI distribution nor EGI mass center over the planar surface can constrain the viewer rotation around the viewer direction. Thus, the other rules should be applied to determine the viewer rotation.

Since the representative attitude has only one observable region, A3 (region distribution) cannot be applied to this S1 representative attitude. A4 (inertia direction) can constrain the viewer rotation up to two directions. Between the two directions, A5 (original face shape) can determine the viewer rotation uniquely. Thus, A1 (EGI mass center), A4 (inertia direction), and A5 (original face shape) are adopted into the tree to determine the viewer direction and the rotation at representative attitude S1.

Representative Attitude S2

This representative attitude has two observable regions of cylindrical surfaces. A1 (EGI mass center) can determine viewer direction, while the viewer rotation cannot be constrained with A1.

Theoretically, the EGI distribution can determine the viewer direction and the rotation uniquely in this representative attitude. However, the determined rotation is very noisy. Thus, we will use the other features to determine the viewer rotation.

Since this representative attitude has two observable regions, A3 (region distribution) is applicable and can constrain the viewer rotation up to two directions. None of A4 (inertia direction), A5 (original face shape), nor A6 (surface characteristic) can constrain the remaining freedom of the viewer rotation. A7 (edge distribution) can determine the viewer rotation uniquely. Thus, A1 (EGI mass center), A3 (region distribution), and A7 (edge distribution) are adopted into the tree.

Representative Attitude S3

Representative attitude S3 has one observable region which mainly consists of three parts: a planar surface patch and two cylindrical surface patches. A1 (EGI mass center) can determine the viewer direction, while the viewer rotation is difficult to determine in practice due to the same reason as with representative attitude S2.

A3 (region distribution) cannot be applied to this representative attitude due to the single observable region. A4 (inertia direction) can constrain the viewer rotation up to two directions. Neither A5 (original face shape) nor A6 (surface characteristic) can constrain the remaining freedom. A7 (edge distribution) can determine the viewer rotation uniquely. Thus, A1 (EGI mass center), A4 (inertia direction) and A7 (edge distribution) are adopted into the tree.

Representative Attitude S4

The features used to determine the viewer direction and the rotation are the same as those of the representative attitude A3.

Representative Attitude S5

Representative attitude S5 has two regions observed separately which come from two planar surfaces. Thus, A1 (EGI mass center) can determine viewer direction, while the viewer rotation is difficult to constrain with A1 for the same reason as with representative attitude S1. Since this representative attitude has two observable regions, A3 (region distribution) is applicable and can constrain the viewer rotation up to two directions. None of A4 (inertia direction), A5 (original face shape), nor A6 (surface characteristic) can constrain the remaining freedom of the viewer rotation. A7 (edge distribution) can determine the viewer rotation uniquely. Thus, A1 (extended Gaussian image), A3 (region distribution), and A7 (edge distribution) are adopted into the tree.

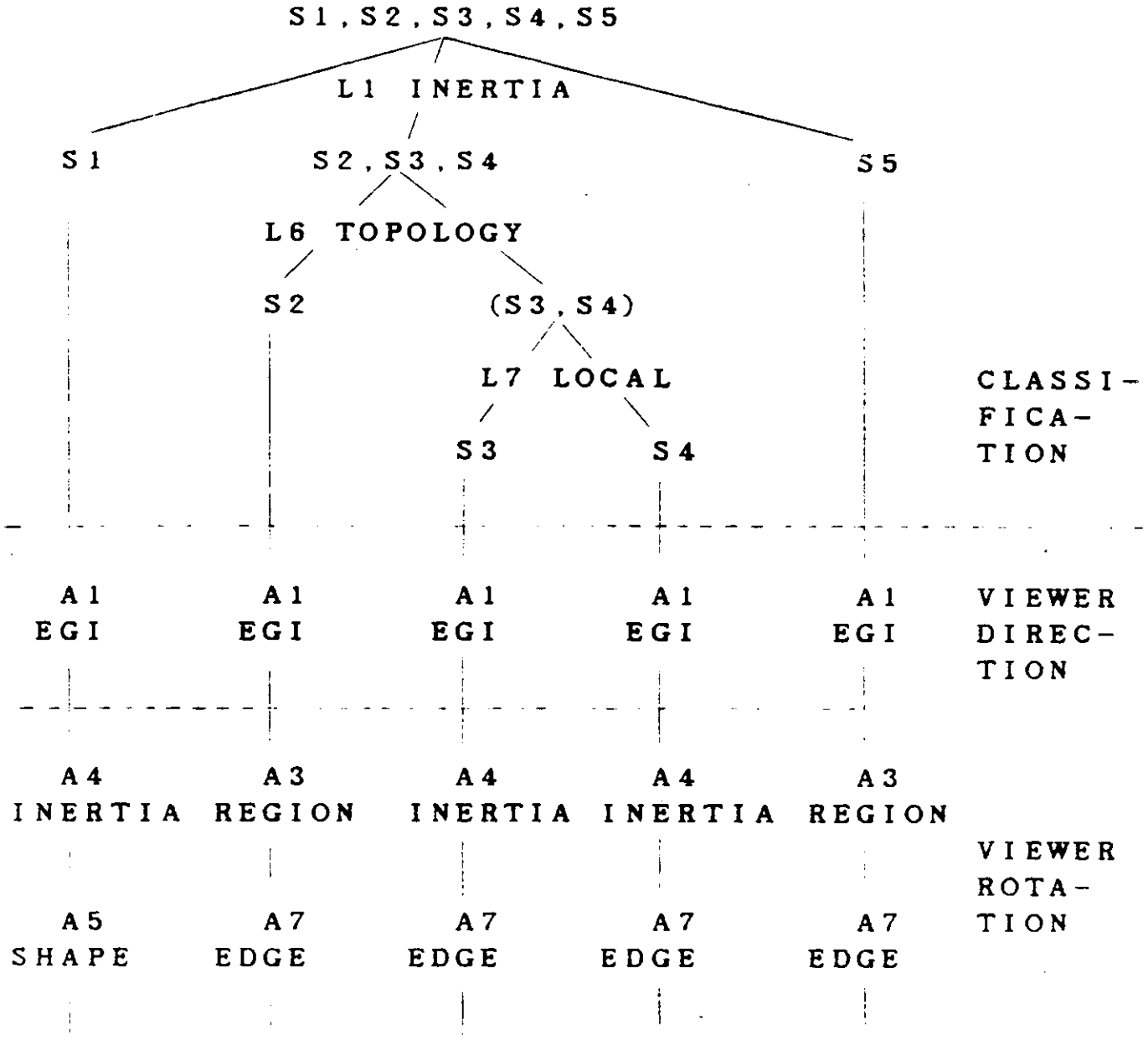


Figure 3. The interpretation tree.

APPLYING THE INTERPRETATION TREE (Run Mode)

Attitude Determination by the Interpretation Tree

The system can use three kinds of maps: edge maps, needle maps, and one depth map. Three maps can be obtained by differentiating three intensity maps also to be used for the photometric stereo. A needle map can be obtained by the photometric stereo system. A depth map can be obtained by comparing a pair of needle maps which are generated by a dual photometric stereo system [Ikeuchi 85]. The edge maps, the needle map, and the depth map are represented in the same coordinate system; that is, all pixels having the same X-Y coordinates correspond to the same physical point.

The highest region is determined from the depth map. This highest region will be sent to the interpretation tree as the target region. The interpretation tree extracts necessary features from the region. These features will be transformed according to the procedures defined in the interpretation tree. These transformed features will be compared with features in the work models defined in the interpretation tree. Following this procedure, the target region will be classified into one of the attitude groups, and then the precise attitude and position determined.

Determining A Collision-Free Grasp Configuration

The obtained attitude and position of an object will be registered in the world of SOLVER. The object model will be placed in this world under the obtained position and attitude. The regions around the target object will be expressed as dodecahedral prisms whose heights correspond to the heights of the regions and whose shapes correspond to the approximations of the regions.

By using this representation, we can get a collision-free configuration to grasp the target region. At each representative attitude, stable grasp configurations are defined prior to execution. These grasp configurations are transformed into the world coordinates using the obtained attitude. Based on this transformed configuration, the gripper work space is generated from the swept out volume of the fingers.

Using SOLVER's intersection-check mechanism, collisions between the work space and obstacle dodecahedral prisms is examined. If none of the prisms intersects the workspace prism, then that stable grasp configuration is a collision-free configuration. If any of the prisms intersects the workspace prism, then one of the fingers collides with the prism, and we cannot take that configuration.

Experiment

Case 1: Attitude group 1

Figure 4 shows one of the input scenes, where the white arrow indicates the highest region. From this scene, the edge map shown in Figure 4b is obtained. The photometric stereo system gives the needle map shown in Figure 4c. Further, the depth map shown in Figure 4d is obtained by the pair of the photometric stereo systems.

This highest region will be given to the interpretation tree. The interpretation tree calculates the inertia moment of the original face observed as the region (L1). The mass center and the region distribution can be obtained over the binary map which has been converted from the needle map to have 1 at the place where the surface orientation is determined, and to have 0 at the place where the surface orientation is not determined. Then, the affine matrix is obtained from the the surface orientation distribution over the region using Equation 1. Finally, the interpretation tree can determine the inertia moment of the original face using the affine matrix and the region distribution with Equation 2 through Equation 4. Figure 5a shows the region distribution and the square which is displayed by the interpretation tree. The square has the same inertia and direction as the original face. The interpretation tree determines that this region belongs to the representative attitude S1 based on the inertia value.

The interpretation tree uses the EGI mass center to determine the viewer direction (A1). This EGI mass center is obtained from the surface orientation distribution over the target region by the interpretation tree.

The interpretation tree determines the viewer rotation up to two directions using the inertia direction (A4). Branch A5 in the interpretation tree requires the original face shape to determine the viewer rotation uniquely. Figure 5b shows the original face shape

Figure 4a. Input scene.The white arrow indicates the highest region.

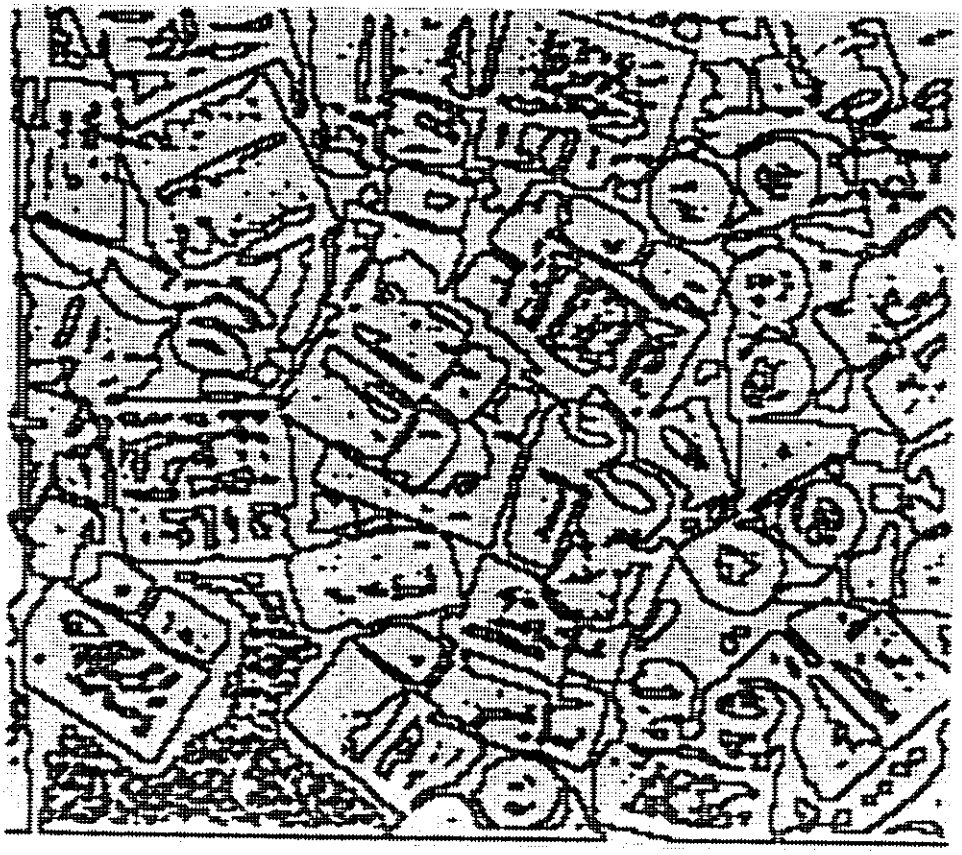


Figure 4b. The edge map obtained from the scene.

Figure 4. Input scene and maps.

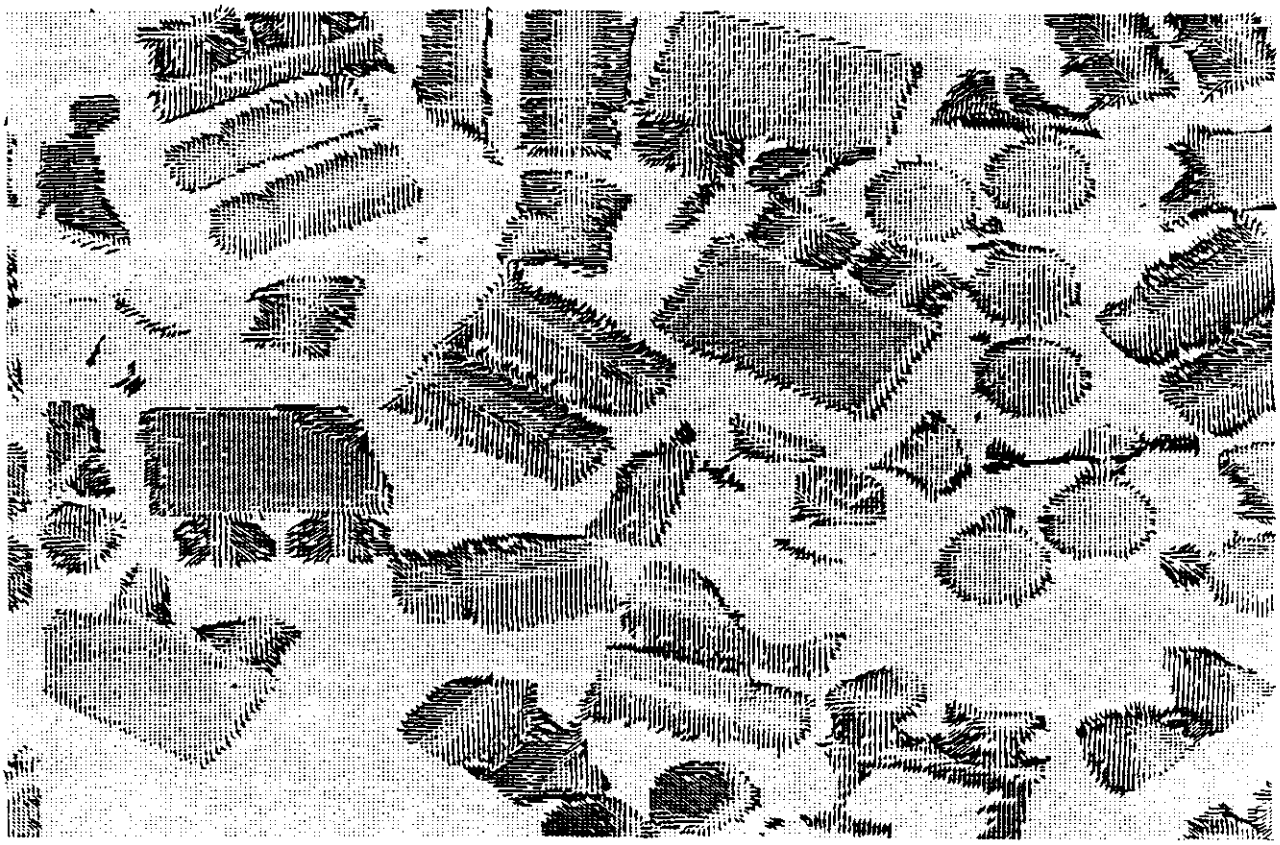


Figure 4c. The needle map obtained by the photometric stereo system.

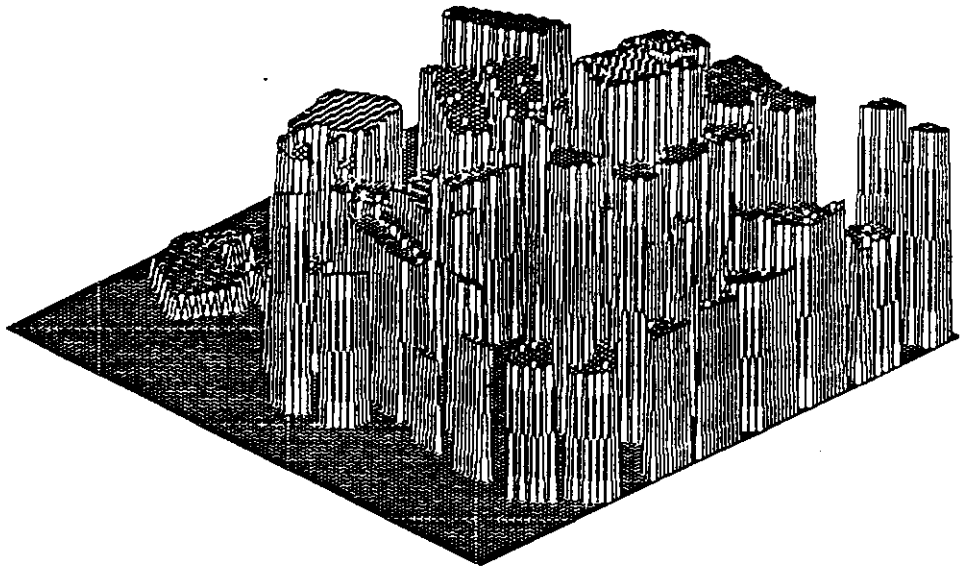


Figure 4e. The depth map obtained by the dual photometric stereo system.

Figure 4 (continued).

obtained from the target region. In this case, however, the interpretation does not measure the difference between the observed shape and the shape from the models in all directions, but only checks the crack direction of the observed region with respect to the inertia direction under the two possible rotations. Since the viewer rotation is constrained up to the two directions, the interpretation tree determines the object attitude in the space by this comparison.

SOLVER represents the object in the world model using the object position and the attitude obtained by the interpretation tree. The object position can be obtained from the depth map. Around the target region, there are a few regions which have not been processed by the interpretation tree at this time. These neighboring regions are expressed as dodecahedral prisms in the world model. The height of a prism agrees with the height of the corresponding region, and the cross section of the prism is an approximation of the region shape by the dodecagon. These dodecahedral prisms are also represented in the world model in SOLVER.

The stable grasp configurations of representative attitude S1 are the following ones.

1. The center of the griper exists on either the maximum axis of inertia or the minimum axis of inertia

2. The normal direction of the griper agrees with either the maximum axis of inertia or the minimum axis of inertia.

3. The approach direction of the griper agrees with the average surface orientation of the target region.

At each stable configuration, a corresponding sweep-out volume of the fingers is generated. The intersections between this generated volume and the obstacle pyramids are examined by SOLVER. In this scene, all stable configurations are collision-free configurations. Figure 5e shows two representative collision-free configurations of the griper.

Case 2: Attitude group 2

Figure 6a shows a second example. The white arrow in the picture indicates the highest region. The interpretation tree calculates the original face inertia of the region from the

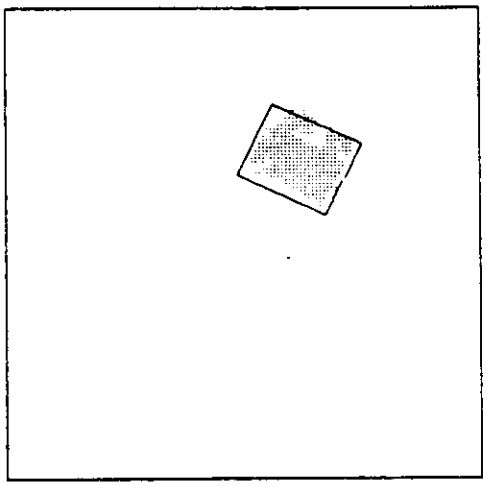


Figure 5a the target region and the original face inertia.

Figure 5b The original face shape recovered by the Affine transformation. The shape is represented using 2D WTS.

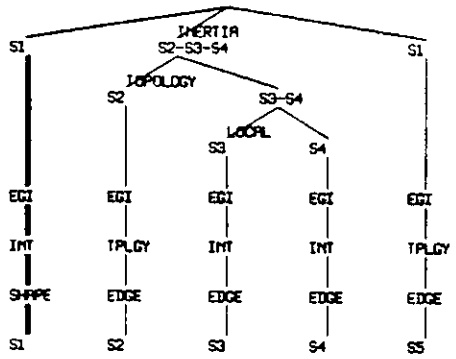


Figure 5c The decision path in the interpretation tree.

Figure 5 An interpretation example: attitude group 1.

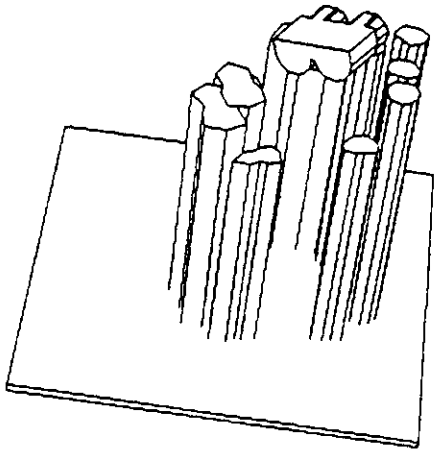


Figure 5d The obtained position, the obtained attitude, and the neighboring regions.

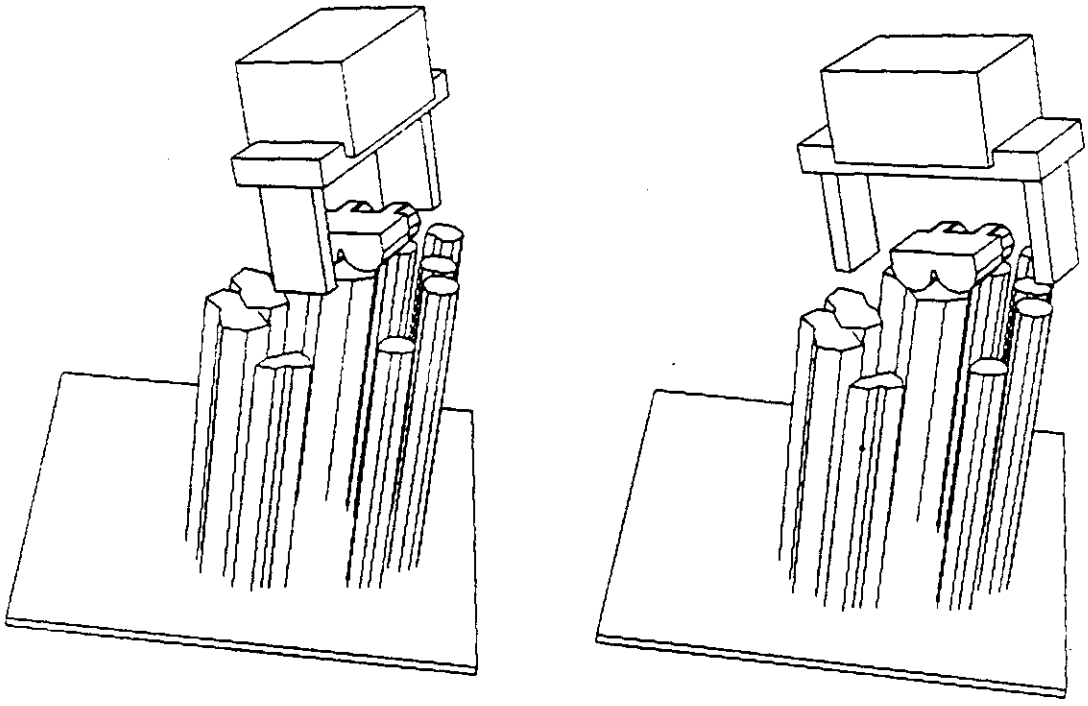


Figure 5e Two collision-free configurations.

Figure 5 (continued).

binary map converted from the needle map and the affine matrix obtained from the needle map over the target region. Figure 6b shows the square which has the same inertia direction and inertia value as the obtained inertia moment. The interpretation tree determines this region to belong to the group of representative attitude (S2, S3, S4) from the inertia value (L1).

The interpretation tree makes the distinction between the representative attitude (S2) and the group (S3, S4) by determining whether a brother region exists having the same inertia direction and the inertia value around the target region. The interpretation tree tries to find such a brother region; it succeeds, as shown in Figure 6b, where the target region and the brother region are connected with a solid line. From this evidence, the interpretation tree determines that the target region and the brother region come from the same object and belong to representative attitude S2 (L6).

The interpretation tree makes an EGI-mass center comparison to determine the viewer direction (A1). From the direction of the brother region, the viewer rotation is determined up to the two directions (A3).

The edge distribution is necessary to determine the viewer rotation uniquely (A7). The interpretation tree only examines the existence of the edge distribution whose direction agrees with the edge direction under one of the two possible rotations, at the place where one of the two rotations is supposed to make the edge distribution. This predicted place and the predicted direction can be obtained by applying the affine transform to the edge representation in the work models. In Figure 6c, the dotted lines indicate the distribution of edges over the target region and the broken lines indicate the search areas for the edge distributions. The solid lines in Figure 6c indicate the edges found to have the supposed directions at the supposed places under two possible rotations of the object. One of the two rotations is determined by the comparison of the edge distributions. The interpretation tree determines the object attitude in the space uniquely up to this point. The decision flow on the interpretation tree is expressed as the bold line in Figure 6d.

Representative attitude S2 has the stable grasp configuration whose normal direction corresponds to one of the two principal inertia axes, and whose approach direction

corresponds to the average surface orientation of the target region. According to SOLVER, one of the stable configurations whose normal direction agrees with the minimum inertia direction hit one of the obstacle prisms (indicated by the arrow) as shown in Figure 6e. Thus, we cannot use this configuration. On the other hand, the stable configurations corresponding to the maximum inertia direction are collision-free. Thus, we can pick up the object using one of these configurations. One of the collision-free configurations is shown in Figure 6e.

Case 3: Attitude group 4

Figure 7a shows the third example classified into attitude group 4. The white arrow indicates the highest region. The interpretation tree determines that the target region belongs to the group of the representative attitude (S2, S3, S4) based on the original face inertia. Figure 7b shows the target region and the obtained moment-compatible square of the original face.

The interpretation tree makes the distinction between the representative attitude S2 and the group (S3, S4) based on the existence of a brother region (L6). Since there are no brother regions around this target region, the region is determined to belong to the group (S3, S4).

The surface characteristic distribution with respect to the edge distribution resolves the ambiguity between S3 and S4 (L7). The interpretation tree examines which attitude has the more consistent surface characteristic distribution. First, the interpretation tree searches the existence of the edge distribution at the supposed places at the supposed directions from the inertia direction as in the S2 case. Figure 7c indicates the edge distribution found as the solid lines. Second, the interpretation tree generates both the surface characteristic distribution of S3 and that of S4 based on the inertia direction and the edge distribution.

Representative attitude S3 has the planar surface at the left region and the cylindrical surface at the right region with respect to the edge distribution shown in Figure 7c. Figure 7d shows the surface characteristic distribution which agrees with the distribution of the representative attitude S3. Note that since no distributions agree with the observed

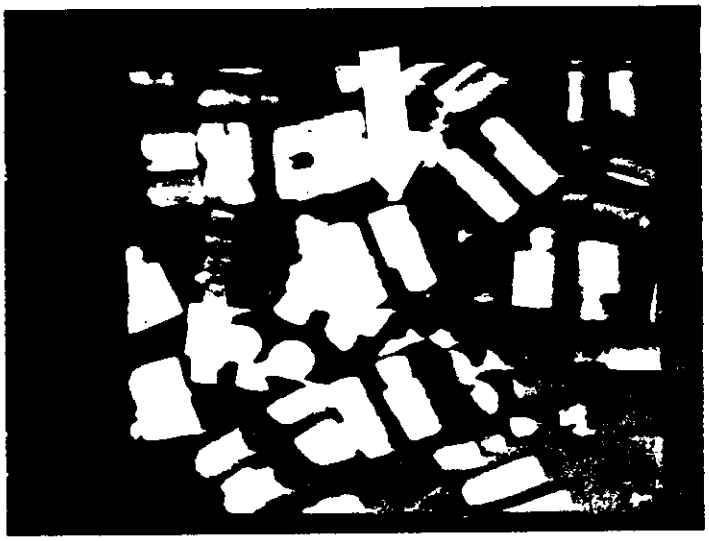


Figure 6a Input scene. The white arrow indicates the highest region.

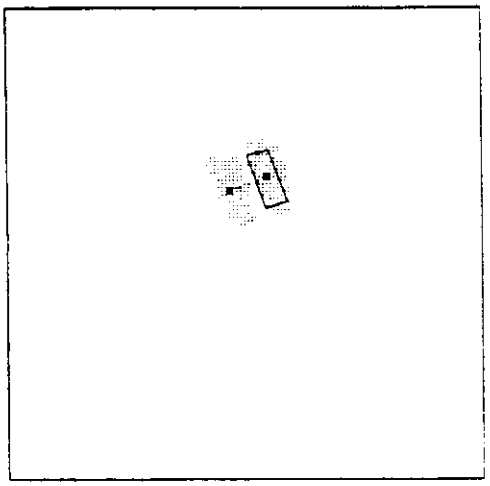
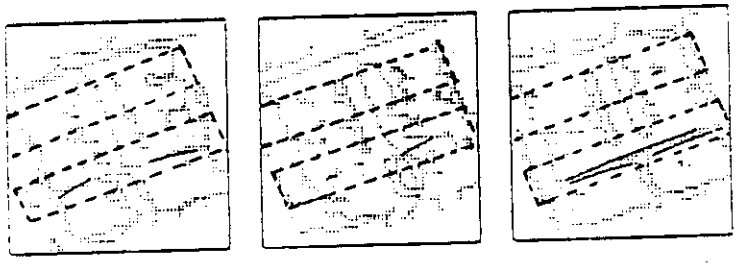


Figure 6b The target region and its brother region found by the algorithm.



Figere 6c Obtained edges. The interpretation tree only examines the existence of the edge distribution whose direction agrees with the edge direction under one of the two possible rotations, at the place where one of the two rotations is supposed to make the edge distribution. The dotted lines indicate the distribution of edges over the target region and the broken lines indicate the search areas for the edge distributions. The solid lines indicate the edges found to have the supposed directions at the supposed places under two possible rotations of the object.

Figure 6 An interpretation example: attitude group 2.

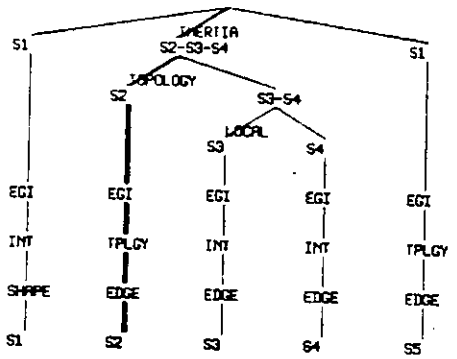


Figure 6d The decision flow on the interpretation tree.

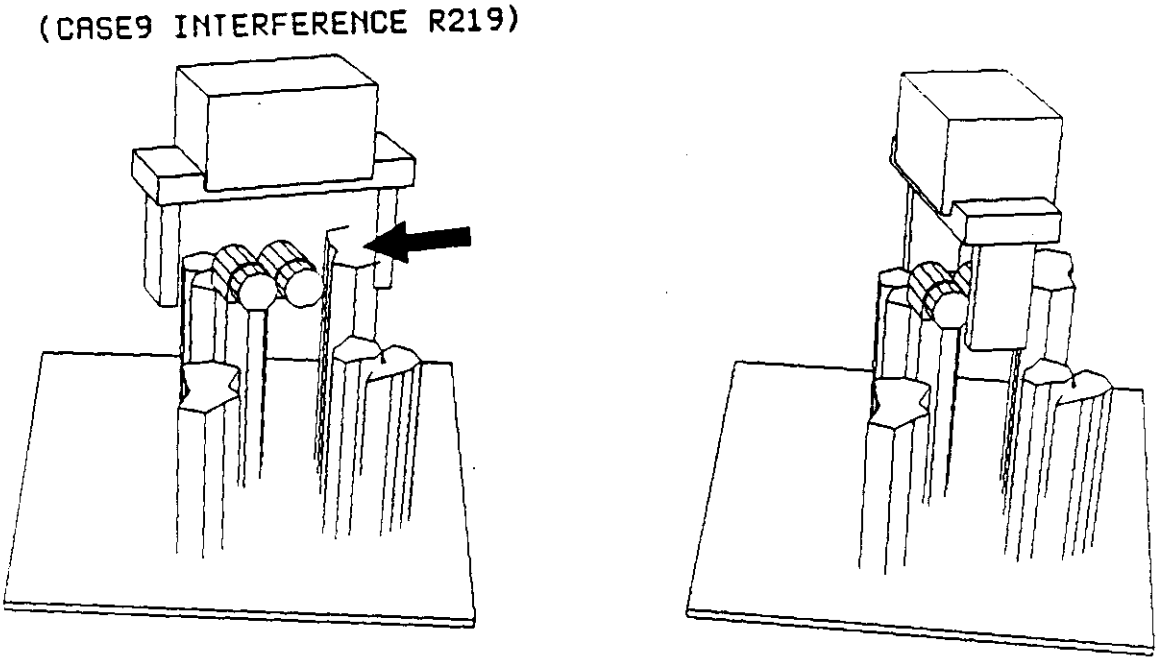


Figure 6e Collision configuration and safe configuratation. This representative attitude has the stable grasp configuration whose normal direction corresponds to one of the two principal inertia axes, and whose approach direction corresponds to the average surface orientation of the target region. According to SOLVER, one of the stable configurations whose normal direction agrees with the minimum inertia direction hit one of the obstacle prisms (indicated by the arrow). The stable configurations corresponding to the maximum inertia direction are collision-free.

Figure 6 (continued).

disbributions, the result figure shows white space. On the other hand, if the target region is assumed to belong to representative attitude S4, the region should have the cylindrical surface at the left region and the planar surface at the right region relative to the edge distribution. Figure 7e shows the characteristic distribution which agrees with representative attitude S4. The interpretation tree determines that the target region belongs to the S4 representative attitude.

The interpretation tree determines the viewer direction from the EGI mass center (A1). The viewer rotation is determined up to the two directions from the inertia direction (A4). To determine the viewer rotation uniquely, the edge distribution is necessary (A7); it had been obtained when the system used rule L7. The interpretation tree determines the object attitude from these comparisons, while the object position is obtained from the depth map. Figure 7f shows the decision flow on the interpretation tree.

Using the object position and attitude, the object is represented in the world model in SOLVER. The collision-free configurations are obtained using this representation as shown in Figure 7g.

CONCLUDING REMARKS

This paper describes a vision system to localize an object by an interpretation tree using a depth map, needle maps, and edge maps.

This system has the following characteristics:

1. The system requires one depth map, needle maps, and edge maps.

2. Representative attitudes are derived from a geometrical modeler, SOLVER, automatically.

3. The interpretation tree controls the localization process to use the most appropriate features at each stage of the localizatio.

4. The obtained attitude and position are represented in the world model in SOLVER for further use.

This paper develops a flexible interpretation by an interpretation tree using multiple

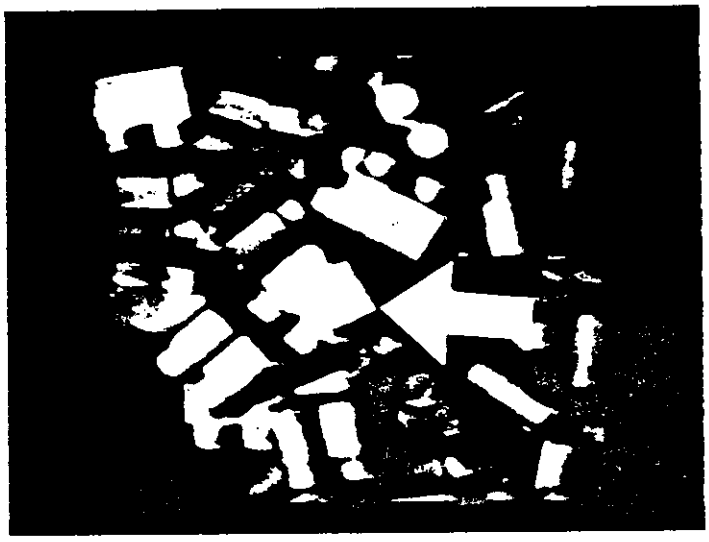


Figure 7a Input scene. The white arrow indicates the highest region.

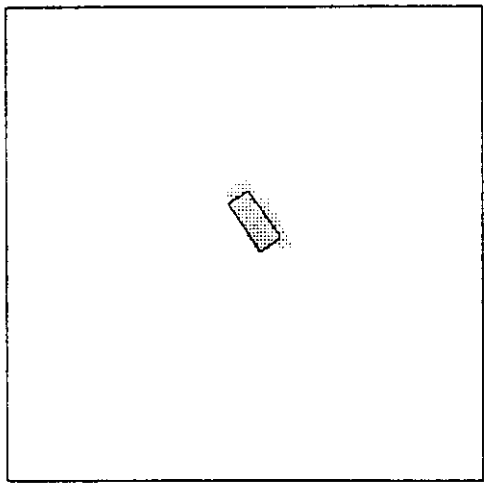


Figure 7b The target region and its original face inertia.

Figure 7 An interpretation example: attitude group 4.

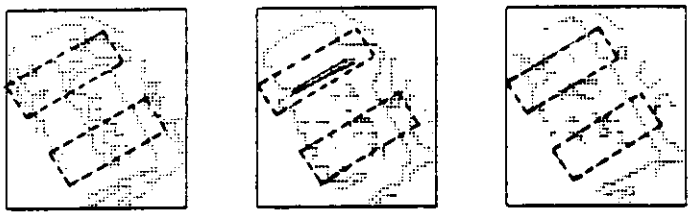


Figure 7c The edge distributions. The dot lines indicate output from an edge operator. The broken lines indicate search areas predicted from the model. The solid lines indicates the edges which corresponds to the model.

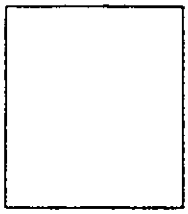


Figure 7d No surface characteristic districutions agree with the distributions of the representative attitude 3.

Figure 7e The characteristic distributions which agrees with representative attitude 4. The target region has the cylindrical surface at the left region and the planar surface at the right region relative to the edge distribution. This distributions correspond to the representative attitude 4.

Figure 7 (continued).

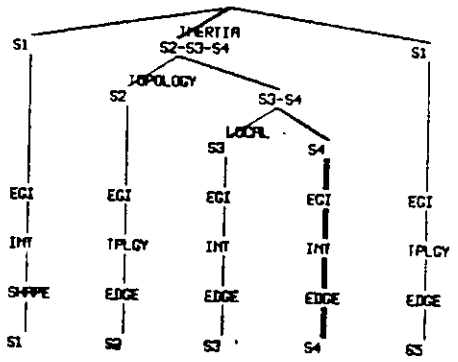


Figure 7f The decision flow on the interpretation tree.

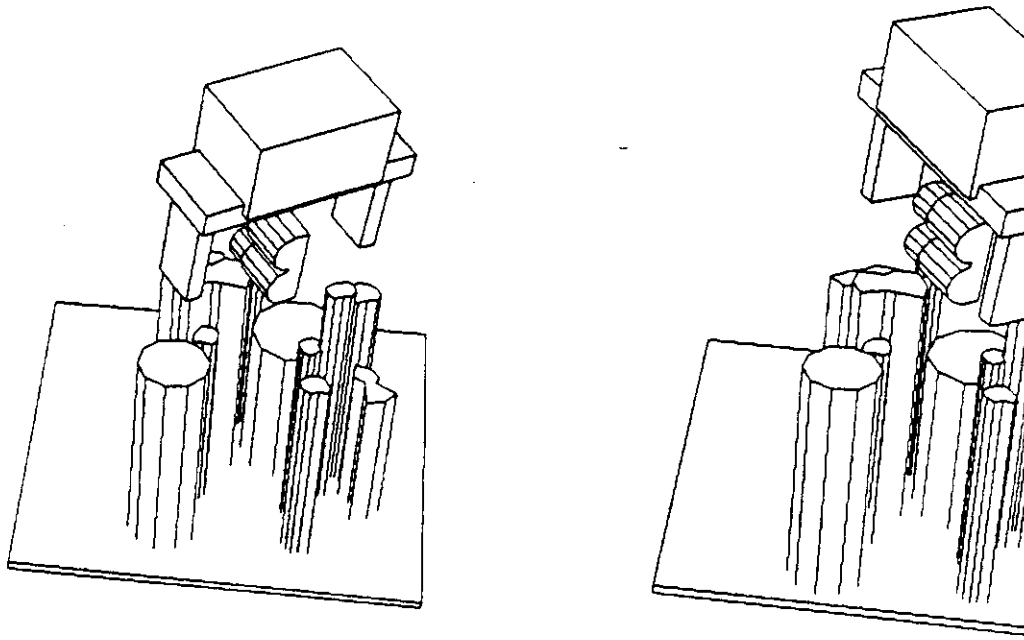


Figure 7g Two collision-free configurations.

Figure 7 (continued).

sensory inputs. Recent work in image understanding has led to techniques for computing surface orientation or surface depth. We can take various sensory inputs from the same scene by these methods. Since each technique has some merits and demerits, we have to select one appropriate feature among many available features in each processing stage. This paper proposes to use the interpretation tree for this purpose. This flexible interpretation matching should be further explored.

A geometrical modeler is used for the recognition problem. Models from a geometrical modeler possess rich geometrical features. Unfortunately however, the distance between the rich information and the information from the observed data is great. This paper uses the work models and the representative attitude to interface them. Effort is required to explore more convenient forms and methods to connect them.

The task of a vision system is to generate a description of the outer world. Some of the representations are symbolic, others use mathematic representations such as extended Gaussian imanges and generalized cylinders [Binford 71, Brooks 81, Shafer and Kanade 83]. However, since the representation is needed for manipulation by other modules such as planning and navigation, the representation must be easy to manipulate. This paper proposes to represent the outer world in the CAD model, because a CAD representation is an easy basis to achieving further tasks. Certainly there are many path finding programs that start from the polyhedral representations [Lozano-Perez 81]. How to express the outer world in such a representation should be explored more.

ACKNOWLEDGMENT

The author thanks Satoru Tomura and Ichiro Ogata of ETL, and Takeo Kanade, S.A. Shafer and R. Muller of CMU for their valuable comments and discussions.

REFERENCES

Ayache, N., Faverjon, B., Boissonnat, J. , and Bollack, B. (1984) "Automatic handling of overlapping workpieces", *Proc. International Conference on Pattern Recognition 84*, pp.837-839.

Baumgart, B. G. (1972) "Winged edge polyhedron representation", *STAN-CS-320*, Stanford Univ. A.I. Laboratory.

Besl, P. J. and Jain, R. C. (1985) "Intrinsic and extrinsic surface

characteristics", *Proc. Computer Vision and Pattern Recognition Conference*, IEEE, San Francisco, pp. 226-233.

Baird, M. L. (1977) "Image segmentation technique for locating automotive parts on belt conveyers", *Proc. 5th International Conference on Artificial Intelligence*, pp.694-695.

Binford, T.O. (1971) "Visual preception by computer", *Proc. IEEE Systems Science and Cybernetics Conf.*.

Birk, J. R. , Kelly, R. B. , and Martines, H. A. S. (1981) "An orienting robot for feeding workpieces stored in bins ", *IEEE Trans.*, Vol. SMC-11, No.2, pp.151-160.

Bolles, R. and Cain, R. A. (1982) "Recognizing and locating partially visible objects: the local-feature-focus method ," *J. Robotics Research*, Vol. 1, No. 3, pp.57-82.

Brady, M., Ponce, J., Yuille, A., and Asada, H. (1985) "Describing surfaces", *Proc. 2nd International Symposium on Robotics Research*, H. Hanafusa and H. Inoue (eds.) MIT Press, Cambridge, MA.

Brou, P. (1983) "Using the Gaussian image to find the orientation of object", *The International Journal of Robotics Research*, Vol.3, No.4, pp.89-125.

Brooks, R.A. (1981) "Symbolic reasoning among 3-D models and 2-D images", *Artificial Intelligence*, Vol.17, Nos. 1-3, pp.285-348.

Brown, C. M. (1979) "Fast display of well-tessellated surface ", *Computer and Graphics*, Vol. 4, No. 2. pp. 77-85.

Chakravarty, I. and Freeman, H. (1982) "Characteristic views as a basis for three-dimensional object recognition", *Proc. The Society for Photo-Optical Instrumentation Engineers Conference on Robot Vision*, Vol. 336, SPIE, Bellingham, Wash., pp. 37-45.

do Carmo, M.P. (1976) *Differential Geometry of Curves and Surfaces*, Prentice-Hall, Englewood Cliffs, New Jersey.

Fukada, Y. (1984) "Recognition of structural industrial parts stacked in bin", *Robotica*, vol. 2.

Goad, C. (1983) "Special purpose automatic programming for 3D model-based vision", *Proc. Image Understanding Workshop*, pp.94-104.

Grimson, W. E. L. , and Lozano-Perez, T. (1984) "Model-based recognition and localization from sparse range or tactile data", *International Journal of Robotics Research*, Vol. 3, No. 3.

Herman, M. and Kanade, T. (1984) "The 3D MOSAIC scene understanding system: incremental reconstruction of 3D scene from complex images", *CMU CS Report*, CMU-CS-84-102.

Horn, B.K.P. (1979) "Sequins and Quills- Representations for surface topography", *MIT AI Memo 536*, MIT Artificial Intelligence Laboratory.

Horn, B.K.P. (1984) "Extended Gaussian images", *Proc. of the IEEE*, Vol.72, No.12, pp.1671-1686.

Horn, B.K.P. (1986) *Robot Vision* MIT Press Cambridge.

Horn, B.K.P. and Ikeuchi, K. (1984) "The Mechanical Manipulation of Randomly Oriented Parts", *Scientific American*, Vol.251, No.2, pp.100-111.

Ikeuchi, K. (1981) "Recognition of 3-D objects using the extended Gaussian image", *Proc. 7th International Joint Conference on Artificial Intelligence*, pp.595-600.

Ikeuchi, K. (1983) "Determining attitude of object from needle map using extended Gaussian image", *AI memo No. 714*, MIT Artificial Intelligence Laboratory, Cambridge, MA.

Ikeuchi, K., Horn, B. K. P., Nagata, S. , Callahan, T, and Feingold, O. (1984) "Picking up an object from a pile of objects", *1st International Symposium on Robotics Research*, M. Brady and R. Paul (eds.) MIT Press, Cambridge, MA.

Ikeuchi, K. (1985) "Region-based stereo on needle maps", *Proc. '85 International Conference on Advanced Robot*, pp.207-214, Robotics Society of Japan, Tokyo.

Ikeuchi, K., Nishihara, H. K., Horn, B. K. P., Sobalvarro, P., and Nagata, S. (1986) "Determining grasp points using photometric stereo and the PRISM binocular stereo system", *J. Robotics Research*, Vol. 5, No. 1, pp.46-65.

Kimura, F. and Hosaka, M. (1977)*Program Package GEOMAP Reference Manual*, Computer Vision Section, Electrotechnical Lab.

Koenderink, J.J., and Van Doorn, A. J. (1979) "Internal representation of solid shape with respect to vision", *Biological Cybernetics*, Vol. 32, No. 4, pp. 211-216.

Koshikawa, K. (1984) *SOLVER reference manual*, RM-85-33J, Computer Vision Section, Electrotechnical Lab. (in Japanese)

Koshikawa, K., and Shirai, Y. (1985) "A 3-D modeler for vision research", *Proc. '85 International Conference on Advanced Robot,* pp.185-190, Robotics Society of Japan.

Little, J.J. (1985) "Determining object attidute from extended Gaussian images", *Proc. of Intern. Joint Conf. on Artificial Intelligence",* pp.960-963.

Lozano-Perez, T. (1981) "Automatic Planning of Manipulator Transfer Movements", *IEEE Trans. Sys. Man. Cyb.*, Vol.SMC-11, No.10, pp.681-689.

Oshima, M. and Shirai, Y. (1985) "A model based vision for scenes with stacked polyhedra using 3D data ", *Proc. '85 International Conference on Advanced Robot,* pp.191-198, Robotics Society of Japan.

Perkins, W. A. (1977) "Model-based vision system for scene containing multiple parts", *Proc. 5th International Joint Conference on Artificial Intelligence*, pp.678-684.

Shafer, S.A. and Kanade T. (1983) "Using shadows in finding surface orientation", *Computer Vision, Graphics and Image Processing*, Vol.24, No.2, pp.182-199.

Smith, D. (1979) "Using enhanced spherical images", *AI Memo 451*, MIT Artificial Intelligence Laboratory.

Sugihara, K. (1979) " Automatic construction of junction dictionaries and their exploitation for analysis for range data", *Proc. 6th International Joint Conference on Artificial Intelligence*, pp.859-864.

Thorpe, C., and Shafer, S. (1983) "Topological correspondence in line drawings of multiple views of objects", *CMU-CS-83-113*, Dept. of Computer Science, Carnegie-Mellon Univ., Pittsburgh, Pa.

Tsuji, S. and Nakamura, A. (1975) "Recognition of an object in a stack of industrial parts", *Proc. 4th International Joint Conference on Artificial Intelligence*, pp.881-818.

Tsuji, S. and Matsumoto, F. (1977) "Detection of elliptic and linear edges by searching two parameter space", *Proc. 5th International Joint Conference on Artificial Intelligence*, pp.569-575.

Yachida, M. and Tsuji, S. (1975) "A machine learning capability", *Proc. 4th International Joint Conference on Artificial Intelligence*, pp.819-826.