International Journal of Computer Vision, 145–165 (1987)

Generating an Interpretation Tree from a CAD Model for 3D-Object Recognition in Bin-Picking Tasks

KATSUSHI IKEUCHI
Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA

This article describes a method to generate 3D-object recognition algorithms from a geometrical model for bin-picking tasks. Given a 3D solid model of an object, we first generate apparent shapes of an object under various viewer directions. Those apparent shapes are then classified into groups (representative attitudes) based on dominant visible faces and other features. Based on the grouping, recognition algorithms are generated in the form of an interpretation tree. The interpretation tree consists of two parts: the first part for classifying a target region in an image into one of the shape groups, and the second part for determining the precise attitude of the object within that group. We have developed a set of rules to find out what appropriate features are to be used in what order to generate an efficient and reliable interpretation tree. Features used in the interpretation tree include inertia of a region, relationship to the neighboring regions, position and orientation of edges, and extended Gaussian images.

This method has been applied in a task for bin-picking objects that include both planar and cylindrical surfaces. As sensory data, we have used surface orientations from photometric stereo, depth from binocular stereo using oriented-region matching, and edges from an intensity image.

1 Introduction

Sensory capabilities will extend the functional range of robots. Without sensing the outer world, robots can only repeat preprogrammed 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 noncontact 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.
3. Assembling something using the object.

This article describes 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 predetermined. Such a system needs the help of another machine or a human for feeding objects at a predetermined place in a predetermined attitude. Since this feeding job is tedious, it is quite unsuitable for a human being. Traditional mechanical feeding methods rely on a

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.

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 [1–10]. 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 have difficulty extracting the contour of a three-dimensional (3D) object from the image of a set of overlapping objects, which is typical in bin-picking.

Birk and others [11] highlight scenes to segment and determine the position and the orientation of an object in a bin. This system is limited to cylindrical workpieces with a metallic surface. Also, their vision system only determines two degrees out of three degrees of freedom in attitude.

We have presented techniques for using photometric stereo and an extended Gaussian image to determine object attitude [12–14]. 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 *uninterpretable* 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-achieving parts. The representation can serve as the starting point to the task-achieving module. Thus, the representation should be somehow a copy of the outer world and be in a convenient form to operate with it.

This report resolves these problems using a geometrical modeler. This system has the following characteristics:

1. An interpretation tree is precompiled from an object model so as to determine attitude by using the optimal features at each determining process.
2. 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.
3. The attitude and the position obtained are represented in the world in a geometrical modeler.

2 Deriving the Interpretation Tree

A 3D object varies its apparent shape depending on the viewer direction and rotation. Two classes of shape changes exist among these possible shape changes of a 3D object: a nonlinear shape change and a linear shape change. Figure 1a shows an example of a nonlinear shape change. Between these two shapes, two sets of visible faces are different. In this nonlinear shape change, topological relationships between faces are different from each other. Figure 1b shows an example of a linear shape change. Between these two

Fig. 1a. An example of nonlinear shape change.

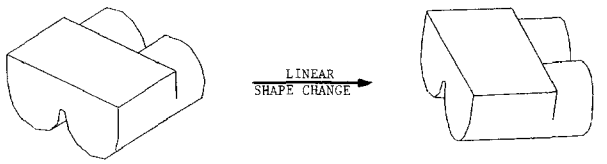

Fig. 1b. An example of linear shape change.

shapes, two sets of visible faces are the same. Only the shape of each face is skewed.

Different features are appropriate for resolving nonlinear shape changes and linear shape changes. Also the features necessary to resolve shape changes are different, depending on the class of shapes. Thus, it is desirable to precompile a geometrical model into an interpretation tree so that the most appropriate features among available features at each determination stage are used to resolve the nonlinear and linear shape changes.

Since there are two classes of shape changes, the interpretation tree also consists of two parts: resolving the nonlinear shape changes and resolving the linear shape changes. In order to derive the interpretation tree, we will follow the next four steps: the first three steps for the nonlinear shape change and the last step for the linear shape change:

1. Extract all possible types of nonlinear shape changes.
2. Derive classification branches of the nonlinear shape changes.
3. Determine features to be used at the branching nodes of the tree.
4. Determine features to be used to determine linear shape change.

2.1 Resolving Nonlinear Shape Change

2.1.1 Representative Attitude (Classification of Nonlinear Shape Change).

The nonlinear shape changes can be categorized with various clues. Some researchers explore this categorization with visible lines [20–22]; others explore this with visible vertices [23]. Occluding boundaries are also explored [24, 25].

This article explores this categorization using faces observable by photometric stereo [35, 41], because they are stable and contain rich geometrical properties. Photometric stereo can determine the surface orientation at the place where the three light sources project their light directly. This article categorizes the nonlinear shape changes based on this observable faces by photometric stereo.

In the following discussion, we use the term *visible* for the sake of simplicity. If we treat the term *visible* as detectable with photometric stereo, we can apply the same discussion to detectable relationship between the viewer direction and the surface orientation. Since the geometry between TV camera and the light sources is fixed in photometric stereo, the detectability of one face is determined by the relationship between the surface orientation of the face and the line of sight of the TV camera. Thus, we can regard the detectability of one face as the same as the visibility of the face, provided that detectable directions of one face become a cone whose center is the surface orientation, while visible directions of one face become a hemisphere.

The object attitude and the viewer configuration have three degrees of freedom. Note that we use the terms *object attitude* and *viewer configuration* interchangeably. Two degrees of freedom exist in the viewer direction—the direction of the line of sight has two degrees of freedom with respect to the object. The remaining freedom exists in the rotation around the line of sight.

Among these three degrees of freedom, some of the changes of the viewer direction cause the nonlinear shape change. On the other hand, changes of the viewer rotation around the line of the sight do not cause the nonlinear shape change; they only cause a linear shape change. Thus, categorization of nonlinear shape changes requires only exploring apparent shapes under possible viewer directions.

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

pose that

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

$(X_1, X_2, \ldots, X_n)$ denotes one label of an apparent shape based on the visible faces. Here, one face corresponds either to a planar surface or to a curved surface. Each viewer direction can be characterized with this label, which will be referred to as a *shape label.*

The set of viewer directions that have the same *shape label* becomes an *attitude group*, which is one equivalent class among nonlinear shape changes of the object. There are two ways to generate attitude groups: an analytic method and an exhaustive method. If the target object is a convex polyhedron, then the analytic method is easy. The face visibility is determined by the relationship between the viewer direction and the surface orientation of the face. Viewer directions have two degrees of freedom and can be described as a point on the Gaussian sphere at whose center a target object is located. In the meantime, surface orientations can be also represented as a point on the same Gaussian sphere. Then, the visible viewer directions of a face are limited by a circle on the Gaussian sphere. The circle center corresponds to the surface orientation of the face, and the radius of the circle is $\pi/2$. The inside area of the circle corresponds to the viewer directions which can observe the face. Drawing these visible circles on the Gaussian sphere, attitude groups can be determined from the combination of the circle covers on the sphere.

If the target object is nonconvex, then the visible circle is distorted due to self-occlusion and the analytic method becomes difficult. A curved surface also makes the analytic method difficult. Thus, in the general case, the exhaustive method is preferable. Essentially, the exhaustive method generates various apparent shapes of the object under various viewer directions, and then examines shape labels of the generated shapes in order to get the attitude groups.

The first task is to sample the Gaussian sphere evenly; a geodesic dome is used to tessellate the Gaussian sphere evenly [26]. Each tessellated triangle corresponds to a particular viewer direction. These sampled viewer directions exist evenly over the Gaussian sphere, and cover the whole sphere surface.

At each sampled viewer direction, an apparent shape of the object is generated using a geometrical modeler. Then, we can sample all possible apparent shapes evenly under all possible viewer directions. One observable shape gives one shape label $X_1, X_2, \ldots, X_n$. After obtaining all shape labels of all generated shapes, attitude groups are generated based on these shape labels so that shapes at each attitude group share the same shape label.

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 that 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 2 shows an example of this process. Figure 2a is a picture of an object. Figure 2b is a model synthesized using a geometrical modeler [15–19]. The Gaussian sphere to represent the possible viewer directions is tessellated into small triangles using the one-frequency dodecahedron shown in figure 2c. Apparent shapes are generated at the viewer directions corresponding to the centers of the triangles. Since the one-frequency dodecahedral geodesic dome has 60 triangles, 60 different shapes are generated as shown in figure 2d, where the faces enclosed with bold lines are observable by the photometric stereo. The observable face is the face where the three light sources project light directly. Note that, even though some faces are visible to humans, they cannot be detected by the photometric stereo because of the geometry of the light sources. Such faces are enclosed with thin lines. Figure 2e shows the larger eight faces used for the shape label among 12 component faces of the object. Note that face 1 and face 2 in figure 2e are treated as one continuous surface even though they are divided into small patches approximating cylindrical surfaces. The shapes in figure 2e are combined

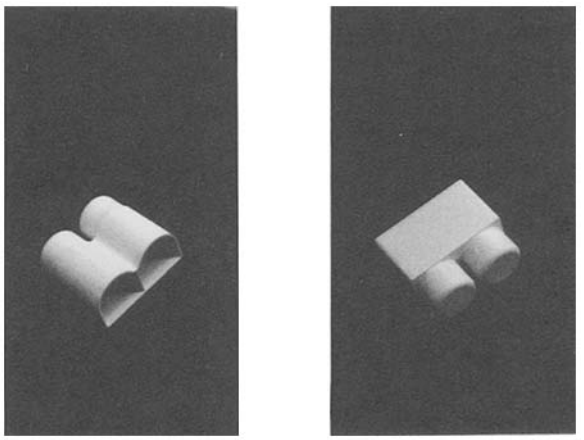

Fig. 2a. An object.

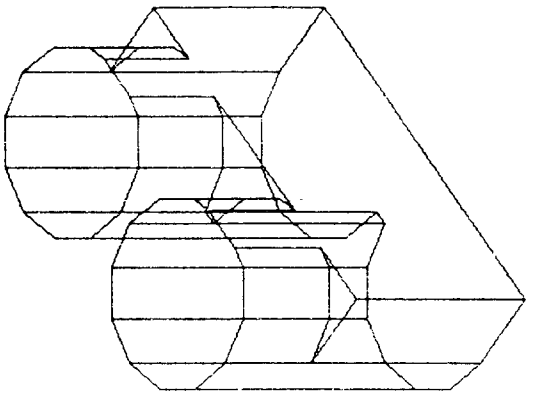

Fig. 2b. A synthesized model of the object in SOLVER [15].

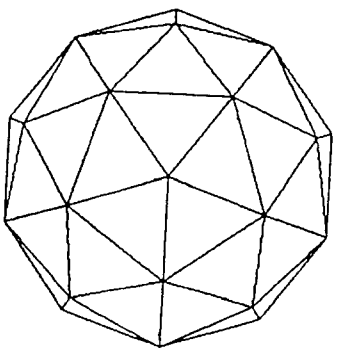

Fig. 2c. One-frequency dodecahedron.

into seven attitude groups as shown in figure 2f. Smaller regions under a certain threshold are regarded as nondetectable. Numbers of the visible faces as well as shape labels are printed under the group number in figure 2f. For example, in group 1, face 1, face 2, and face 3 are observable. Thus, the shape label of group 1 becomes 11100000. For face group 1 to group 5, five representative attitudes are generated as shown in figure 2g. Group 6 corresponds to a hole region of the object and such a steep convex area cannot be detected by photometric stereo. Group 7 has too small a visible area. Thus, no representative attitudes are generated from the groups 6 and 7.

2.1.2 Classifying into Representative Attitude. The previous section gives the final stage in shape classification: the deepest leaf of the interpretation tree. Yet we have not determined the branches of the interpretation tree from the root to the leaves. Therefore, the next step is to generate branches from the root to the attitude groups. Branches are generated using the shape label. 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 that generate the shape label. 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 $g_1 = \{f_1\}$, $g_2 = \{f_1, f_2\}, \ldots, g_n = \{f_1, f_2, \ldots, f_n\}$. The sequence of attitude groups given by this sequence of face groups generates a tree that contains only the representative attitude as the leaves.

Note that only valid attitude groups are generated among possible combinations of shape label at each face group. These valid attitude groups and valid shape labels are obtained from a geometrical modeler using the same method as generating the representative attitude described in the previous section. If we follow a brute-force method to generate a tree whose branches correspond to conditions, whether faces of the object are visible or not, without considering validity of each shape label, the method generates the tree of 2^n leaves. On the other hand, since our method generates only a vaild shape label of each face group based on the object, the method generates only leaves corresponding to the representative attitudes.

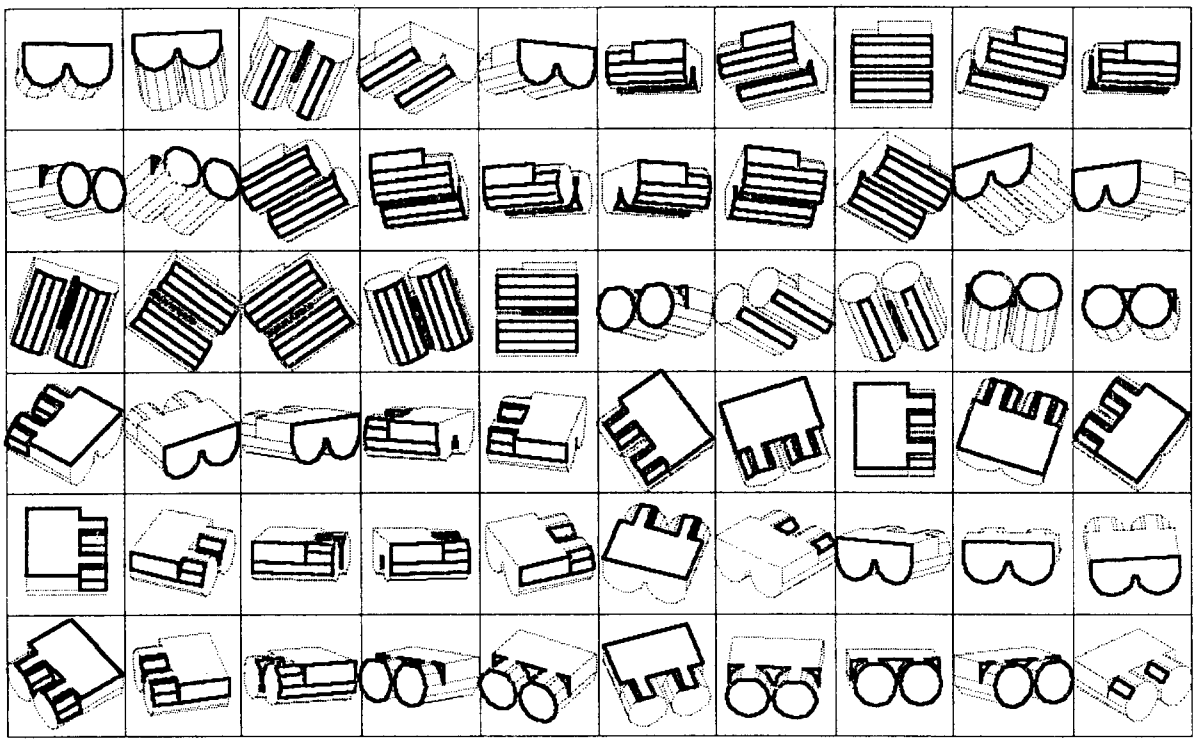

Fig. 2d. Sixty apparent shapes of the object.

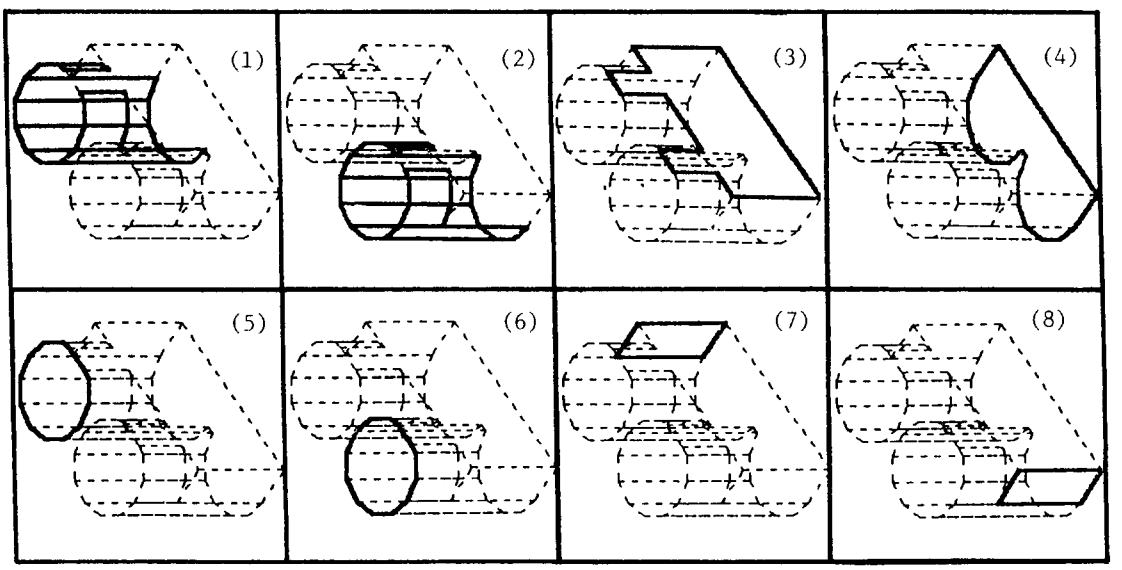

Fig. 2e. Eight identifying faces.

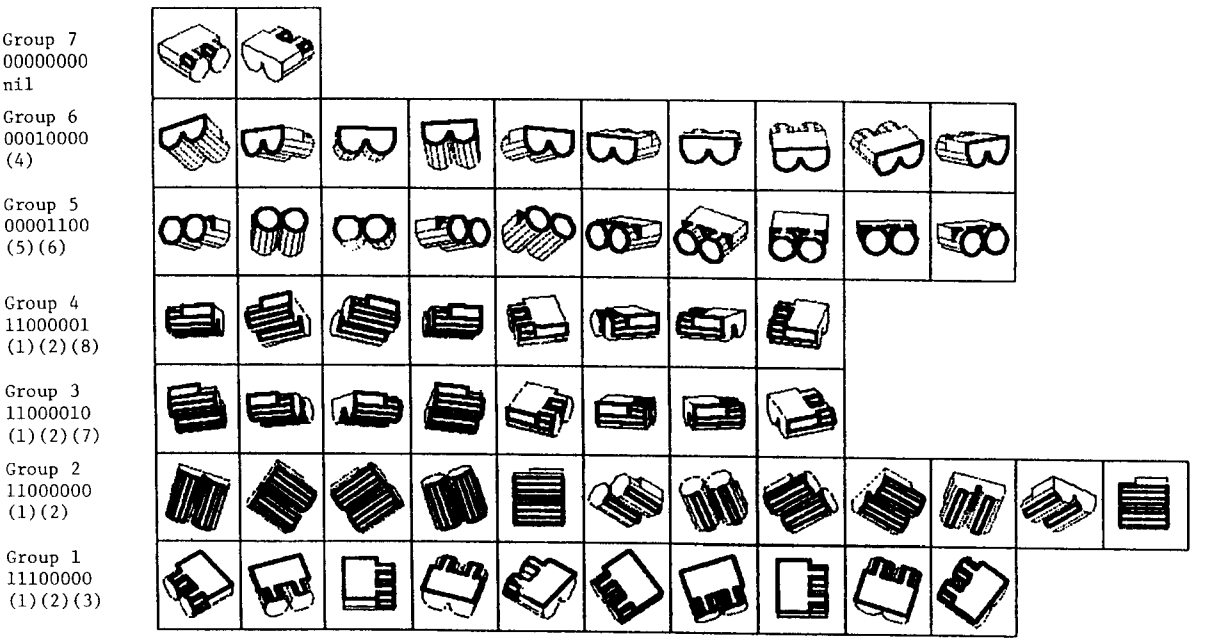

Fig. 2f. Seven attitude groups.

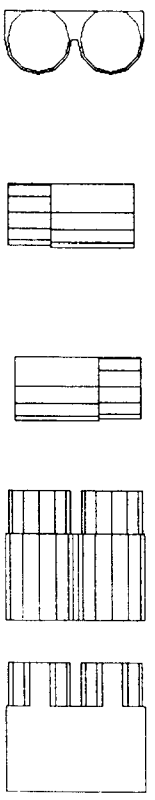

Fig. 2g. Five representative attitudes.

The fact that the sequence of attitude groups generates a tree that contains only the representative attitudes at the leaves can be proved inductively:

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

(Step n) Next we will consider the relationship between the attitude group from g_i and the attitude group from g_{i-1}. The attitude group of g_i is obtained by dividing the attitude group of g_{i-1} based on the visibility of the face, f_i. Thus, if the number of attitude group increases from $i - 1$ to i, new attitude groups at i come from only division of attitude groups at i; no new attitude groups at i come from combining one part of one attitude group at $i - 1$ and one part of the other attitude group at $i - 1$.

This division sequence generates a tree structure

that gradually reaches the final attitude groups. Since the representative attitudes are generated using the shape label of g_n, there is always one and only one leaf corresponding to one representative attitude in the final tree. This tree structure will be used as the structure of the interpretation tree.

Figure 3 shows the branches obtained from the object shown in figure 2. In the application, it often occurs that two faces have the same area. Since our method of tree generation is based on the area size of each face, the method becomes unstable at that branch. 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.

2.1.3 Work Models. The work models consist of physical face information. 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 a geometrical modeler 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

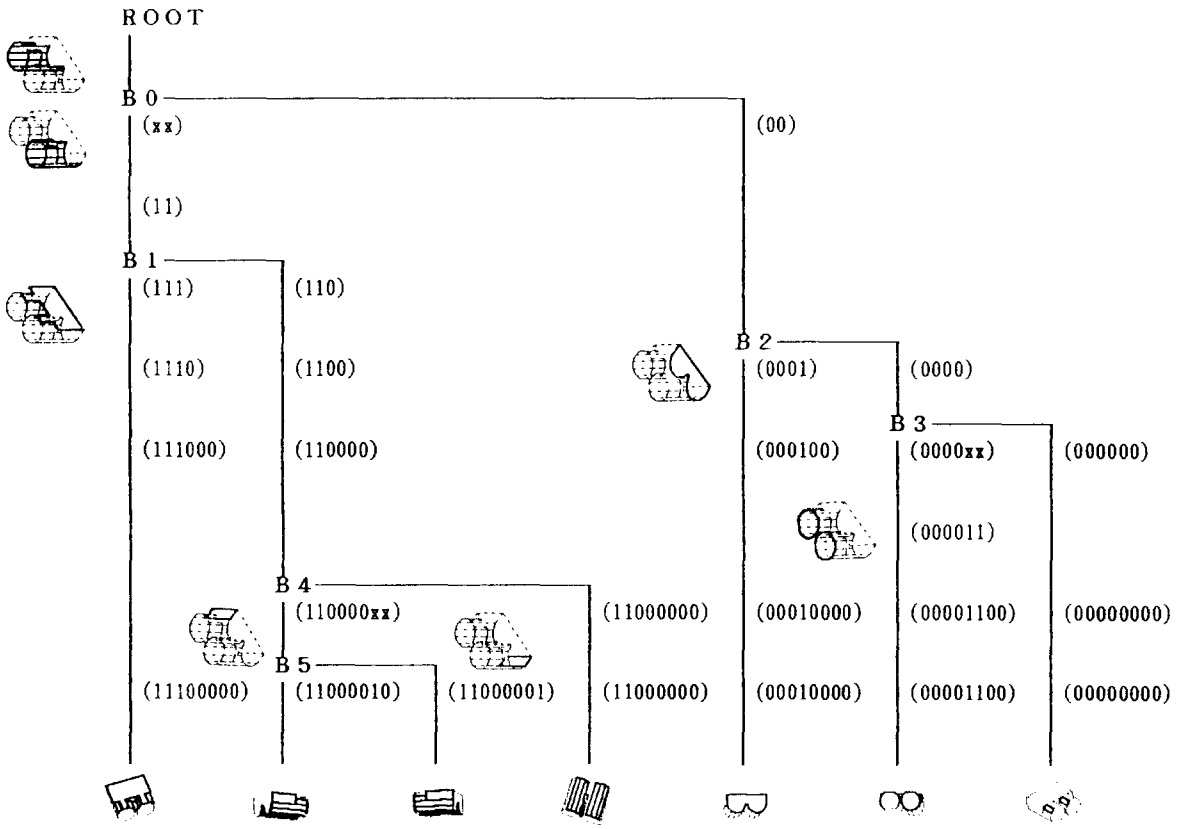

Fig. 3. Branches based on shape label.

the observed region information using an affine transform, where we assume the camera model as the orthographic projection. For example, when the surface orientation, the affine matrix, and the observed region shape 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 that represents one attitude group.

Let (p, q) be the surface orientation of one face.

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

gives the affine matrix to recover the original face information from the observed face information. Thus, given (p, q) from photometric stereo can derive T and transform apparent features to original features.

Our work models consist of original face inertia, original face relationship, original face shape, original edge relationship, extended Gaussian image, and surface characteristic distribution.

Original face inertia: The inertia moments of one face in the directions of least and most inertia direction. These inertia moments give the rough shape information of the face. See the appendix for more details.

Original face relationship: A nonconvex 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 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 [26].

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 [28–32]. 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 [33, 34]. The characteristics, defined in terms of the Gaussian curvature and the mean curvature, are independent of the viewer direction and the rotation. Distribution of the characteristics are used as work models. See the appendix for more details.

2.1.4 Classification Rules. This section gives rules to generate the classification part of the interpretation tree. At each branch, we examine whether one of the rules can discriminate between the attitude groups. If one of the rules can discriminate, the rule is registered at that branch. The decision of whether the rule can divide them or not is made by humans at present.

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.*

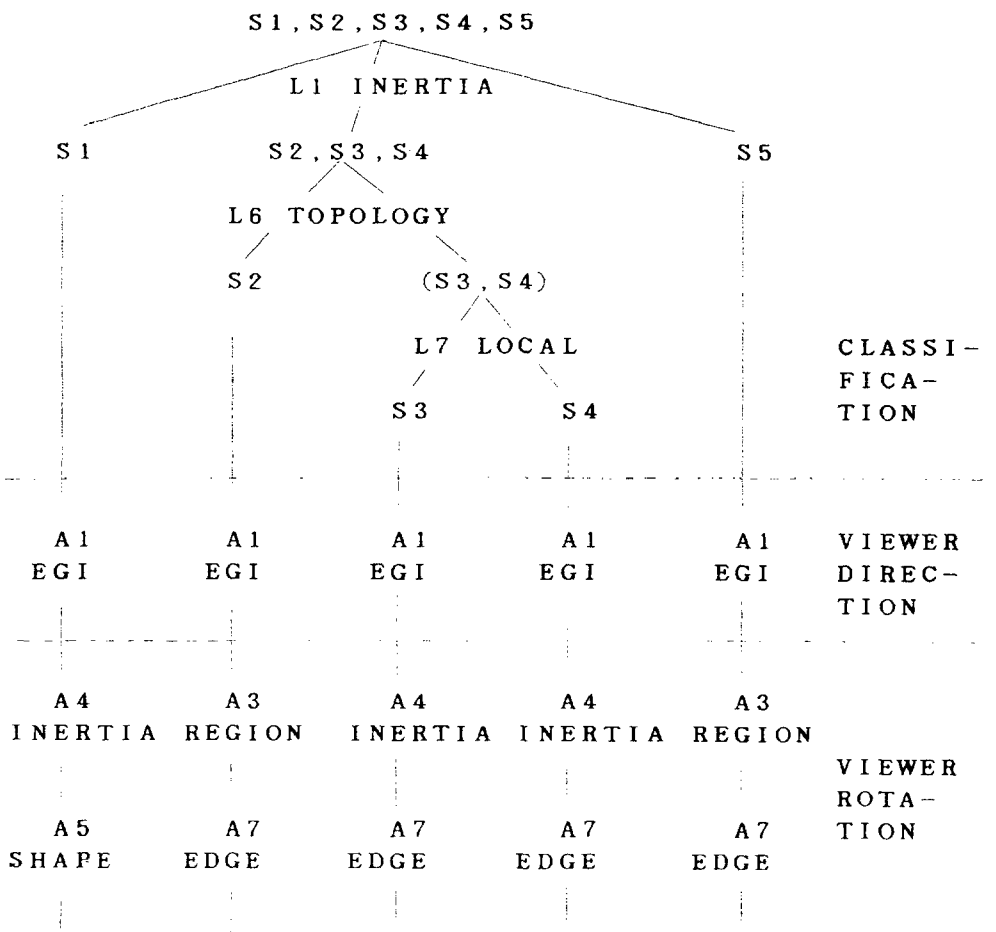

Fig. 4. The interpretation tree.

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.

2.1.5 Deriving the Classification Part of the Tree. The classification part of the interpretation tree, figure 4, is generated for the object shown in figure 2a. At the B0 branch, the rule L1 (original face inertia) can divide all the 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), or 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, and 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.

2.2 Resolving Linear Shape Change

2.2.1 Determination Rules. This section gives the rules to generate the part of the interpretation tree that determines the viewer direction and the rotation. Each rule that can reduce some of remaining freedom in the viewer direction and rotation will be adopted into the tree. The decision of whether the rule can reduce the freedom or not is made by humans 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.

2.2.2 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 [29] 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.

2.2.3 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 unqiuely 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.

2.2.4 Representative Attitude S3. Representative attitude S3 has one observable region that 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.

2.2.5 Representative Attitude S4. The features used to determine the viewer direction and the rotation are the same as those of the representative attitude A3.

2.2.6 Representative Attitude S5. Representative attitude S5 has two regions observed separately that 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 4 shows the interpretation tree obtained.

3 Applying the Interpretation Tree

3.1 Attitude Determination by the Interpretation Tree

The system can use three kinds of maps: edge maps, needle maps, and one depth map. Three edge 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 that are generated by a dual photometric stereo system [35]. 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.

3.2 Case 1: Attitude Group 1

Figure 5 shows one of the input scenes, where the white arrow indicates the highest region. From

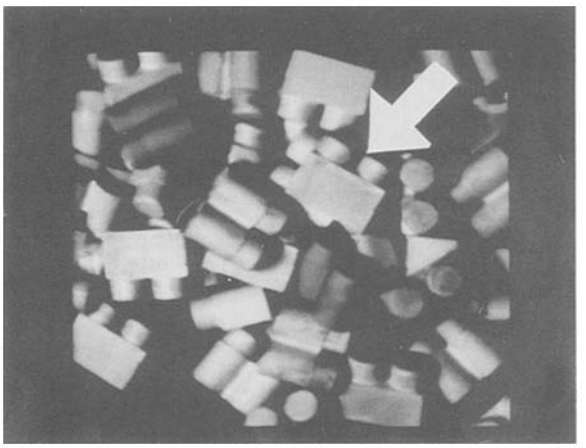

Fig. 5a. Input scene. The *white arrow* indicates the highest region.

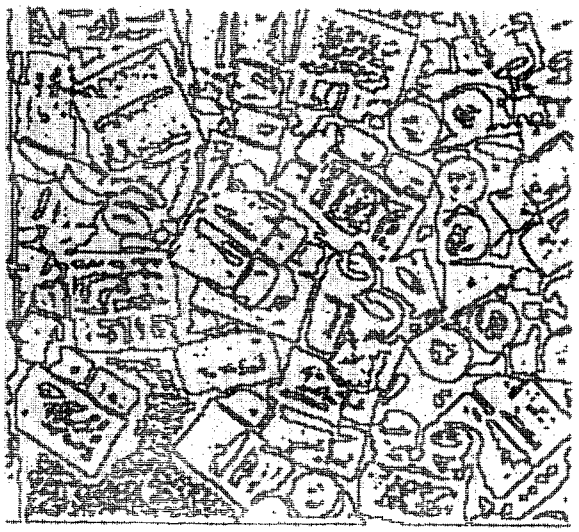

Fig. 5b. The edge map obtained from the scene.

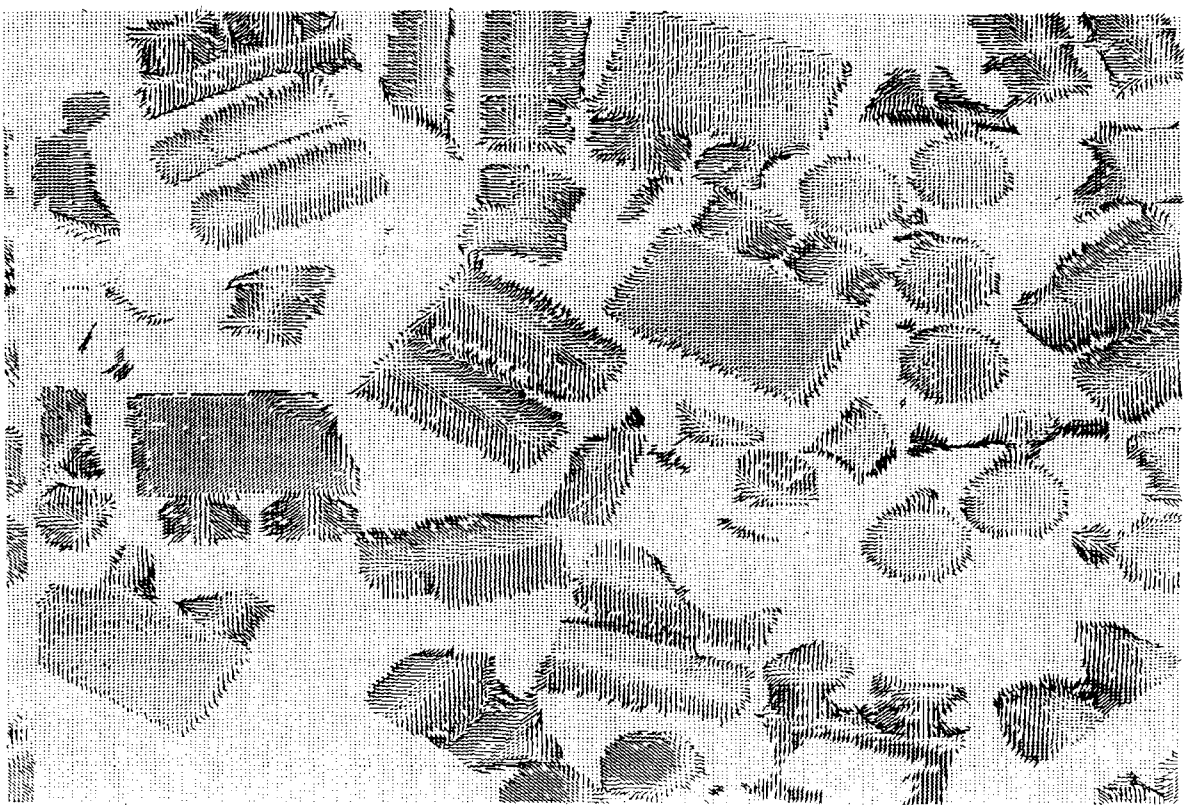

Fig. 5c. The needle map obtained by the photometric stereo system.

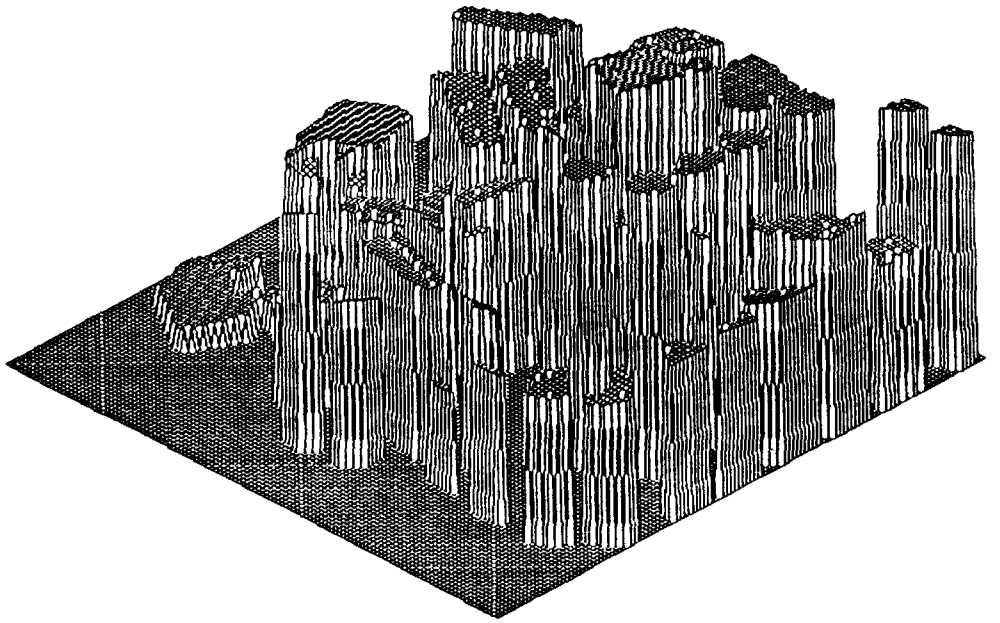

Fig. 5d. The depth map obtained by the dual photometric stereo system [35].

this scene, the edge map shown in figure 5b is obtained. The photometric stereo system gives the needle map shown in figure 5c. Further, the depth map shown in figure 5d is obtained by the dual photometric stereo system.

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 that has been converted from the needle map to have 1 at the places where the surface orientation is determined, and to have 0 at the places

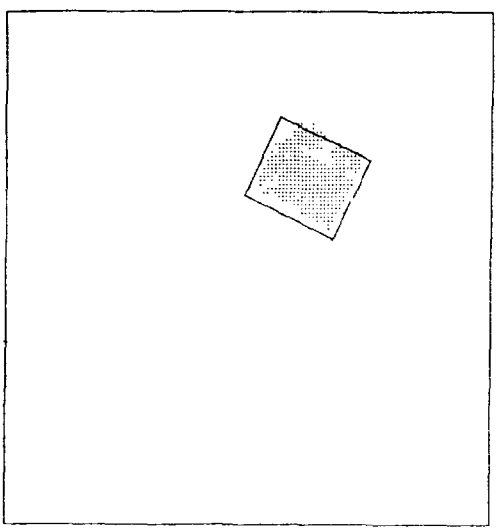

Fig. 6a. The target region and the original face inertia.

Fig. 6b. The original face shape recovered by the affine transformation. The shape is represented using 2D WTS.

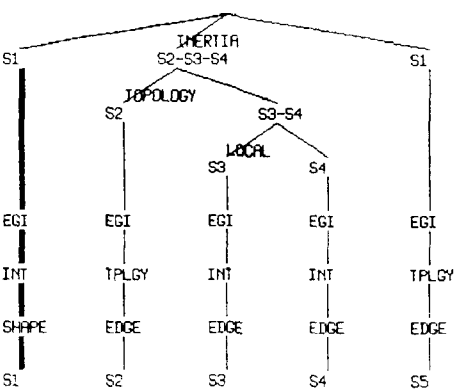

Fig. 6c. The decision path in the interpretation tree.

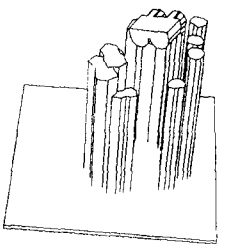

Fig. 6d. The obtained position, the obtained attitude, and the neighboring regions.

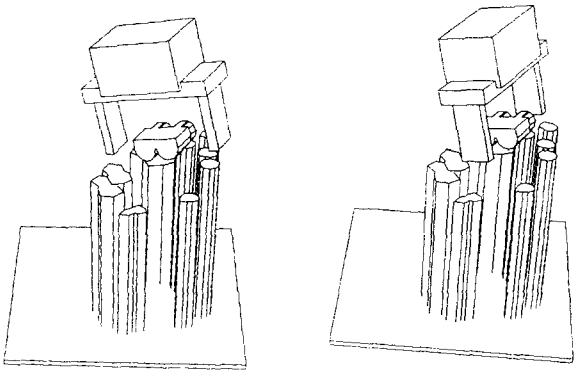

Fig. 6e. Two collision-free configurations.

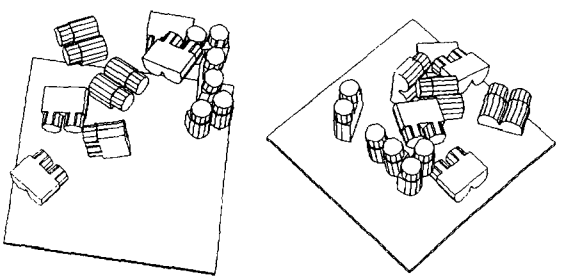

Fig. 6f. Scene description.

where the surface orientation is not determined. Then, the affine matrix is obtained from the surface orientation distribution over the region. Finally, the interpretation tree can determine the inertia moment of the original face using the affine matrix and the region distribution. Figure 6a shows the region distribution and the square that 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 6b shows the original face shape 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. Figure 6c shows the decision flow on the interpretation tree.

A geometrical modeler 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 that 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 a geometrical modeler as shown in figure 6d. By using this representation, we can calculate collision-free configurations as shown in figure 6e. In the meantime, if we repeat the recognition-and-representation loop, the system finally obtains the representation shown in figure 6f.

3.3 Case 2: Atttitude Group 2

Figure 7a 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 binary map con-

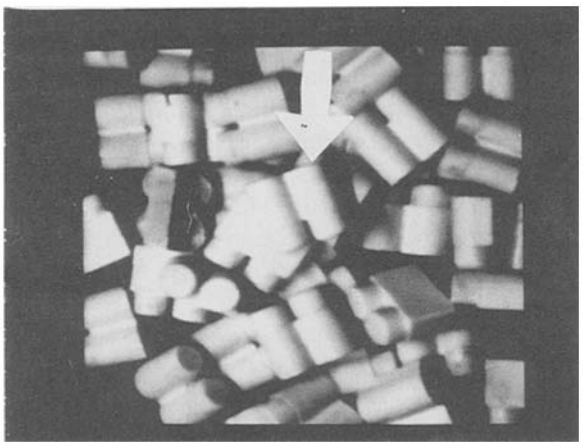

Fig. 7a. Input scene. The *white arrow* indicates the highest region.

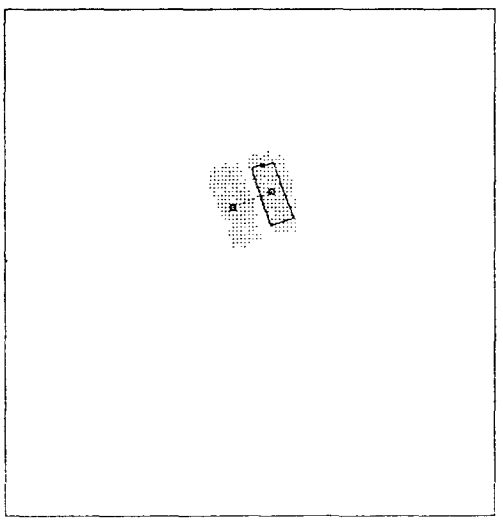

Fig. 7b. The target region and its brother region found by the algorithm.

verted from the needle map and the affine matrix obtained from the needle map over the target region. Figure 7b shows the square that 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 the 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

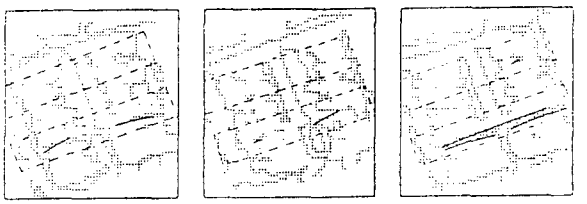

Fig. 7c. Obtained edges. The interpretation tree only examines the existence of the edge distribution whose direction agrees with the edge direction under one of the two possible rotations, at the place where one of the two rotations is supposed to make the edge distribution. The *dotted lines* indicate 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.

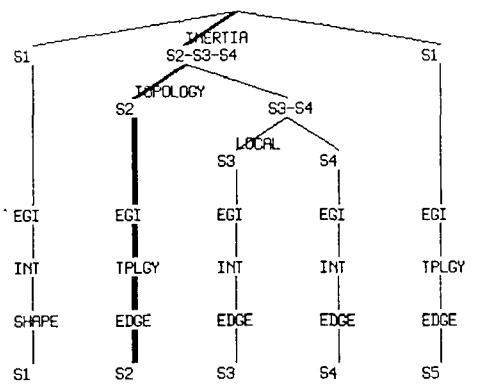

Fig. 7d. The decision flow on the interpretation tree.

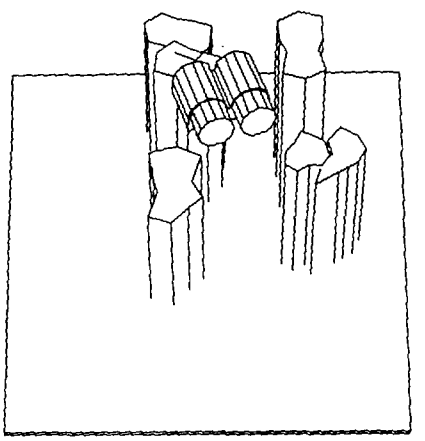

Fig. 7e. Obtained description.

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 7b, 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 7c, 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 7c 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 7d. Figure 7e shows the object attitude obtained by this process.

3.4 Case 3: Attitude Group 4

Figure 8a 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 8b shows the target region and the obtained moment-com-

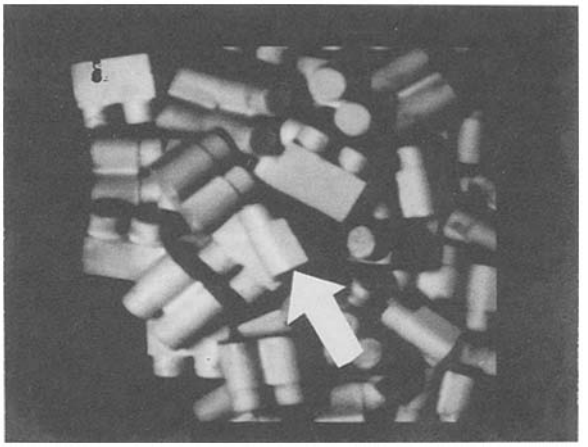

Fig. 8a. Input scene. The *white arrow* indicates the highest region.

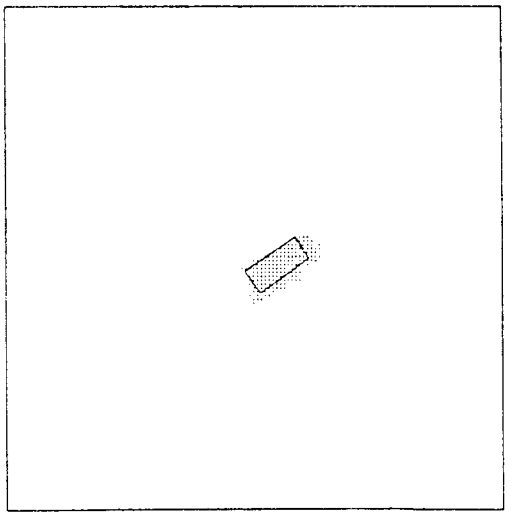

Fig. 8b. The target region and its original face inertia.

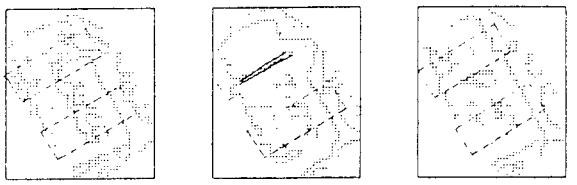

Fig. 8c. The edge distribution. The *dotted lines* indicate output from an edge operator. The *broken lines* indicate search areas predicted from the model. The *solid lines* indicates the edges that corresponds to the model.

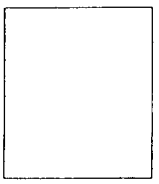

Fig. 8d. No surface characteristic distribution agree with the distributions of the representative attitude 3.

Fig. 8e. The characteristic distribution that 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 distribution corresponds to the representative attitude 4.

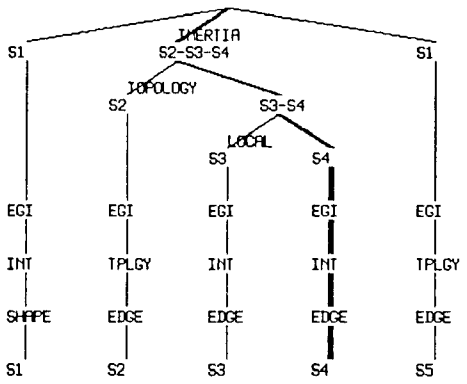

Fig. 8f. The decision flow on the interpretation tree.

patible 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 8c 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.

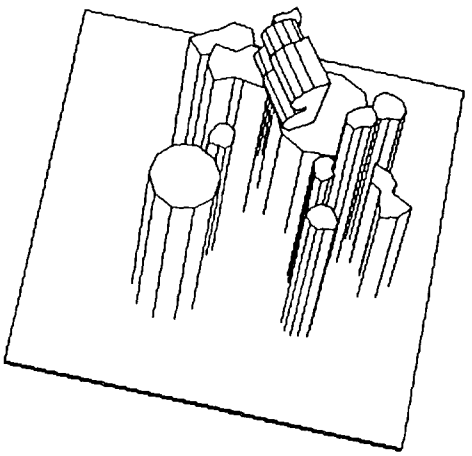

Fig. 8g. Obtained description.

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 8c. Figure 8d shows the surface characteristic distribution that agrees with the distribution of the representative attitude S3. Note that, since no distributions agree with the observed distributions, 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 8e shows the characteristic distribution that 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 8f shows the decision flow on the interpretation tree. Using the object position and attitude, the object is represented in the world model in a geometrical modeler shown in Figure 8g.

4 Concluding Remarks

This article describe a vision system to localize an object by an interpretation tree. This system has the following characteristics:

1. Representative attitudes are derived from a geometrical modeler, automatically.
2. The interpretation tree controls the localization process to use the most appropriate features at each stage of the localization.
3. The obtained attitude and position are represented in the world model in a geometrical modeler for further use.

This article assumes that the low-level operations are reliable, and does not emphasize backtracking. This assumption works well in our situation because

1. The interpretation tree only analyzes the highest region, which is usually not occluded and exhibits all information necessary to be recognized.
2. The interpretation tree only uses the most reliable features at each matching stage.
3. The interpretation tree also contains some of the verification process and returns the target region as unrecognized. Thus, if the interpretation failed to verify the target region, the region is discarded and the second highest region is given to the interpretation tree by a higher-level flow controller. This iteration is repeated until one of the regions passes the examination.

However, an active backtracking schema would be necessary to apply this method to the analysis of occluded objects and to increase the efficiency of the interpretation process. Certainly, the next step is to explore how to include backtracking control in the interpretation tree.

This article develops a flexible interpretation by an interpretation tree using multiple 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 faults, we have to select one appropriate feature among many available features in each processing stage. This article proposes using the interpretation tree for this purpose. This flexible interpretation matching should be further explored. Right now, the choice of discriminators used at nodes of the interpretation tree is made by "hand." In order to choose discriminators automatically, it is necessary to measure the uncertainty of each discriminator at each stage. This direction should be 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 article 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 images and generalized cylinders [36–38]. However, since the representation is needed for manipulation by other modules such as planning and navigation, the representation must be easy to manipulate [39]. This article proposes representing the outer world in the CAD model, because a CAD representation is an easy basis for achieving further tasks. Certainly there are many path-finding programs that start from the polyhedral representations [40]. How to express the outer world in such a representation should be explored more.

5 Appendix: Work Model

5.1 Original Face Inertia

One work model is the original face inertia. The original face inertia gives the rough shape in-

formation 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,

$$I_{xx} = \int m(x', y')dx'dx'$$
$$I_{xy} = \int m(x', y')dx'dy'$$
$$I_{yy} = \int m(x', y')dy'dy'$$

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:

$$I_{\max} = (I_{xx} + I_{yy} + \sqrt{(I_{xx} + I_{yy})^2 - 4(I_{xx}I_{yy} - I_{xy}I_{xy})})/2$$
$$\alpha = (\tan^{-1}\{(2I_{xy})/(I_{xx} - I_{yy})\})/2$$

5.2 *Surface Characteristic Distribution*

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

$$E = (1 + p^2)$$
$$F = pq$$
$$G = (1 + q^2)$$

The second fundamental forms e, f, g are

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

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

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

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.

Acknowledgments

The author thanks Takeo Kanade, Steve A. Shafer, Richard Mickelsen, and the members of Image-Understanding-System Group of Carnegie Mellon University for their valuable comments and discussions.

References

1. S. Tsuji and A. Nakamura, "Recognition of an object in a stack of industrial parts," in PROC. 4TH INT. JOINT CONF. ARTIF. INTELL., 1975, pp. 881–818.
2. S. Tsuji and F. Matsumoto, "Detection of elliptic and linear edges by searching two parameter space," in PROC. 5TH INT. JOINT CONF. ARTIF. INTELL., 1977, pp. 569–575.
3. M. Yachida and S. Tsuji, "A machine learning capability," in PROC. 4TH INT. JOINT CONF. ARTIF. INTELL., 1975, pp. 819–826.
4. M.L. Baird, "Image segmentation technique for locating automotive parts on belt conveyers," in PROC. 5TH INT. CONF. ARTIF. INTELL., 1977, pp. 694–695.
5. W.A. Perkins, "Model-based vision system for scene containing multiple parts," in PROC. 5TH INT. JOINT CONF. ARTIF. INTELL., 1977, pp. 678–684.
6. R. Bolles and R.A. Cain, "Recognizing and locating partially visible objects: the local-feature-focus method," INT. J. ROBOTICS RES. vol. 1, no. 3, pp. 57–82, 1982.
7. Y. Fukada, "Recognition of structural industrial parts stacked in bin." ROBOTICA vol. 2, 147–154, 1984.

8. C. Goad, "Special purpose automatic programming for 3D model-based vision," in PROC. IMAGE UNDERSTANDING WORKSHOP, 1983, pp. 94–104.
9. N. Ayache, B. Faverjon, J. Boissonnat, and B. Bollack, "Automatic handling of overlapping workpieces," in PROC. INT. CONF. PATTERN RECOGNITION 84, 1984, pp. 837–839.
10. W.E.L. Grimson and T. Lozano-Pérez, "Model-based recognition and localization from sparse range or tactile data," INT. J. ROBOTICS RES. vol. 3, no. 3, pp. 3–35, 1984.
11. J.R. Birk, R.B. Kelly, and H.A.S. Martines, "An orienting robot for feeding workpieces stored in bins," IEEE TRANS. SMC vol. SMC-11, no. 2, pp. 151–160, 1981.
12. K. Ikeuchi, B.K.P. Horn, S. Nagata, T. Callahan, and O. Feingold, "Picking up an object from a pile of objects," in PROC. FIRST INT. SYMP. ROBOTICS RES., M. Brady and R. Paul (eds.). M.I.T. Press: Cambridge, MA, 1984.
13. K. Ikeuchi, H.K. Nishihara, B.K.P. Horn, P. Sobalvarro, and S. Nagata, "Determining grasp points using photometric stereo and the PRISM binocular stereo system," INT. J. ROBOTICS RES. vol. 5, no. 1, pp. 46–65, 1986.
14. B.K.P. Horn and K. Ikeuchi, "The mechanical manipulation of randomly oriented parts," SCIENTIFIC AMERICAN vol. 251, no. 2, pp. 100–111, 1984.
15. K. Koshikawa, SOLVER REFERENCE MANUAL, RM-85-33J [in Japanese]. Computer Vision Section, Electrotechnical Lab., 1984.
16. K. Koshikawa and Y. Shirai, "A 3-D modeler for vision research", in PROC. '85 INT. CONF. ADVANCED ROBOT, Robotics Society of Japan, 1985, pp. 185–190.
17. M. Oshima and Y. Shirai, "A model based vision for scenes with stacked polyhedra using 3D data," in PROC. '85 INT. CONF. ADVANCED ROBOT, Robotics Society of Japan, 1985, pp. 191–198.
18. F. Kimura and M. Hosaka, PROGRAM PACKAGE GEOMAP REFERENCE MANUAL, Computer Vision Section, Electrotechnical Lab., 1977.
19. B.G. Baumgart, "Winged edge polyhedron representation," Stanford Univ. A.I. Lab., STAN-CS-320, 1972.
20. I. Chakravarty and H. Freeman, "Characteristic views as a basis for three-dimensional object recognition," in PROC. SOC PHOTO-OPTICAL INSTRUMENTATION ENGINEERS CONF ROBOT VISION, vol. 336. SPIE: Bellingham, WA, 1982, pp. 37–45.
21. J.J. Koenderink and A.J. Van Doorn, "Internal representation of solid shape with respect to vision," BIOL. CYBERNETICS, vol. 32, no. 4, pp. 211–216, 1979.
22. K. Sugihara, "Automatic construction of junction dictionaries and their exploitation for analysis for range data," in PROC. 6TH INT. JOINT CONF. ARTIF. INTELL., 1979, pp. 859–864.
23. C. Thorpe and S. Shafer, "Correspondence in Line Drawings of Multiple Views," in PROC. 8TH INT. JOINT CONF. ARTIF. INTELL., 1983, pp. 959–965.
24. M. Herman, "Matching three-dimensional symbolic description obtained from multiple views," in PROC. IEEE COMPUT. SOC. CONF. COMPUT. VISION AND PATTERN RECOGNITION, San Francisco, June 1985.
25. M. Hebert and T. Kanade, "The 3D profile method for object recognition," in PROC. IEEE COMPUT. SOC. CONF. COMPUT. VISION AND PATTERN RECOGNITION, San Francisco, June 1985.
26. C.M. Brown, "Fast display of well-tessellated surface," COMPUT. GRAPHICS. vol. 4, no. 2, pp. 77–85, 1979.
27. D. Smith, "Using enhanced spherical images," M.I.T. Artif. Intell. Lab., Cambridge, MA, A.I. Memo. 451, 1979.
28. B.K.P. Horn, "Extended Gaussian images," PROC. IEEE, vol. 72, no. 12, pp. 1671–1686, 1984.
29. K. Ikeuchi, "Recognition of 3-D objects using the extended Gaussian image," in PROC. 7TH INT. JOINT CONF. ARTIF. INTELL., 1981, pp. 595–600.
30. K. Ikeuchi, "Determining attitude of object from needle map using extended Gaussian image," M.I.T. Artif. Intell. Lab., Cambridge, MA, A.I. Memo 714, 1983.
31. P. Brou, "Using the Gaussian image to find the orientation of object," INT. J. ROBOTICS RES. vol. 3, no. 4, pp. 89–125, 1983.
32. J.J. Little, "Determining object attitude from extended Gaussian images," in PROC. 9TH INT. JOINT CONF. ARTIF. INTELL., 1985, pp. 960–963.
33. M. Brady, J. Ponce, A. Yuille, and H. Asada, "Describing surfaces", in PROC. 2ND INTERNATIONAL SYMPOSIUM ON ROBOTICS RESEARCH, H. Hanafusa and H. Inoue (eds) M.I.T. Press: Cambridge, MA, 1985.
34. P.J. Besl and R.C. Jain, "Intrinsic and extrinsic surface characteristics," in PROC. COMPUTER VISION AND PATTERN RECOGNITION CONFERENCE. IEEE: San Francisco, 1985, pp. 226–233.
35. K. Ikeuchi, "Determining a depth map using a dual photometric stereo," INT. J. ROBOTICS RES. vol. 6, no. 1, pp. 15–31, 1987.
36. T.O. Binford, "Visual perception by computer," in PROC. IEEE SYSTEMS SCIENCE CYBERNETICS CONF., 1971.
37. R.A. Brooks, "Symbolic reasoning among 3-D models and 2-D images," ARTIF. INTELL. vol. 17, nos. 1–3, pp. 285–348, 1981.
38. S.A. Shafer and T. Kanade, "The theory of straight homogeneous generalized cylinder and a taxonomy of generalized cylinders," Carnegie Mellon University Computer Science Department, Pittsburgh, PA, CMU-CS-83-105, 1983.
39. M. Herman, and T. Kanade, "The 3D MOSAIC scene understanding system: incremental reconstruction of 3D scene from complex images," CMU-CS Report, CMU-CS-84-102, 1984.
40. T. Lozano-Pérez, "Automatic planning of manipulator transfer movements," IEEE TRANS. SYS. MAN. CYBERNETICS vol. SMC-I, no. 10, pp. 681–689, 1981.
41. R.J. Woodham, "Reflectance map techniques for analyzing surface defects in metal castings," M.I.T. Artif. Intell. Lab., Cambridge, MA, A.I.-TR-457, 1978.