Received 19 May 86
AAAI entry May 86

Representing Object Configurations for Bin-picking Tasks Guided by an Interpretation Tree Derived from a CAD Model

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 has two major modes; deriving a interpretation tree and applying the interpretation tree. When deriving an interpretation tree, a geometrical modeler is used to generate various apparent shapes of an object under various viewer directions. Secondly, shape groups are generated from these apparent shapes based on the observable faces under each viewer direction. Thirdly, representative attitudes are determined at each shape groups. Fourthly, various features are extracted using the geometrical modeler at each representative attitude. These features contain the original face inertia, the original face shape, the original edge distribution, the surface characteristic distribution which can be recovered from the observed shape with the affine matrix. Here, the affine matrix can be obtained from the surface orientation distribution. Fifthly, an interpretation tree is generated to classify an observed shape into one of the representative attitudes, and to determine the attitude using generation rules. Here, the interpretation tree determines maps, areas, and features for attitude determination. When applying the interpretation tree, the system uses the depth map, the needle map, and the edge maps, which are obtained by the pair of the photometric stereo systems. The interpretation tree determines the attidue and the position of the object observed as a target region in the image. The attitude and the position obtained is represented in the world in the geometrical modeler.

INTRODUCTION

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

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

A manipulator without vision can only pick up an object whose position and attitude is pre-determined. Such a system needs the help of another machine and/or human for feeding objects at a pre-determined place in a pre-determined attitude. Since this feeding job is tedious, the job is quite unsuitable for a human being. Some researchers have aimed at solving this feeding problem by introducing mechanical vibration methods. These methods may cause defects in objects due to collisions. Other researches have proposed a

This paper proposes a method to solve this problem by visual guidance. This method has following characteristics:

1. The system uses a depth map, a needle map, and edge maps which are obtained by the pair of the photometric stereo (Ikeuchi:1985).

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

3. The interpretation tree classifies one target region into a representative attitude and then to determine 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 (Koshikawa:1984).

DERIVING INTERPRETATION TREE

REPRESENTATIVE ATTITUDE

A three-dimensional object varies its apparent shapes depending on the viewer direction and viewer rotation. These apparent shapes of an object fall into groups such that each group consists of roughly same shapes. Some researchers explore this characterization with visible lines (Koenderink & Van Doorn:1979, Sugihara:1979, Chakravarty & Freeman:1982). This paper explore the characterization with observable faces under photometric stereo. Here, "roughly same" means that almost same faces can be observed in the almost same condition under the photometric stereo.

The number of observable regions of a non-convex object depends on the viewer direction under the photometric stereo. 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 of detectable regions which are isolated with each other by shadowing and/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 the object attitude has 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 affects the number.)

Each viewer direction can be characterized with visible faces from the direction. Let us suppose that

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

$(X_1, X_2, \ldots, X_n)$ 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. If this label has the same combination of 1 as another shape, then these two shapes are contained in the same shape group. In other words, the viewer directions which have the same visible face label becomes a shape group.

One representative attitude will be selected from each shape group. Each shape group is represented as one representative attitude. Namely, the viewer directions over one particular range is represented by one representative attitude. Usually, the viewer direction which gives the largest sectional area among a shape group is determined as the viewer direction for the representative attitude. The viewer rotation for the representative attitude is determined so as to have the maximum inertia direction which agrees with the x axis on the image plane.

Fig. 1 shows an example of this process. Fig.1a is a picture of an object. Fig.1b is a model synthesized using SOLVER. Fig.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. The sixty directions come from the face centers of 1 frequency dodecahedral geodesic dome which can devide the spatial angle into sixty directions almost evenly. These shapes are fallen into 7 shape groups as shown in Fig.1d. Through face group 1 to group 5, five representative attitudes are generated as shown in Fig.1e. Since group 6 corresponds to a hole region of the object and group 7 has too small visible area, no representative attitudes are generated from the group 6 and group 7.

WORK MODEL

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, 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 modeling process, and are derived from needle maps and/or edge maps in 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 affine transform. For example, when the surface orientation, the affine matrix, and the observed region shape is known, the original face shapes can ber recovered from the skewed region shape with the affine transform. Only one face information is necessary at each shape group in which

detectable faces are the same and they are reachable with each other by the affine transformation. The work models are, thus, generated at each representative attitude which represent one shape group.

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 the pixel where the surface orientation can be obtained there, and *0* at the pixel where the surface orientation cannot be obtained there. The original face inertia can be obtained from the obtained binary map and the affine matrix.

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 with 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. In case that 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:1979).

Original Edge Distribution

Some of the prominent edge information is also used. In some case 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 with the starting position and the ending position. These positions are denoted relatively with the mass

center of the face and the maximum inertia direction. In application, this 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 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. 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 as the extended Gaussian image (EGI) of the object (Ikeuchi:1981).

Surface Characteristic Distribution

The surface characteristic distribution is available from the surface orientation distribution. A surface patch has a characteristic such as planer surface, cylindrical surface, elliptic surface, or hyperbolic surface. The first and the second fundamental forms can be obtained from the surface orientation and its derivatives. The first and second fundamental forms give the gaussian curvature and the mean curvature. The characteristic can be defined with the gaussian curvature and the mean curvature (do Carmo:1976). The surface characteristics are independent on the viewer direction and the rotation.

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 comparing the most appropriate feature of feature in work models with the feature obtained from the observed data over one target region at each step. 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. The second part determines the viewer direction of the region 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 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

This section gives rules to generate the classification part of the interpretation tree. At each rule, if the rule can divide a group of the representative attitudes into smaller groups, the rule will be adopted into the tree. The decision whether the rule can divide them or not is made by human 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.*

If one observed shape of an object cannot be classify into one particular representative attitude with these rules, it means that the object is observed as the same number of regions whose area, inertia moment, edge distribution, and the surface characteristic distribution are completely same in two different attitude. We treat such kinds of objects over the scope of this technique presented in this paper.

The classification part of the interpretation tree is generated to the object using these rules.

Determining Viewer Direction and Viewer Rotation

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 parts of remaining freedoms in the viewer direction and the rotation, the rule will be adopted into the tree.

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 rotation of original face shape.*

A5: *Using the position of the surface characteristics distribution.*

A6: *Using the position of the edges.*

A7: *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 as the same number of regions whose area, inertia moment, edge distribution, and the surface characteristic distribution are completely same in two different attitude. We treat such kinds of objects over the scope of this technique presented in this paper.

The viewer direction and the rotation is determined at each representative attitude using the most effective featur at each step. The most powerful rule for determining the viewer direction and the 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. Using these rules, the determination part of the interpretation tree is derived. Fig.2 shows the interpretation tree derived for the object using these rules.

APPLYING INTERPRETATION TREE

This experiment checks the ability of the interpretation tree. A strategy is necessary to select the region interpreted for the bin picking tasks. The highest region is selected as the target region. In this scene the region which will be classified into S2 representative attitude is highest.

The system can use three kinds of maps: edge maps, needle maps, and one depth map. Fig.3 shows one of the edge maps (Fig.3b), one of the needle maps (Fig.3c), and one depth map (Fig.3d) to be used by the interpretation tree. The system also generate a binary map from the needle map which has 1 at the place where the surface orientation is determined and has 0 at the place where the surface orientation cannot be determined. Isolated regions are obtained from this binary map using the labelling operation.

Fig.4 shows the determination process of a region to be classified into S2 representative attitude, while a bold line in Fig.4a indicates the trace of the determining process of this region in the interpretation tree. The arrow in Fig.3a indicates the target region given to the interpretation tree. The interpretation tree calculates the original face inertia of the region from the binary map converted from the needle map and the affine matrix obtained from the needle map over the target region. Fig.4b 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 examining whether a brother region exists to have the same inertia direction and the inertia value around the target region or not. The interpretation tree tries to find such a brother region, and succeeds to find the brother region as shown in Fig.4b, where the target region and the brother region is connected with a solid line in Fig.4b. From this evidence, the interpretation tree determines that the target region and the brother region come from the same object and belong to S2 representative attitude (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 rotation, at the place where one of the two possible rotations are 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 Fig.4c, the dot 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 Fig.4c 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 object is represented in the world model in SOLVER using the object position and the attitude. The neighboring regions around the target region and the brother region are expressed as dodeca prisms as before. The final representation in SOLVER is shown as in Fig.4d. Fig.4e shows the all regions obtained successfully by the interpretation tree.

CONCLUSION

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

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 localization,

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

for further use.

This paper develops a flexible interpretation by an interpretation tree using multiple sensory inputs. Recent work in image understanding has let to techniques for computing surface orientation and/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 explored further more.

A geometrical modeler is used to the recognition problem. Models from a geometrical modeler possesses rich geometrical features. Unfortunately however, the distance between the rich information and the information from the observed data is far. This paper uses the work model and the representative attitude to interface them guided by the interpretation tree. The effort is required to explore more convenient forms and method to connect them as well as to develop method of automatic generation and aquisition of the interpretation tree from CAD models.

The task of a vision system is to generate a description of the outer world. Some of the representation uses symbolic representation, others uses mathematics representations such as EGI and GC. However, since the representation is needed to manipulate it by other modules such as planning and navigation, the representation is easy to manipulate. This paper proposes to represent the outer world in the CAD model, because a CAD model is easy to achieve further task from that representation. Certainly there are many path finding programs to start from the polyhedral representations. How to express the outer world in such a representation should be explored more.

ACKNOWLEDGMENT

This research is done in part at the Electrotechincal Laboratory, MITI and in part at Department of Computer Science, Carnegie-Mellon University. Discussions with Nakashima, Tomura, Ogata of ETL, Kanade of CMU are helpful.

REFERENCES

Bolles, R. and Cain, R. A. (1982:) Recognizing and locating partially visible objects: the local-feature-focus method , *J. Robotics Research*, 1: 3, 57-82.

Brown, C. M. (1979:) Fast display of well-tessellated surface , *Computer and Graphics*, 4: 2, 77-85.

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

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*, 336, SPIE, Bellingham, Wash., 37-45.

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

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

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

Ikeuchi, K. (1986:) Deriving Interpretation Tree from a CAD Model for Bin-Picking Tasks, *CSD-technical paper*, Dept. of Computer Science, Carnegie-Mellon Univ., Pittsburgh, Pa., (in preparation).

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*, 5: 2.

Koenderink, J.J., and Van Doorn, A. J. (1979:) Internal representation of solid shape with respect to vision, *Biological Cybernetics*, 32: 4, 211-216.

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

Sugihara, K. (1979:) Automatic construction of junction dictionaries and their

SOLVER

exploitation for analysis for range data, *Proc. 6th International Joint Conference on Artificial Intelligence*, 859-864.

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

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

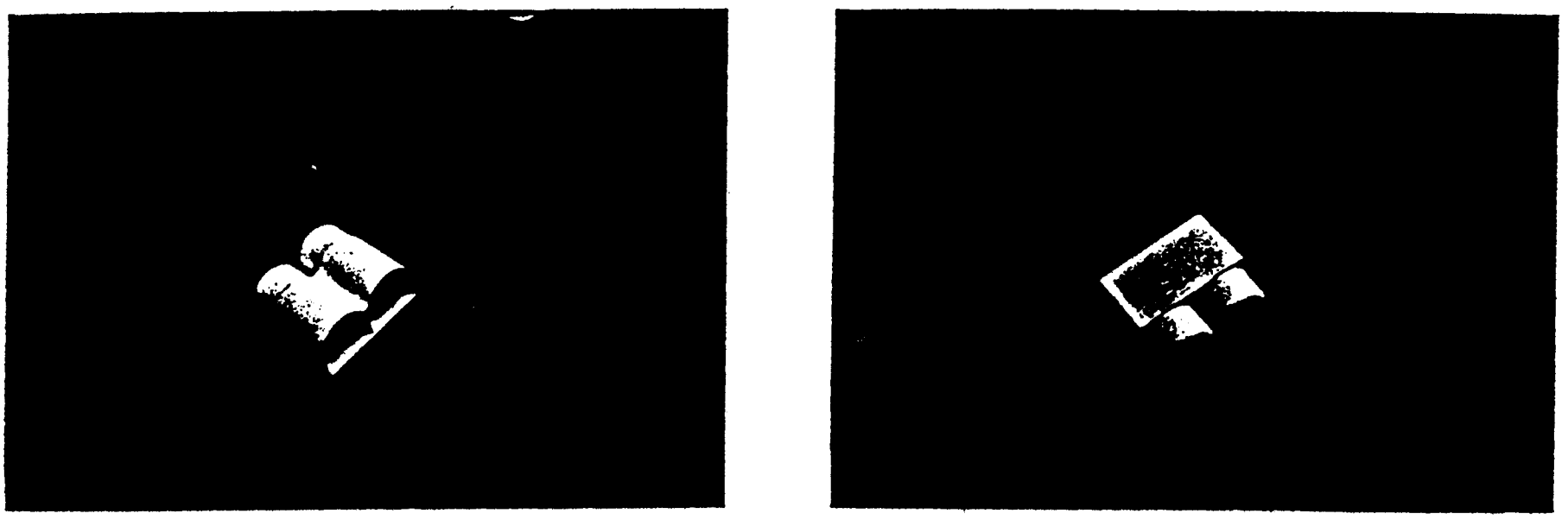


Fig 1a The Object

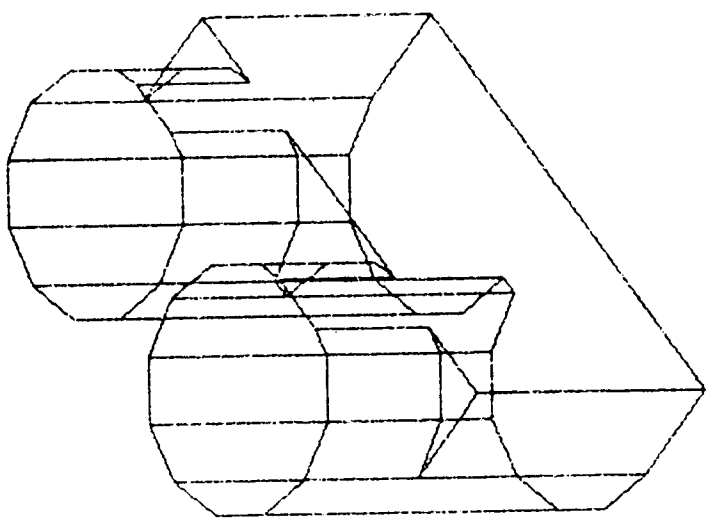


Fig 1b

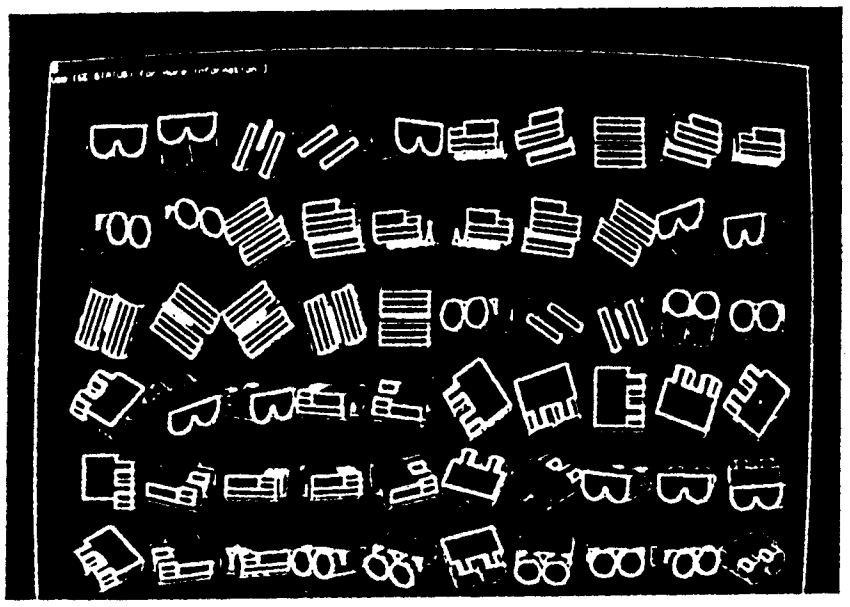


Fig 1 c
60 attitudes of the
object

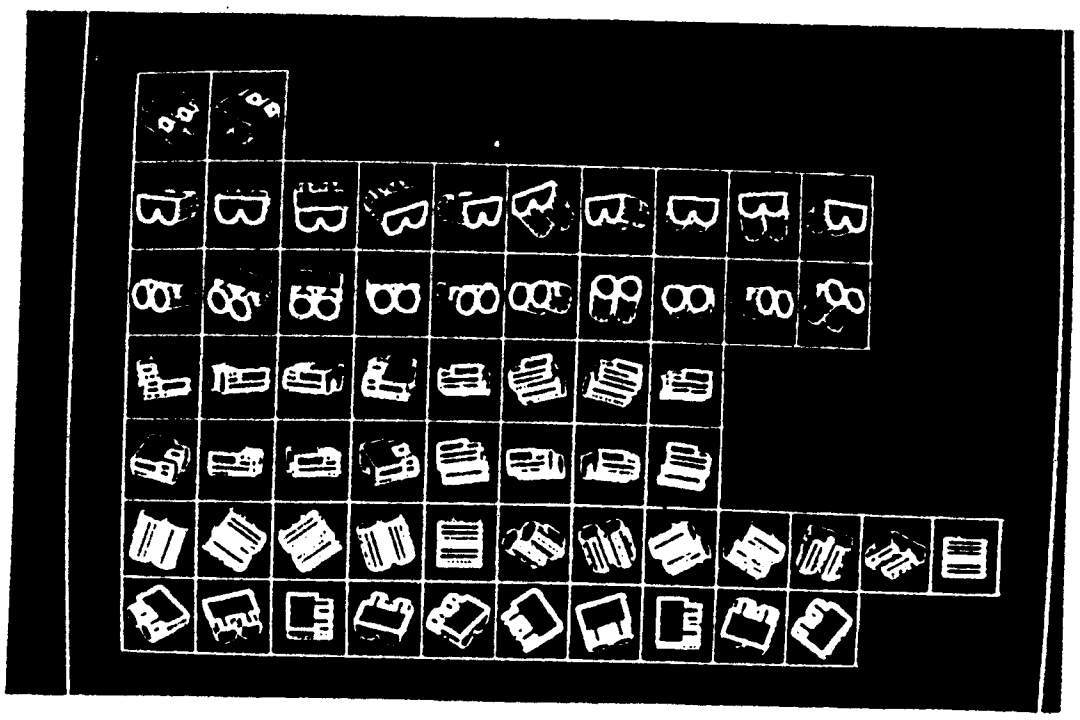


Fig 1 d
7 attitude groups

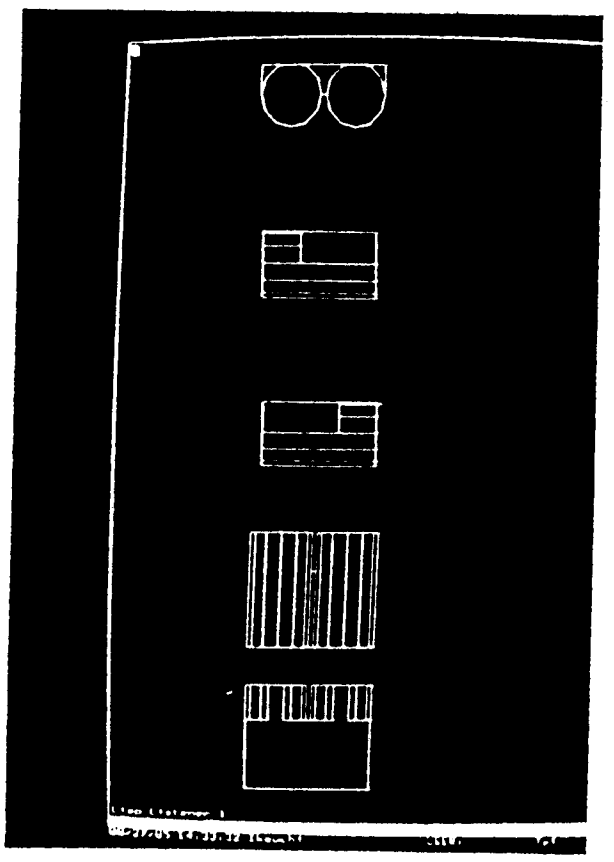


Fig 1 e
5 representative attitudes

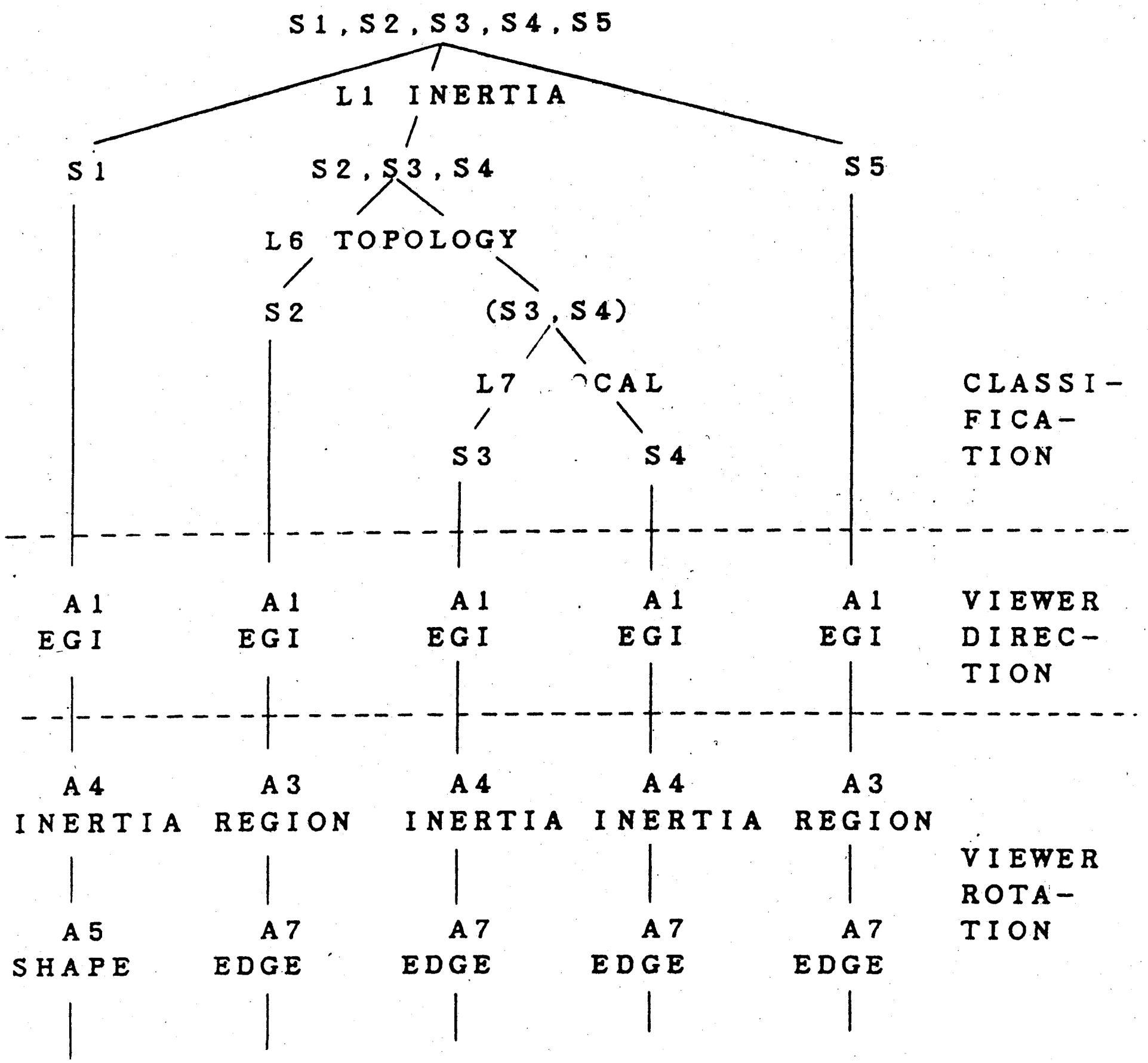
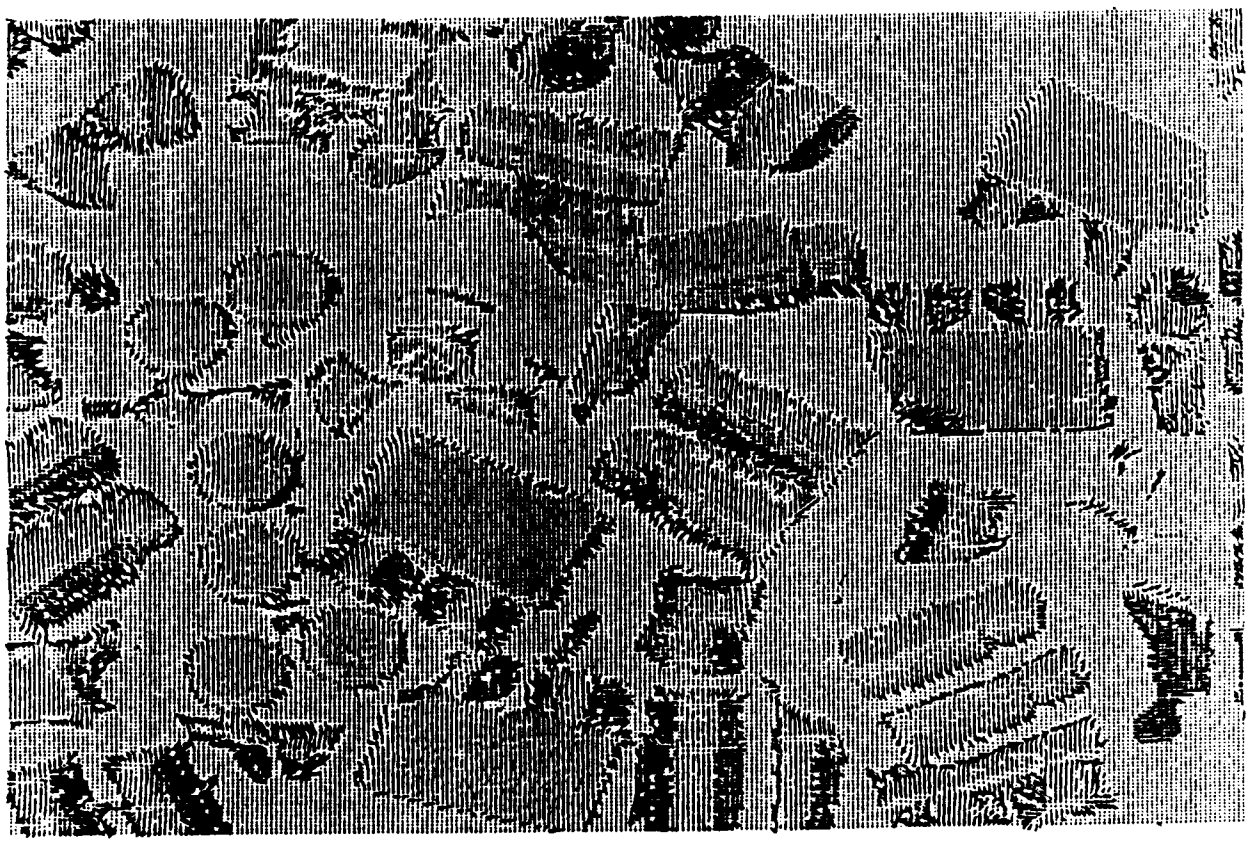
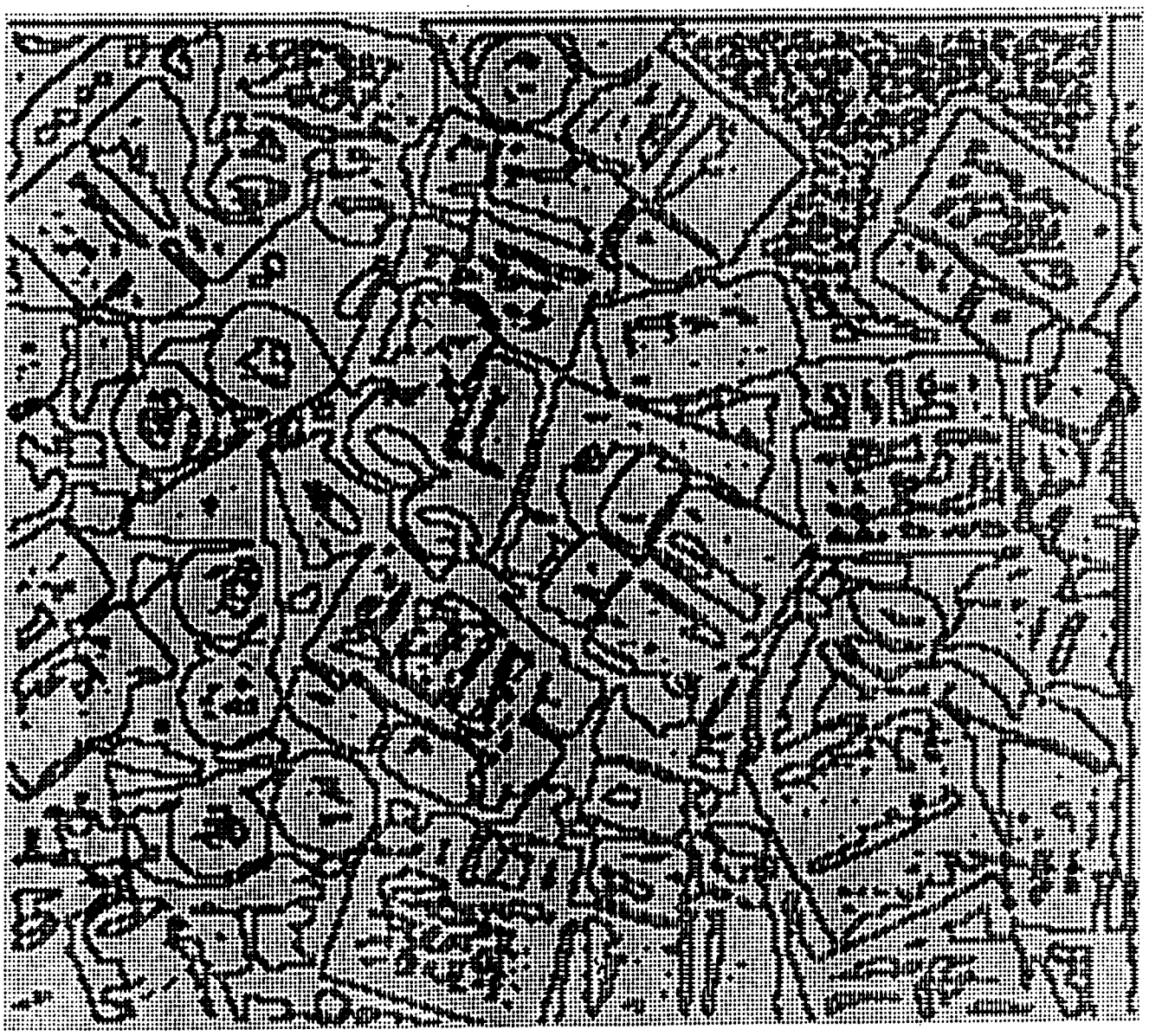


Fig 2.

The interpretation tree derived using the rules

Fig 3a input scene.





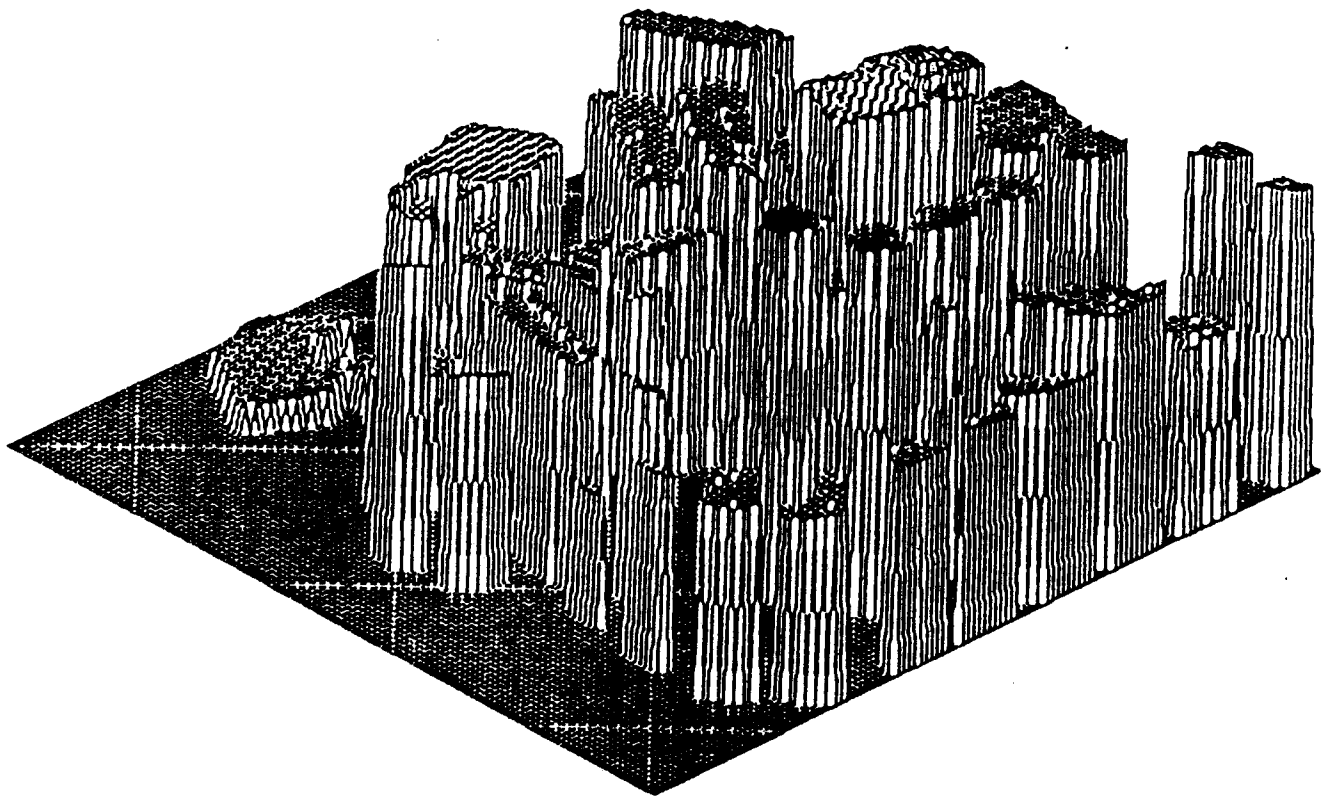


Fig 3d Depth map

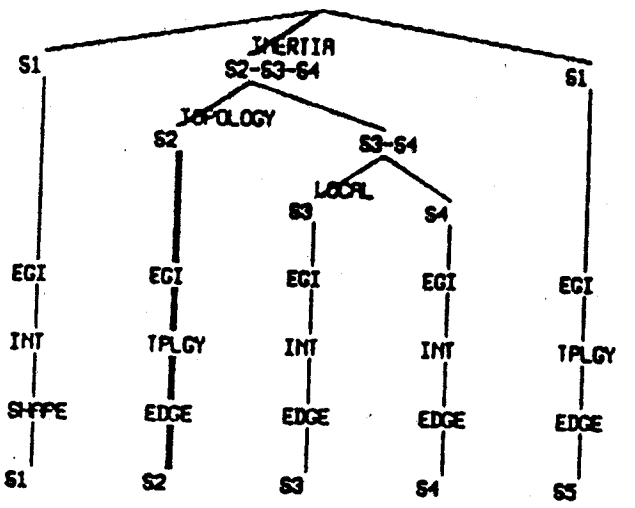


Fig 4a decision path

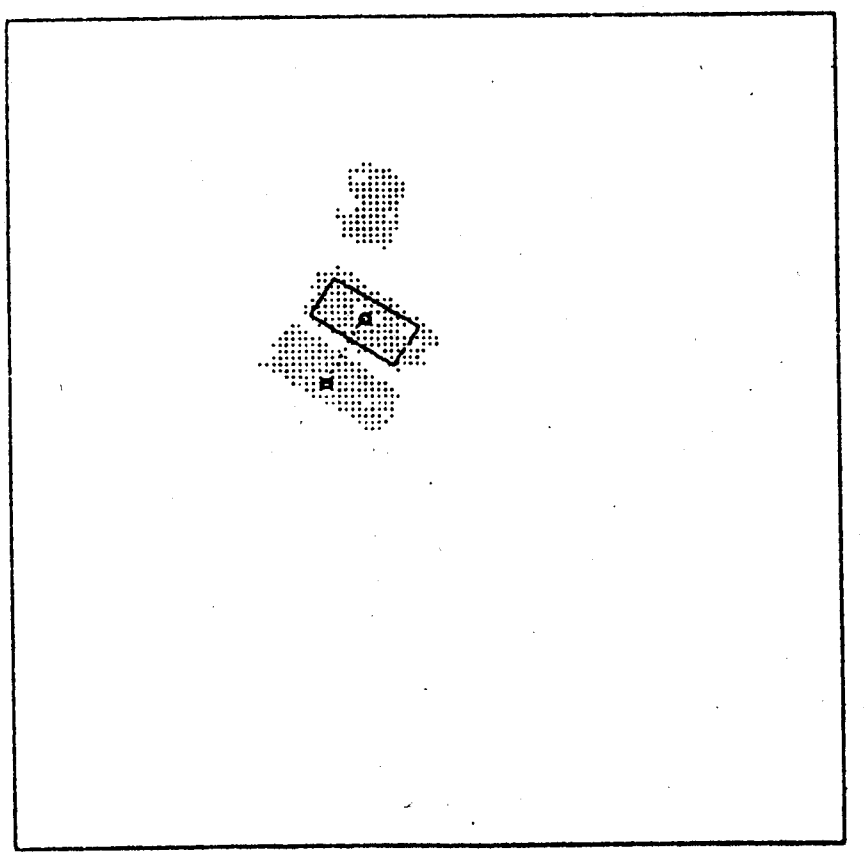


Fig 4b target region and another region

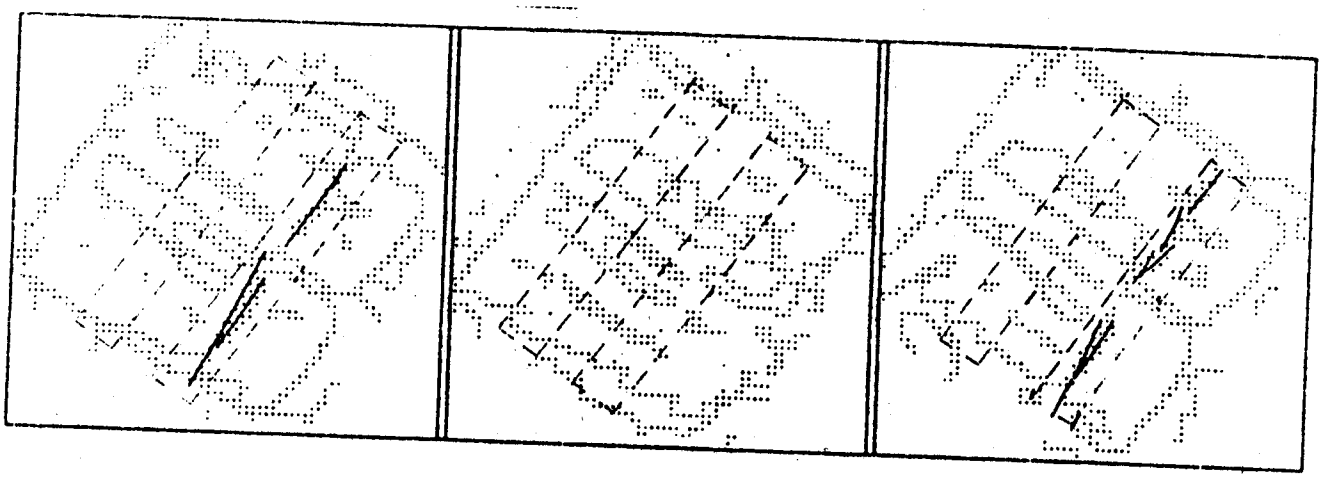


Fig 1 c edge search

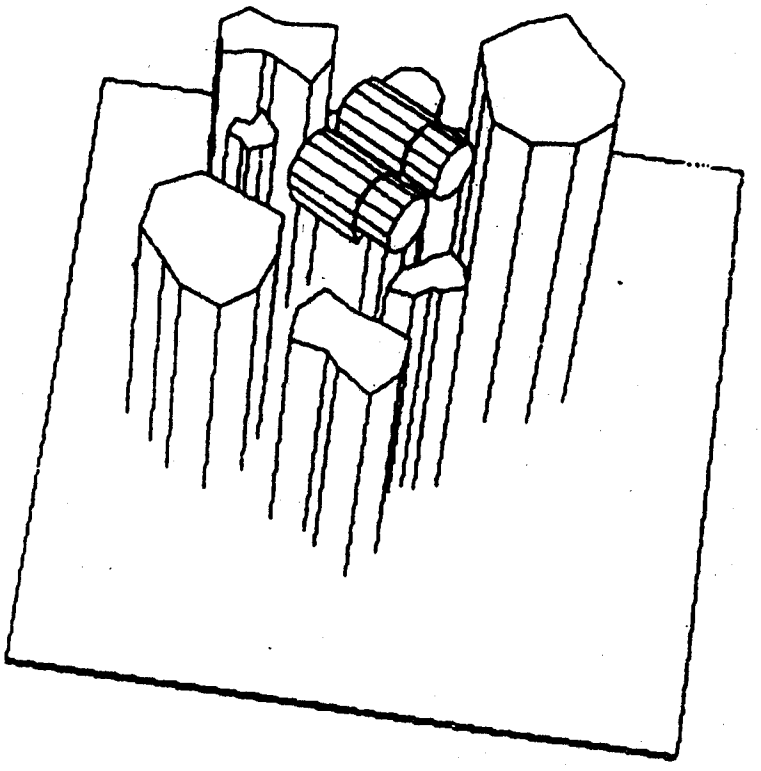


Fig 4 d configuration determined

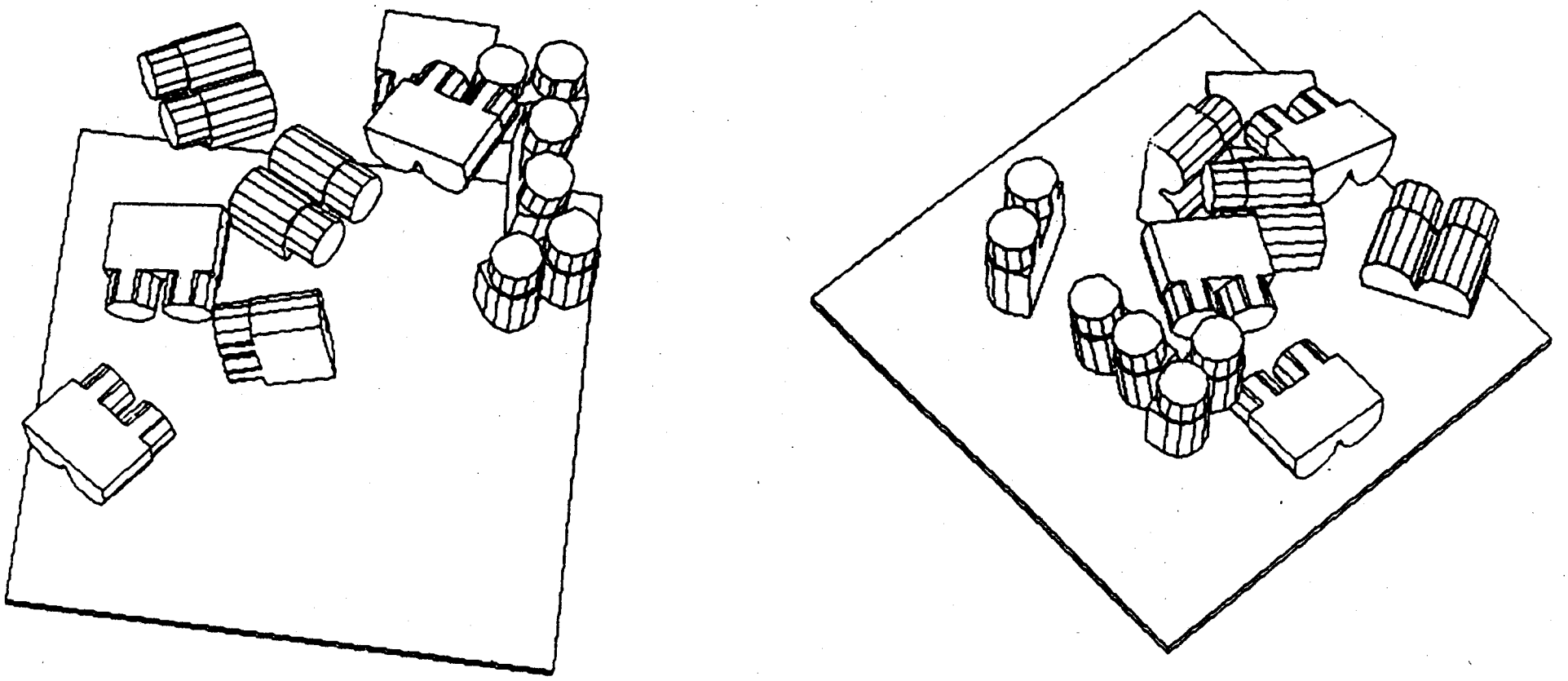


Fig 4e object configurations determined successfully.