INCREMENTAL CONSTRUCTION OF 3-D MODELS FROM A SEQUENCE OF FRAMED VIEWS : MATCHING PARTIAL OBJECTS

Shun-en Xie and Thomas W. Calvert

LCCR. School of Computing Science. Simon Fraser University
Burnaby. B.C.. V5A 1S6. Canada

ABSTRACT

In the future intelligent mobile robots will be called upon to play many important roles. In many realistic situations. the knowledge of the structure and placement of objects in an environment should be learned rather than built in. Thus the mobile robot must often construct 3-dimensional models for the objects by analysing sensed multiple views.

In this paper. we describe an approach to the incremental construction of 3-D body models in a practical office or warehouse environment by matching planned multiple views. In particular. we discuss the following aspects:

1. the decomposition of a framed view and the construction of partial 3-D descriptions of the view:
2. the matching of partial 3-D descriptions of a view with the built-in model of the robot environment:
3. the matching of partial descriptions of bodies derived from the current framed view with partial models constructed from previous views:
4. the identification of the new information in the current view and the updating of the models:
5. the identification of the unknown parts of the models which are being constructed so that further vantage viewpoints can be planned.

This approach combines such intelligent robot functions as attention. planning. sensing. learning and knowledge rectification. A prototype system for matching and constructing 3-D body models has been implemented and tested with synthesized images using C-PROLOG under Berkeley UNIX on a VAX 11/750.

INTRODUCTION

In the future computer-controlled robots will be called upon to play many important roles in industrial. business and domestic situations. If these robots are to work in complex environments it will be necessary to develop knowledge-based sensory systems. In simple situations. the robot vision system can have built-in models of both the environment and all objects within it: this allows a relatively simple recognition process. In more realistic situations. however. although the geometry of the surrounding environment may be known (i.e. the dimensions of the room. warehouse. etc, in which the robot operates). the type and position of the objects in the environment will generally be unknown. Thus knowledge of the structure and placement of these objects must be learned. To do this the mobile robot must first construct 3-dimensional models for the objects it encounters. It should then be possible to classify these objects by comparing their structural properties with those of generally known classes of objects such as benches. chairs. tables. etc.

In analysing a single framed view of part of a large scene. the problems which will generally stand in the way of constructing the 3-D body models include:

1. partial features:
2. self-occlusion:
3. occlusion:
4. accidental alignment and special alignment:
5. undetermined geometric parameters.

An approach to understanding a scene from image sequences by incrementally constructing body models seems promising. However. even to-day. the information processing load involved in analysing a sequence of images presents a serious technical problem. Dynamic selection of a minimal set of vantage viewpoints and effective selection of only the necessary information will be essential if the burden of computation is to be lightened. Fortunately. a mobile robot. by its nature. offers a good foundation for gathering information from different points of view. Thus combining a vision system with a planner. so that a scene can be analysed from planned multiple views. is both natural and necessary.

In this paper. we describe a system which incrementally constructs 3-D object models of an office or warehouse scene from planned multiple views. In particular. we address the matching and construction of 3-D partial models.

To limit the scope of the immediate research problem the following assumptions have been made:

1. The bodies in the environment are static. rigid. weakly externally visible. and have vertices formed by at most three surfaces. Edges are formed by two surfaces. which can be planar. conical. cylindrical or spherical.

This work was supported in part by grants from the NSERC of Canada. Mr. Xie was supported by the C. D. Nelson Memorial Scholarship.

2. The only lights in the environment are one point source and one diffuse source.
3. The shape and dimensions of the robot environment are given.
4. A pinhole spherical camera model is used to acquire the images.
5. There is a preprocessor which deals with early and intermediate vision processing of the visual data. The output of the preprocessor is equivalent to a complete 2 1/2 D sketch. The categories of facets and lines, the orientations of planes and the rough depths of junctions have already been extracted from the 2 1/2 D sketch.

Partially constructed models which remain incomplete after a sequence of views are analysed by a viewpoint planning system, which is described in a companion paper[1]; using this system new views can be chosen to resolve the ambiguities. In any realistic situation, we would expect that the task assigned to the robot would also provide input to the planning system so that a decision could be made to ignore incomplete objects which were irrelevant to the current task.

The body models (either partial or complete) which are constructed from the multiple views have Boundary Representation (BR) like representations. Once a complete model has been constructed, a rule based conversion system which has been described elsewhere[2] is used to transform the BR representation into our new "Constructiv Solid Geometry - Extanded Enhanced Spherical Image" (CSG-EESI) representation which provides both structural and geometric information for the bodies. Higher level 3-D models can be more easily derived from the CSG-EESI representation. This facilitates object classification by comparing a structure with those for prototype objects which might be expected to be in the environment.

BACKGROUND

There has been considerable research on the segmentation and labeling of images. After Guzman[3], Huffman[4], Waltz[5] and Turner[6], Chakravarty[7] generalized a line and junction labeling scheme that deals with planar-faced or curved-surface solid bodies, having vertices formed by three surfaces. In this scheme, 3 types of lines and 8 types of junctions were defined. By dealing with regions and lines, objects can be correctly labeled by the set of junction types.

An object must often be observed from several directions in order to form an assessment of what it looks like or to form a 3-D model of it. In order to form a 3-D model of an object from sequential views, Underwood and Coates[8] developed a program which forms a 3-D description of a planar convex object when the object is rotated in space. In their match algorithm, the connections between surfaces, the number of edges which bound a surface and the clockwise ordering of edges form the deterministic factors. Later, Preiss[9] described an approach which interprets a standard engineering drawing of a planar object for construction of its 3-D representation. This approach consisted of three main steps:

1. interpretation of projected faces,
2. interpretation of dashed lines,
3. assembling them into a body.

The connectedness properties and the geometric relationships between junctions or faces (such as coplanar relationships), were used for matching junctions. In this approach, the final 3-D representation of a body is complete, and consists of each of its faces, edges and vertices along with the three coordinates of each vertex.

Several researchers have investigated the problem of matching multiple views of a block's world in front of a featureless background. Using wide-angle stereo, Ganapathy[10] designed a scheme which uses some heuristic rules, such as "Single Match", "Order Match", "Connectivity" and "Table Match", to choose an initial match between corresponding vertices in order to reduce the search space. His program finally stops after building up the 3-D coordinates of the vertices. Shapira and Freeman[11, 12] developed a program for constructing a description of solid bodies from a set of pictures taken from different vantage points. A heuristic procedure was devised for establishing matches between junctions in the different pictures and determining the validity of doubtful junctions. It first establishes matches among junctions by using the constraints of projection and the connectedness between junctions; then it establishes matches among lines by using the cyclic-order property and fills in missing connections between junctions and missing junctions. The final description reported by the program involves bodies made up of their face groups which are described in terms of triples of matched lines. Asada[13] developed a system which describes 3-D motions of jointed trihedral blocks. In this system, a Huffman-like labeling scheme and an object-to-object matching method are first used to segment the line-drawing images into individual blocks and to find the possible correspondence of their junctions between closely consecutive frames. A transition table of junction labels and contextual information is used to analyse structural changes of the line drawings. Then, the shape rigidness property of three vertices on a block is used to evaluate geometrical parameters, such as the 3-D coordinates of the vertices and motion parameters.

It is only in recent years that attempts have been made to match multiple views in a complex environment in order to incrementally construct some kind of model of a scene. Herman, Kanade and Kuroe[14] described the 3-D MOSAIC project whose goal is to incrementally acquire a 3-D model of an urban scene from images. Their method is to first extract 3-D shape information from the images by stereo analysis, then to match two views based on junction matching and finally to generate an approximate model of the scene by using task-specific knowledge. Crowley[15] described a navigation system for an intelligent mobile robot which included techniques for the construction of a line segment description of a recent sensor scan and the integration of such descriptions to built up a model of the immediate environment using a list of directed line segments. Herman[16] described an algorithm which matches vertices in two 3-D descriptions. The algorithm consists of three main steps: first, initial matches are obtained for each vertex based on local properties; next, a Waltz-filtering procedure is applied which propagates topological constraints to reduce the set of matches; finally, a tree-type search which uses both topological and geometrical constraints gives globally consistent sets of unique matches.

OVERVIEW OF THE SYSTEM

The system which we have developed for incrementally constructing 3-D models of objects is illustrated schematically in Figure 1.

The incremental construction of object models from planned multiple views involves the following principal elements:

1. the decomposition of a framed view and the construction of partial 3-D descriptions of the view;
2. the matching of partial 3-D descriptions of a view with the built-in model of the robot environment;
3. the matching of partial descriptions of bodies derived from the current framed view with those partial models constructed from the previous views;
4. the identification of the new information in the current view and the updating of the models;
5. the identification of the unknown parts of the models which are being constructed so that further vantage viewpoints can be planned;
6. the finding of the relationships between bodies and

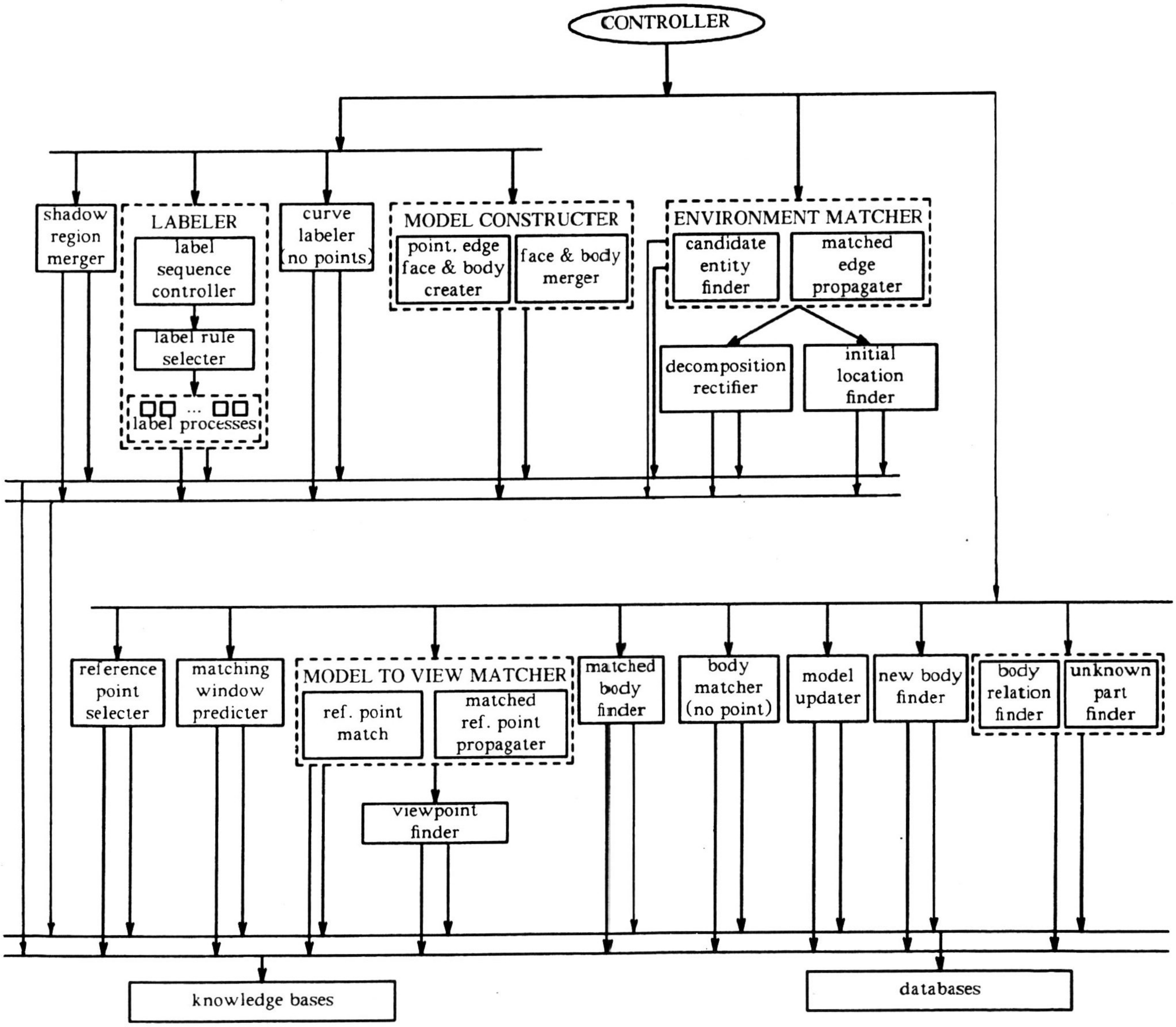

Figure 1

A schematic sketch of the system fo matching and model construction.

their environment and constructing a partial map of the scene.
In the following sections we discuss elements 1 - 5 in turn. Each of these elements is more or less related to matching. In our approach, matching is based on the rules (contraints) derived from geometry, topology, photometrics, triangulation and problem assumptions.

Strategies for recognition can be data-directed (bottom-up), knowledge-directed (top-down) or some mixture of the two. In our approach, a bottom-up and top-down mixed strategy is used to match partial 3-D descriptions of a view with a built-in model of the robot environment and a data-directed strategy is used to incrementally construct 3-D body models for the objects in the environment. We are interested in exploring how far we can go with a data-directed strategy.

In the scene learning process the robot vision system generally produces incomplete and erroneous knowledge of the objects in the environment; consequently, it is important to identify the unknown parts, to rectify the erroneous knowledge and to assimilate the new information. Our approach combines such functions of an intelligent robot as attention, planning, sensing, learning and knowledge rectification.

DECOMPOSITION

For each view, the partial 3-D descriptions of bodies are derived by labeling and segmenting the image. In general, the decomposition process first merges the regions that are separated by shadow lines, and then it labels the junctions and lines in the image. After labeling, internal representations are created for real vertices, edges and faces. At that time, those edges separated by virtual junctions are united, and the partly viewed edges and faces are identified. In the last step, faces are combined on internal edges to form bodies. Some relationships between bodies, such as "touch by" or "occluded by" will also be identified. Thus, hierarchical internal representations which are used as partial 3-D descriptions are constructed for each body in a view.

Starting from the Chakravarty and Waltz labeling schemes, a modified and extended labeling scheme has been devised for labeling scenes containing shadows and certain curved objects. The images are labelled by an expert system; knowledge of label (production rules) is stored separately in micro-knowledge-bases according to the categories of related junctions. The top level of the expert system controls the sequence of the labels. It first arranges the addresses of the junctions that need to be labeled in an "ORDER QUEUE". The junctions which are generally easier to label, such as "p", "w" and "y" types, are arranged at the front of "ORDER QUEUE". When a junction and its related lines have been successfully labeled, the junction is deleted from the "ORDER QUEUE". Meanwhile, its immediatly adjacent junctions will be inserted in a "PRIOR QUEUE". The top-level expert then propagates the labeled junctions to their immediate neighbours, thus the label procedure can take advantage of derived facts and the labeling time can be reduced.

The second level of the expert system selects the appropriate micro-knowledge-base, according to the category of the junction given by the top-level expert, and sequentially selects the production rules from the selected micro-knowledge-base in order to label the junction and its related lines. At the lowest level of the system are the processes which carry out the following tasks:

1. fetch the related facts from the appropriate micro-databases;
2. match the facts with the condition of a rule;
3. incorporate the label results into the appropriate micro-databases;
4. add the addresses of the immediate neighbours of the successfully labeled junction into the "PRIOR QUEUE".

The decomposition is conservative, i.e., it favours the separation of objects on concave edges. For a curve on which there is no feature point, a label is given according to its convexity: a concave curve is assigned as a concave boundary and a convex curve is assigned as a convex internal edge. The errors caused by an incorrect decomposition are expected to be rectified by facts collected later or by the knowledge stored at higher levels.

MATCHING THE ENVIRONMENT MODEL

The initial location of the robot in the environment is not known *a priori*. In order to determine the initial coordinates of the robot in a fixed coordinate system keyed to the environment, it is necessary to identify which entities in a framed view correspond to parts of the environment model. For this purpose, at least some edges of an entity should be matched with a connected part of the environment. Edges are stable, relative features which contain dimensional information and are at the lowest level (except for vertices). Since the geometry (shape and dimensions) of the environment are known, an edge-based matching process has been devised for matching the environment.

The process first matches the completely visible real edges of entities from a framed view with the built-in environment model; this is done according to their attributes, the categories (e.g. planar or conical) and the directions of their adjacent faces. The edge attributes consist of:

1. category, e. g., straight line, circle or other curve;
2. type, e.g., shadow, occluding boundary, concave internal, convex internal, concave boundary, clipping line or limb;
3. convexity;
4. approximate length.

If an edge in a view is matched with several edges in the environment model, then a "matching confidence" will be assigned to it which is proportional to the inverse of the number of matched pairs. An entity is considered to be a candidate for part of the environment model if at least its visible and well labelled internal and occluding edges match with edges in the environment model. From these candidates, entities will be designated as being parts of the environment on the basis of the following properties:

1. at least two matched edges;

2. a maximum number of matched edges;
3. a maximal sum of confidences for matched edges.

Following this identification, a top–down analysis process, which propagates the matched facts according to the built–in model of the environment, will be applied to those entities in order to:
1. Further verify the matched facts and find more matching facts. If in the propagation an inconsistent fact is discovered, then the initially matched entity will be rejected.
2. Identify the matched vertices and determine the position of the current viewpoint of the robot.
3. Rectify the result of the decomposition of the current view. When those concave edges, which were initially labeled as the concave boundaries, are revealed as the concave internal edges of the environment, their labels are revised and the corresponding bodies are merged together.

Since the 3-D coordinates of two known feature points and their spherical coordinates in a view can be used to determine the position of a viewpoint, the position can be determined from any two matched edges. From a pair of matched edges and the approximate depths of their related points (e.g. end points), the process identifies two pairs of the best matching points. Since the 3-D coordinates of the environment points are known, the possible position of the current viewpoint can be calculated from the two pairs of points. After another pair of matching edges has been discovered by the matching propagation procedure or when the other pair of matched edges is used for propagation, the new facts will be used to confirm or rectify the position of the viewpoint and make it more precise.

MATCHING PARTIAL BODY MODELS

Once the environment model has been matched, the partial 3-D descriptions from the first view will be used as the initially constructed partial models of the bodies in the scene.

In order to match the partial descriptions of bodies derived from a new view with those partial models constructed in the previous views, a multi-level feature matching approach has been used. This approach first matches the partially constructed models to those 3-D descriptions in the current view by selecting those reference vertices from the object models and the environment model which have the following features:
1. they are valid vertices or Shadow Intersection Points (SIP);
2. they are within the new view frame (though sometimes they may be occluded and unseen);
3. for each reference vertex, either the directions of the two constituent faces are known or the projection of the vertex is a boundary point of an unknown area in the current constructed map;
4. they are related to the objects of current interest.
Following the selection of the reference vertices a prediction process determines the possible matching windows in the new view; this is done on the basis of the approximate position of the new viewpoint and the coordinates of the reference vertices which may only have approximate values stored in the models. The process finds the candidates for matching in the new view which are located within the matching windows. The widow sizes are determined by the tolerance errors of the robot movement, the errors of the coordinates of the refernce points and the relative positions of the new viewpoint and the reference points.

In the knowledge base, there is a "Junction Family Dictionary". In the dictionary, each family consists of the possible junction types for a specific kind of vertex, when it is viewed from different positions. Using the Junction Family Dictionary, and the categories and directions of the constituant faces, the matching process assigns each candidate a confidence. The candidate which has the uniquely highest confidence will be chosen as the matched vertex for a reference vertex, and its corresponding faces will be considered as matched faces. After finding a matched pair of vertices, the matching process propogates the fact along the emanating edges to adjacent vertices. A depth-first search is used at each pair of matched vertices, and when partial edges, unmatched vertices caused by occlusion or already matched facts appear, the match propagation for that direction will stop. Thus, different levels of features (i.e. faces, edges and vertices) can be matched in the same propagation process.

For those constructed body models which do not have any vertex or feature point (e.g. SIP), the edges and faces will be chosen as the basic matching elements. According to their categories, types, shape parameters and approximate positions, the corresponding feature elements can be found. Also the matched facts will be propagated to their related feature elements.

After the faces have been matched, matched bodies can be derived from these faces. Following this, the model updating process searches the matched facts starting with high level features and moving to low level features; the low level features which do not exist in the body model are now filled in by the known parts of the current 3-D body description which are matched. After matching the partially constructed models to those 3-D descriptions in a new view, unmatched bodies in the new view are identified. In order to test whether these are newly discovered bodies, a reverse direction matching process is used to check whether any vertex in an unmatched body has a corresponding vertex in the body models or the environment model. If this is not the case, then the body is new and is added to the database of the body models; otherwise the appropriate matched model will be found. Meanwhile, features separated in the new view or in the constructed models may be merged into one if their correspondence is unique in one of the two 3-D representations. The related revision will also be done.

POSITION ADJUSTMENT

In practice, data gathered by a robot vision system always includes certain tolerance errors. Although the relative positions of the viewpoints can be derived from a robot servo system, this information is generally imprecise. Since any two views form a pair of wide angle stereo images, the matching process provides the information necessary to calculate the position of the sensor (the robot camera) quite precisely. This information can be used to correct the position calculated by the movement control servos and used for dynamically adjusting the robot movement.

IDENTIFICATION OF UNKNOWN PARTS

The ambiguities caused by special alignments and accidental alignments generally can be distinguished by using multiple views. For example, when a strange junction type occurs in a view, if from other views the matched points belong to the same family of junctions, then it is caused by a special alignment, otherwise it is caused by accidental alignment. The ambiguity caused by accidental alignments can be ignored. For a special alignment, the ambiguity can often be solved by a correct decomposition, though sometimes higher knowledge of the scene may be needed.

Inside a body, self-occlusion may result in unknown occluded parts. Between bodies, an occlusion may cause the occluded bodies to be unidentified. For these two cases, unknown parts only occur at the occluding edges. Besides, a concave boundary edge indicates that the two related bodies are touching, and hence the touching parts cannot be seen if there is no means to change the status of the bodies.

In the system described here, when a model of a body has been created, only the internal, occluding and concave boundary edges which are the real edges of the body are created. The model also contains a list of its boundary edges and a list of the bodies which occlude it. When a newly discovered surface is added into a body model, it is necessary to change the types of those occluding edges in the body model, which are matched with the edges of the added surface. These edges become the internal edges of the body and are deleted from the boundary list of the partial model of the body. Thus boundary occluding edges of a body model always indicate the self-occlusion of parts and the need for further attention.

The "t" type junctions caused by occlusion are kept in the input image databases. Although they are not the vertices of a body, they are important points for the construction of a map of the scene and for discovering the unknown parts caused by occlusion. For an occluded body, discovering its occluded parts is accompanied by a search for its "t" type points and those incompletely seen edges and surfaces which relate to the "t" type points.

All of the above outcomes will be organized and analysed by a view planning system in order to further resolve the ambiguities. This componant has not yet been developed and implemented.

EXPERIENCE

The system for matching and constructing 3-D body models has been implemented by using C-PROLOG under the UNIX operating system on a VAX 11/750.

Figure 2 shows two synthesized views from the scene shown in Figure 3; they have been successfully analyzed by the system described above.

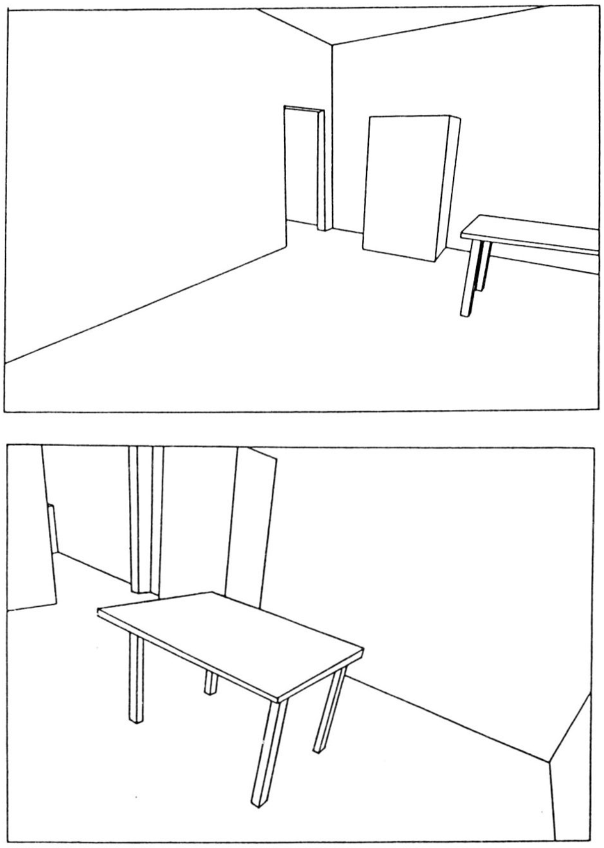

Figure 2

The two synthesized views which have been successfully analyzed by the described system.

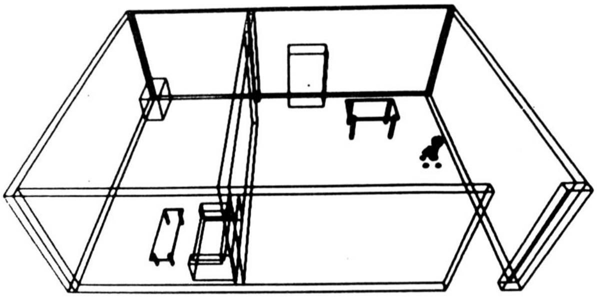

Figure 3

An example of a typical scene (two views of this scene are used in Figure 2).

CONCLUSION

Under the assumptions described in the introduction, the system described here can incrementally construct 3-D body models in an office or warehouse environment by matching planned multiple views. No prior knowledge of the objects is required by this system. The system includes the following important features:

1. a framed view is decomposed and partial 3-D descriptions of the view are constructed;
2. partial 3-D descriptions of a view are matched with the built-in model of the robot environment;
3. partial descriptions of bodies derived from the current framed view are matched with those partial models constructed from the previous views;
4. the new information in the current view is identified and the models are updated;
5. the unknown parts of the models which are being constructed are identified so that further vantage viewpoints can be planned.

Together with a view planning system[1] and the CSG-EESI 3-D model conversion system[2], this system offers a good basis for constructing a higher level image understanding system for an intelligent robot.

As noted above, the system has been implemented in C-PROLOG under the UNIX operating system on a VAX 11/750, and has been tested successfully with synthesized images. While C-PROLOG provides a good environment for testing the ideas used in this system, any practical implementation would have to be more efficient.

REFERENCES

1. S. Xie, T. W. Calvert and B. K. Bhattacharya, "Planning views for the incremental construction of body models", *submitted to The 8th International Conference on Pattern Recognition*, Paris, France, 1986.

2. S. Xie and T. W. Calvert, "The CSG-EESI scheme for representing solids with a conversion expert system", *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Francisco, Ca., 1985, pp. 124-129.

3. A. Guzman, "Decomposition of a visual scene into three-dimensional bodies", *Proc. AFIPS Fall Joint Compt. Conf., vol. 33*, 1968, pp. 291-304.

4. D. A. Huffman, *Impossible objects as nonsense sentence,* Edinburgh Univ. Press, Edinburgh, U.K., 1971.

5. D. L. Waltz, *Understanding line drawings of scenes with shadows,* McGraw-Hill, New York, 1975, pp. 19-91.

6. K. J. Turner, "Computer perception of curved objects", *presented at the AISB Summer Conf.*, Univ. Sussex, Brighton, England, 1975.

7. I. Chakravarty, "A generalized line and junction labeling scheme with applications to scene analysis", *IEEE Trans. on PAMI*, Vol. PAMI-1, 1979, pp. 202-205.

8. S. A. Underwood and C. L. Coates, "Visual learning from multiple views", *IEEE Trans. on Computers*, Vol. C-24, No. 6, June 1975, pp. 651-661.

9. K. Preiss, "Algorithms for automatic conversion of a 3-view drawing of a plane-faced part to the 3-D representation", *Computers in Industry*, Vol. 2, 1981, pp. 133-139.

10. S. Ganapathy, *Reconstruction of scenes containing polyhedra from stereo pair of views,* PhD dissertation, Stanford Artificial Intelligence Laboratory, Memo AIM-272 , December 1975.

11. R. Shapira, "Reconstruction of curved-surface bodies from a set of imperfect projections", *Proc. of the 5th IJCAI*, 1977, pp. 628-634.

12. R. Shapira and H. Freeman, "Computer description of bodies bounded by quadric surfaces from a set of imperfect projections", *IEEE Trans. on Computers*, Vol. C-27, No. 9, Sept. 1978, pp. 841-854.

13. M. Asada, *Understanding of three-dimensional motions in trihedral world,* PhD dissertation, Dept. of Control Engineering, Osaka University, Japan, Jan. 1982.

14. M. Herman, T. Kanade, and S. Kuroe, "Incremental Acquisition of a Three-Dimensional scene model from images", *IEEE Trans. on PAMI*, Vol. PAMI-6(3), May 1984, pp. 331-340.

15. J. L. Crowley, "Dynamic world modeling for an intelligent mobile robot using a rotating ultra-sonic ranging device", *Proc. IEEE Inter. Conf. Robotics and Automation*, St. Louis, Miss., 1985, pp. 128-137.

16. M. Herman, "Matching three-dimensional symbolic description obtained from multiple views of a scene", *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, San Francisco, Ca., 1985, pp. 585-590.