Planning Sensing Strategies in a Robot Work Cell with Multi-Sensor Capabilities

SETH A. HUTCHINSON, MEMBER, IEEE, AND AVINASH C. KAK, MEMBER, IEEE

Abstract—**In this paper, we describe an approach to planning sensing strategies dynamically, based on the system's current best information about the world. Our approach is for the system to automatically propose a sensing operation, and then to determine the maximum ambiguity which might remain in the world description if that sensing operation were applied. The system then applies that sensing operation which minimizes this ambiguity. To do this, the system formulates object hypotheses and assesses its relative belief in those hypotheses to predict what features might be observed by a proposed sensing operation. Furthermore, since the number of sensing operations available to the system can be arbitrarily large, we group together equivalent sensing operations using a data structure that is based on the aspect graph. Finally, in order to measure the ambiguity in a set of hypotheses, we apply the concept of entropy from information theory. This allows us to determine the ambiguity in a hypothesis set in terms of the number of hypotheses and the system's distribution of belief amongst those hypotheses.**

I. Introduction

WITH CURRENT TECHNIQUES in geometric modeling, it is possible to create very detailed representations of objects, containing a large number of features, and expressing a large number of relationships between those features. Likewise, the current state of computer vision (both 2-D and 3-D) and tactile sensing make it possible to extract large feature sets from sensory data. Unfortunately, for the purpose of object recognition and localization, large feature sets can require exponential computational resources. Furthermore, it is often the case that the set of features extracted by a single sensor will yield ambiguous results. For these reasons, a sensing system should be able to choose which sensing operations to apply, so that it extracts the minimum number of features required to uniquely identify the object and its pose. Furthermore, the system should base its decisions on its current hypotheses about the identity and pose of the object. This implies that the system should be able to dynamically assess its hypotheses about object identities and poses, and select a sensing operation which will yield the greatest reduction in the ambiguity in that information.

Previous work in planning sensing strategies has been divided into two distinct areas. One of these is concerned with sensor placement, that is, placing the sensor so that it can best observe some feature (which is predetermined) or region of 3-space. The other is choosing a sensing operation which will prove the most useful in object identification and localization.

Research on optimum sensor placement has been reported by Connolly in [3], Kim *et al.* in [19], and Cowan and Kovesi in [4]. In [3], a system is described which builds up a complete model of a scene by filling in an octree data structure. Since no single viewpoint can observe an entire scene, a sequence of viewpoints is chosen. In [19] a system is described which determines successive camera viewpoints so that the most distinguishing features of the object can be observed. By using the object's aspect graph, the selection of viewing direction is reduced to finding which node in the graph corresponds to the best view of the desired feature. Note, that in this system, the best distinguishing feature is predetermined, and the problem is merely determining the best viewpoint from which to observe this feature. A similar problem has been discussed in [4], where sensing strategies are selected based on object and camera models such that a number of constraints are simultaneously satisfied, for example, the spatial resolution must be better than some minimum value, the surface to be viewed must lie within the camera's field of view. The region of viewpoints in 3-space which satisfies each constraint is determined, and the intersection of these regions defines the set of allowable viewpoints.

Work on automatically determining optimal sensing strategies has been reported by Ikeuchi [17], Hanson and Henderson [11], Magee and Nathan [21], and Hager and Mintz [10]. Ikeuchi's work is based on the automatic synthesis of interpretation trees which are used to guide feature selection. In his approach, the higher level nodes in the interpretation tree yield the aspect of the object, and then the lower level nodes are used for computing the precise pose of the object. This scheme makes use of the fact that for most objects the set of features useful for discriminating between aspects differs from the set of features useful for determining the exact pose once the aspect has been determined.

In the work reported by Hanson and Henderson, a set of filters is used to select the best identifying features (based on rarity, robustness, cost, etc.) for each aspect. These features and their associated aspects are compiled into a strategy tree which, in purpose, is similar to Ikeuchi's interpretation tree. The strategy tree has two levels. Each node at the first level allows aspect hypotheses to be invoked on the basis of certain features and their values. For each hypothesis at a first level node, there exists a Corroborating Evidence Subtree, which is

Manuscript received July 1, 1988; revised May 20, 1989. This work was supported by the National Science Foundation under Grant DCR 8803017 to the Engineering Research Center for Intelligent Manufacturing Systems.

The authors are with the Robot Vision Laboratory, School of Electrical Engineering, Purdue University, West Lafayette, IN 47907.

IEEE Log Number 8930466.

used to guide the search for evidence that supports that hypothesis and for carrying out the computations for determining the object's pose.

The work reported by Magee and Nathan describes a system which is capable of selecting disambiguating features. The method is based on model differencing. With this approach, a potential disambiguating feature is tested to see if it could be instantiated in more than one candidate object model. If not, it is selected as a disambiguating feature, and the sensor is repositioned to observe that feature. The limitation to this approach is that the disambiguating feature is chosen based on its ability to discriminate between only two sets of hypotheses: one set containing a single candidate model (to which the disambiguating feature belongs), and the other set containing the remaining candidate models. This system does not appear able to select features which could simultaneously discriminate between more than two sets of hypotheses.

In the work reported by Hager and Mintz decision theoretic techniques are applied to the problem of selecting optimum sensing strategies. Their methods are aimed at finding the value of some parameter associated with the object. They accomplish this by treating sensors as noisy information sources, and then associating a risk function with each sensing operation. Selection of a sensing operation is achieved by minimizing the risk function. This work assumes that the identity of the object being viewed is known and that the value of the parameter is known *a priori* to lie within some confidence interval.

In this paper, we present the work that we have done in dynamic sensor planning, which extends the work cited above in a number of directions. First, we give the system the ability to choose sensing strategies based on current hypotheses about the identity and position of an object which is being examined. It is possible that each such hypothesis will correspond to a different object. Furthermore, the choices of sensing strategies are not limited by the use of a single type of sensor. The sensor types currently incorporated in the system include a 3-D range scanner, 2-D overhead cameras, a manipulator held 2-D camera, a force/torque wrist-mounted sensor, and also the manipulator fingers for estimating the grasp width. The vision sensors can be used to examine objects from arbitrary viewpoints, while the manipulator and force/torque sensor can be used to measure other features such as weight, depth of occluded holes in the object, etc.

It is important to realize that with these additional sensory inputs, we can discriminate between object identities, aspects, and poses that would otherwise appear indistinguishable to just a fixed viewpoint vision-based system. Our system is capable of dynamic viewpoint selection if that is what is needed for optimum disambiguation between the currently held hypotheses.

We attack the problem of viewpoint and sensor-type selection as follows. Once the system has developed a working set of object hypotheses, candidate sensing operations are automatically proposed and evaluated with regard to their potential effectiveness. (In our system, an object hypothesis is merely a set of matches between sensed and model features. This will be explained in more detail in the remainder of the paper.) This evaluation is performed as follows. For each hypothesis in the current hypothesis set, the system determines the set of features that would be observed by the candidate sensing operation if that hypothesis were correct. Using these predicted features, the system determines the hypothesis set that would be formed if these features were actually found by some sensing operation. The ambiguity of this predicted hypothesis set is calculated and noted. This is repeated for each hypothesis in the hypothesis set, and the maximum value of the ambiguities is associated with the proposed sensing operation. The sensing operation which minimizes this maximum ambiguity is then selected for application.

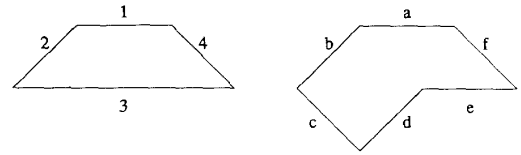


Fig. 1. Two 2-D object models.

In the remainder of the paper, we will describe each of the above steps in some detail. First, in order to give the reader a clear understanding of the problem, in Section II we present an example with 2-D objects. In Section III, we introduce our object representation. This representation is used both to quantize the space of sensing operations and to predict the features which would be observed by a candidate sensing operation. Section IV describes how our system generates and refines hypothesis sets, as well as how evidential reasoning is implemented in the system. In Section V, we define the measure of ambiguity which our system uses. The measure that we describe is based on entropy from information theory. Section VI contains a discussion of how object position hypotheses are generated, and how these are used (in conjunction with the aspect graph of the object) to generate a set of predicted features for a particular viewpoint/sensor combination. In Section VII, we describe the types of sensors our system uses, and the types of features that they can detect. Section VIII brings together the previously discussed concepts and presents a formal algorithm for selecting the next best sensing operation. Finally, in Sections IX and X, we describe some of our experimental results and provide a summary of the contributions made.

II. An Illustration of the Problem

This section of the paper introduces a simple two-dimensional (2-D) example which will be used throughout the paper to illustrate the various aspects of our approach to sensor planning. In this example, it is assumed that the sensory system is capable of observing a 2-D object from an arbitrary viewpoint in 2-space. Such an observation will yield the set of edges in the scene which are visible from that viewpoint. The information reported by the sensory system for a particular edge includes the location, orientation, and length of the edge, each of these quantities being subject to experimental error. In our example, we assume that there are two possible objects, which are shown in Fig. 1. In the interest of clarity, the edges of the left model object have been assigned the integer labels 1

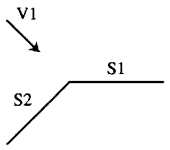


Fig. 2. Two edges, as observed from viewpoint $V1$.

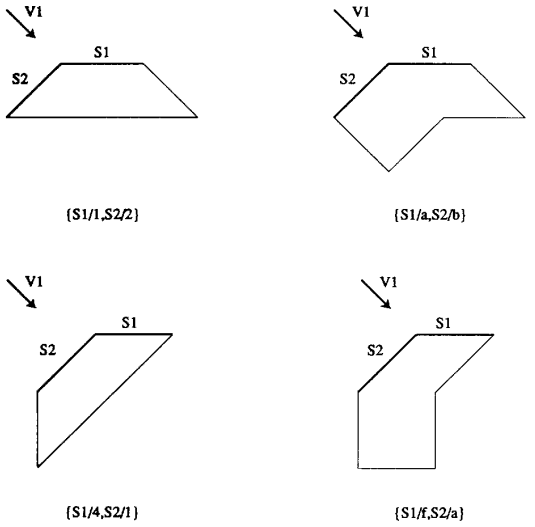


Fig. 3. Edges visible from viewpoint $V1$, and corresponding hypothses.

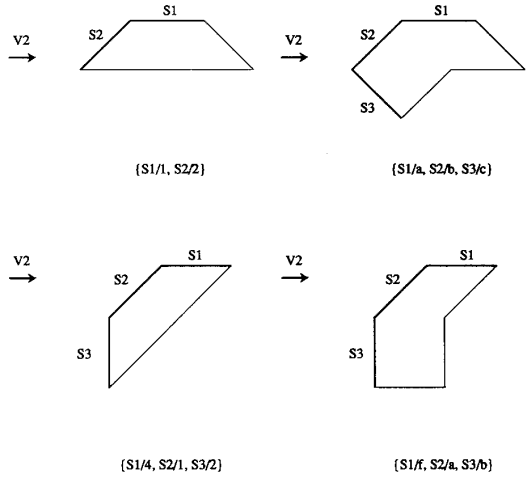


Fig. 4. Edges visible from proposed viewpoint $V2$, and possible resulting hypotheses.

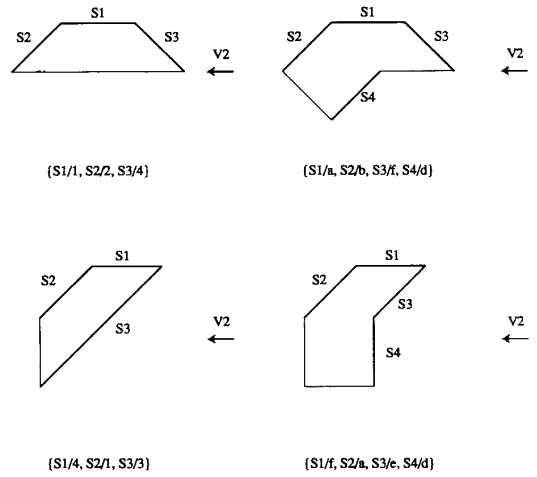


Fig. 5. Edges visible from proposed viewpoint $V2$, and possible resulting hypotheses.

through 4, while the edges in the right model object have been assigned the alphabetic labels a through f.

Assume that an arbitrary viewpoint $V1$ is selected for the first sensing operation, and that the edges observed from this viewpoint are as shown in Fig. 2. By matching the two observed edges ($S1$ and $S2$) to edges in the two model objects, and then using the relational constraints in the models to prune away impossible pairs of matches (this process will be described in greater detail in Sections IV and VII), four possible hypotheses are derived, as shown in Fig. 3. Note, a hypothesis about the identity and position of an object in the scene is represented by a set of matches between sensed and model features, since such a set of matches contains an explicit hypothesis about the identity of the sensed features (and therefore of the object) and an implicit hypothesis about the position of the object. For example, the first hypothesis in Fig. 3 is denoted by $\{S1/1, S2/2\}$, meaning that sensed edge $S1$ is matched to model edge 1, and sensed edge $S2$ is matched to model edge 2, both model edges belonging to the left object in Fig. 1.

Since there are four possible hypotheses about the identity and position of the object in the scene, a second sensing operation is required for disambiguation. In this example, the problem is selecting a viewpoint from which the set of features that are observed will uniquely identify the object and its position, regardless of which of the four hypotheses in Fig. 3 is correct. Figs. 4 and 5 illustrate two possible viewpoints. In each of these figures, the edges that would be observed for each of the four hypotheses are indicated, as are the edges which have already been observed (from $V1$). If the viewpoint in Fig. 4 is chosen, the observed edges will not be sufficient to distinguish between the third and the fourth hypotheses, since, if either the third or fourth hypothesis is correct, a single new edge will be observed, at an angle of 45° from $S2$. However, if the viewpoint shown in Fig. 5 is chosen the observed edges will uniquely identify the object and its orientation, regardless of which hypothesis is actually correct. Therefore, the viewpoint shown in Fig. 5 is a better choice.

III. Object Representation

The object representation used in our system plays two key roles. First, it allows us to quantize the space of sensing operations. This is a result of the fact that the representation groups together sets of object features which can be viewed from a single viewpoint (such a set of features is referred to as an aspect). This, in turn allows us to group together viewpoints which observe the same aspect. Second, the

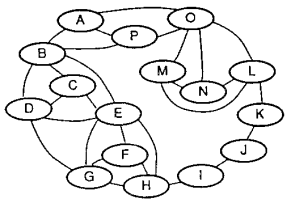


Fig. 6. Aspect graph of the rightmost object of Fig. 1.

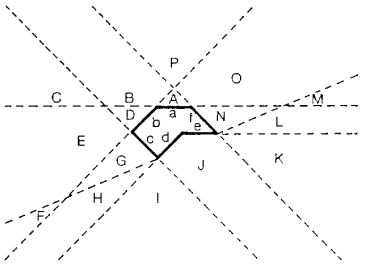


Fig. 7. Regions which view the different aspects of the rightmost object of Fig. 1.

representation allows us to easily determine the features of an object which will be observed by a particular sensor from a particular viewpoint relative to the object. This is done by determining which aspect will be observed from the viewpoint, and then looking up the object features which are associated with that aspect. In this section we will describe the aspect graph representation and how aspect graphs are derived by our system.

The aspect graph was originally developed by Koenderink and van Doorn [20] (who referred to it as the visual potential) to characterize the visual stimulus produced by an object when viewed from different relative positions. This function is defined in terms of the invariant properties of the object and the relative positions of the viewer and the object. The local behavior of the function is defined in terms of the deformation of the retinal images through changing perspective. The global behavior of the function is defined in terms of its singularities. Two types of singularities have been considered: point singularities, which determine a system of protrusions facing the observer, and line singularities, which correspond to the curve on an object that divides its surface into visible and nonvisible regions. An aspect is characterized by the structure of these singularities for a single view. From most vantage points, an observer may execute small movements without affecting the aspect. When an observer's movement causes the structure of the singularities to change, an event is said to have occurred, and a new aspect is brought into view. An aspect graph is created by mapping aspects to nodes and mapping the events that take the viewer from one aspect to another to arcs between the corresponding nodes.

We are interested in features which can be observed by various sensors. Thus we characterize aspects, not in terms of the singularities in the function which defines the visual inflow, but in terms of observable features. In particular, we define an aspect to be a set of features which can be observed simultaneously from a single viewpoint. When a change of viewpoint causes a previously visible feature to no longer be visible, or a new feature to come into view, an event occurs. We use the aspect graph to group viewpoints that see the same aspect into equivalence classes. Associated with a node in the aspect graph is the set of viewpoints from which that aspect can be observed. Arcs in the graph connect nodes with adjacent viewpoints. Also, with each node in the aspect graph, we associate a principal viewpoint.

As an example, Fig. 6 shows the aspect graph for the 2-D object on the right in Fig. 1. Fig. 7 shows the object, and the regions of 2-D space from which each aspect can be observed. For example, aspect B contains features a and b, and aspect I contains features c, d, and e. Since the regions which view aspects B and P are adjacent, there is an arc between B and P in the aspect graph.

Aspect graphs for objects can be generated analytically or by exhaustive examination of the object. Analytic techniques have been reported by Castor and Crawford [1], Stewman and Bowyer [28], and Gigus and Malik [8]. In [1], two algorithms are briefly described which generate aspect graphs for 2 1/2-D solids. Aspect graphs for concave solids are generated by first finding the aspect graph for the convex hull of the solid. The convex hull is then deformed to recover the original object. A corresponding transformation to the aspect graph produces the aspect graph of the original object. In [28] the aspect graph of a convex planar object is constructed using its boundary surface representation. This is done by using the object's bounding planes to segment 3-space into viewing cells. Nodes in the aspect graph are established by determining which faces of the object are visible from each viewing cell. A face is visible from a viewing cell if the cell lies to the outside of the bounding plane for that face (where the inside of the plane is the side of the plane on which the object lies). In [8], a method for constructing aspect graphs of polyhedral objects under orthographic projection is described. Aspects are defined in terms of the qualitative structure of the line drawing of an object, and a catalog of all visual events which can occur for polyhedral objects is provided.

We generate our aspect graphs exhaustively. This is done by creating a CAD model of the object, centered within a tessellated viewing sphere. The geometric modeler is then used to view the object from the center point of each tessel, and the set of visible features is recorded. Using this information, it is a simple matter to generate the aspect graph. Tessels from which the same feature set is observed are grouped together into nodes. The arcs between nodes are generated using tessel adjacency (i.e., if two tessels are adjacent, and correspond to distinct aspects, an arc is added to the aspect graph connecting the two corresponding nodes). Finally, each aspect is assigned a principal viewpoint, which is defined as the average location of the centers of the viewing tessels associated with the aspect, with the constraint that it lies within a tessel that observes the aspect.

IV. Generating Object Hypotheses

As we mentioned in Section II, a set of feature matches defines an object hypothesis, in that it contains an explicit hypothesis about the identity of the object, as well as an

implicit hypothesis about the position of the object. The example of Section II, however, made a number of simplifications from real-world problems, perhaps the most severe being that all matches between sensed and model features were considered to be of equal quality, and therefore all object hypotheses were treated as being equally likely. In the system described in this paper, belief is assigned to an object hypothesis based on quality of feature matches, object consistency, similarity of relations in sensed data to relations in model objects, and aspect consistency.

A number of approaches to reasoning with partial evidence have been proposed in the literature. These include Bayesian methods, fuzzy set theory, certainty factors, and the Dempster–Shafer (DS) theory. For our application, the DS theory appears to be the best choice, for a number of reasons. First, in many cases we will need to express ignorance with regard to specific choices for the identity of a feature observed in a scene (for example, when several model features are identical in appearance). The DS theory is particularly well suited to this, allowing belief to be assigned not only to single propositions, but also to sets of propositions. Second, our system will have no *a priori* probabilities. The DS theory does not need these. Finally, as new features are found (by invoking additional sensing operations) new object hypotheses will be generated. These new hypotheses will be extensions of the old hypotheses, created by adding new matches between sensed and model features to those old hypotheses. This process corresponds nicely to the concept of refining a frame of discernment in the DS theory.

One of the primary objections to the DS theory is that the worst case time complexity for the implementation of Dempster's combination rule (the mechanism used to combine belief from two separate sources of evidence) is exponential in the size of the hypothesis set. Fortunately, the characteristics of the object hypotheses generated by our system allow for a polynominal time implementation of the combination rule.

In the remainder of this section, we elaborate the process of formulating, and assigning belief to, object hypotheses. For the sake of those not well acquainted with the DS theory, Section IV-A provides an introduction to the theory, and an explanation of how it is applied in our domain. A thorough explanation of the DS theory can be found in [25]. In Section IV-B, we describe how sensory measurements are converted to object hypotheses. Finally, in Section IV-C, we show that our implementation of Dempster's combination rule has polynomial time complexity.

A. The Dempster–Shafer Theory

The reasoning process used by our system requires a formalism for representing beliefs in hypotheses about an object's identity and position. Furthermore, since these beliefs will come from a number of independent sources, the formalism must include a method of combining beliefs from distinct sources to obtain a set of aggregate beliefs. The Dempster-Shafer (DS) theory provides such a formalism. In this section, we will provide the reader with an introduction to the DS theory and develop a connection between the terminology of the DS theory and that used in the context of this paper.

1) An Introduction to the Dempster–Shafer Theory: In the DS theory, the set of all possible outcomes in a random experiment is called the *frame of discernment* (FOD), usually denoted by Θ. For example, if we roll a die, the set of outcomes could be described by a set of statements of the form: "the number showing is i," where $1 \leq i \leq 6$; therefore, $\Theta = \{1, 2, 3, 4, 5, 6\}$. The $2^{|\Theta|}$ subsets of Θ are called propositions and the set of all propositions is denoted by 2^{Θ}. In the die example, the proposition "the number showing is even" would be represented by the set $\{2, 4, 6\}$.

In the DS theory, *probability masses* are assigned to propositions, i.e., to subsets of Θ. This is a major departure from the Bayesian formalism in which probability masses can be assigned only to singleton subsets (i.e., elements) of Θ. The interpretation to be given to the probability mass assigned to a subset of Θ is that the mass is free to move to any element of the subset. Under this interpretation, the probability mass assigned to Θ represents ignorance, since this mass may move to any element of the entire FOD. When a source of evidence assigns probability masses to the propositions represented by subsets of Θ, the resulting function is called a *basic probability assignment* (bpa). Formally, a bpa is function $m:2^{\Theta} \rightarrow [0,1]$ where

$$0.0 \leq m(\cdot) \leq 1.0 \qquad m(\varnothing) = 0$$

and

$$\sum_{X \subseteq \Theta} m(X) = 1.0. \tag{1}$$

Subsets of Θ which are assigned nonzero probability mass are said to be *focal elements of* $m(\cdot)$. The *core* of $m(\,)$ is the union of its focal elements.

A belief function, Bel(X), over Θ is defined by

$$\text{Bel}\,(X) = \sum_{Y \subseteq X} m(Y). \tag{2}$$

In other words, our belief in a proposition X is the sum of probability masses assigned to all the propositions which imply X (including X itself).

Dempster's rule of combination states that two bps'a, $m_1(\cdot)$ and $m_2(\cdot)$, corresponding to two independent sources of evidence, may be combined to yield a new bpa $m(\cdot)$ via

$$m(X) = K \sum_{X_i \cap X_j = X} m_1(X_i) m_2(X_j) \tag{3}$$

where

$$K^{-1} = 1 - \sum_{X_i \cap X_j = \varnothing} m_1(X_i) m_2(X_j). \tag{4}$$

This formula is commonly called *Dempster's rule* or *Dempster's orthogonal sum.* In this paper, we will also use the notation

$$m = m_1 \oplus m_2 \tag{5}$$

to present the combination of $m_1(\cdot)$ and $m_2(\cdot)$.

Since Dempster's rule may only be applied to bpa's which have the same domain (i.e., bpa's which discern the same

frame), if $m_1(\cdot)$ and $m_2(\cdot)$ discern different frames (i.e., $\Theta_1 \neq \Theta_2$), they must be mapped to a common frame before they can be combined. As will be clear from the next section, each sensory operation will have a unique frame of discernment. Therefore, before beliefs in pose/identity hypotheses can be modified by combining the results of different sensory operations, their individual frames of discernment must be mapped to a common frame. The process of mapping disparate frames of discernment to a common frame is called *refining* by Shafer, and the common frame thus obtained is called a *refinement*.

Refining the frames of discernment Θ_1, Θ_2, $\cdots$ to a common frame Ω is accomplished by specifying the mapping functions

$$\omega_i : 2^{\Theta_i} \rightarrow 2^{\Omega} \tag{6}$$

which must possess the following properties:

$$\omega_i(\{\theta\}) \neq \varnothing, \qquad \text{for all } \theta \in \Theta_i \tag{7}$$

$$\omega_i(\{\theta\}) \cap \omega_i(\{\theta'\}) = \varnothing, \qquad \text{for } \theta \neq \theta' \tag{8}$$

$$\bigcup_{\theta \in \Theta_i} \omega_i(\{\theta\}) = \Omega. \tag{9}$$

Equation (7) states that any proposition that is discerned in Θ_i must be discernible in Ω. Equation (8) requires that the mapped propositions in Ω be disjoint. Finally, (9) specifies that if Ω is a refinement of Θ_i, then no proposition in Ω can be outside the range of mappings corresponding to the different propositions in Θ_i.

To assess belief in a proposition in Ω, the beliefs represented by $m_i(\cdot)$ must be mapped to beliefs in subsets of Ω. This is accomplished using

$$m_i'(\omega_i(A)) = m_i(A) \tag{10}$$

where the bpa $m_i'(\cdot)$ maps $m_i(\cdot)$'s belief in a subset of Θ_i to belief in the corresponding subset of Ω.

2) Object Hypotheses and their Representation Using DS Terminology: In our application, an experiment consists of extracting features from sensory data and matching those sensed features to features of model objects. The possible outcomes of such an experiment are sets of possible matches between sensed and model features. For example, if N sensed features, $S_1 \cdots S_N$, have been extracted, the possible outcomes are of the form

$$\theta_i = \{S_1/f_1^i, S_2/f_2^i, \cdots, S_N/f_N^i\} \tag{11}$$

where the jth element of θ_i indicates that the sensed feature S_j is matched to model feature f_j^i. In other words, f_j^i denotes the model feature which is matched to sensed feature S_j in object hypothesis θ_i. For example, the four hypotheses shown in Fig. 3 are represented by $\{S1/1, S2/2\}$, $\{S1/a, S2/b\}$, $\{S1/4, S2/1\}$, and $\{S1/f, S2/a\}$.

In our system, such a set of feature matches defines an object hypothesis. This representation for an object hypothesis is explicit about the identity of the object—the model features matched in the hypothesis must belong to the object—and is implicit about the pose of the object, assuming, of course, that hypothesis contains a sufficient number of feature matches to estimate the object's pose (this will be discussed in Section VI).

A frame of discernment will be the set of all possible object hypotheses for a particular set of sensed features.

$$\Theta = \{\theta_1, \theta_2, \cdots\}. \tag{12}$$

Clearly, if there is only a single sensed feature, say S_k, the frame of discernment reduces to

$$\Theta = \{\{S_k/f_k^1\}, \{S_k/f_k^2\}, \cdots, \{S_k/f_k^n\}\} \tag{13}$$

where each element of Θ indicates a possible match of sensed feature S_k to some model feature f_k^i. In the case of a single sensed feature, we simplify this notation to the form

$$\Theta = \{S_k/f_k^1, S_k/f_k^2, \cdots, S_k/f_k^n\}. \tag{14}$$

To illustrate, consider the example discussed in Section II. Assume that the two sensed edges, $S1$ and $S2$, correspond to the edges 1 and 2, as shown in the upper left of Fig. 3. Further, assume that belief is assigned to feature matches for the two edges based on the difference in lengths of the sensed and model edges. Since the model features 1, a, and e are of same length, there is no evidence that discriminates between them; in other words, we are ignorant about which one of these three might actually be in the scene. The three must therefore be grouped together into a single proposition—something that would not be allowed in a Bayesian formalism but is easily accomplished in the DS formalism. Similarly, the features 2, 4, b, c, d, and f are identical and must be incorporated in a single proposition. Therefore, a reasonable bpa for $S1$ is

$$\begin{aligned} m_1(A_1) &= 0.5 \\ m_1(A_2) &= 0.45 \\ m_1(A_3) &= 0.05 \end{aligned} \tag{15}$$

where

$$\begin{aligned} A_1 &= \{S1/1, S1/a, S1/e\} \\ A_2 &= \{S1/2, S1/4, S1/b, S1/c, S1/d, S1/f\} \\ A_3 &= \{S1/3\}. \end{aligned} \tag{16}$$

Similarly, for $S2$ a likely bpa is

$$\begin{aligned} m_2(B_1) &= 0.45 \\ m_2(B_2) &= 0.5 \\ m_2(B_3) &= 0.05 \end{aligned} \tag{17}$$

where

$$\begin{aligned} B_1 &= \{S2/1, S2/a, S2/e\} \\ B_2 &= \{S2/2, S2/4, S2/b, S2/c, S2/d, S2/f\} \\ B_3 &= \{S2/3\}. \end{aligned} \tag{18}$$

In this example, the focal elements of $m_1(\cdot)$ are A_1, A_2, and A_3. The core of $m_1(\cdot)$ is $A_1 \cup A_2 \cup A_3$. The belief in the proposition represented by $A_1 \cup A_2$ is equal to 0.95 (the sum of the basic probability masses for A_1 and A_2), and the belief in the proposition represented by $\{S1/a\}$ is 0, since no subset of $\{S1/a\}$ is assigned a nonzero basic probability mass. At first glance, this may seem unusual, since $\{S1/a\}$ is a subset of A_1, and A_1 has a positive belief value. However, the bpa $m_1(\cdot)$ does not give us any evidence for $\{S1/a\}$ as the correct assignment for $S1$. It merely says that the assignments $\{S1/1, S1/a, S1/e\}$ are equally valid, and that there is no evidence available to discriminate between the three.

In order to combine $m_1(\cdot)$ and $m_2(\cdot)$, their respective frames of discernment must be mapped to a common frame. This is done by specifying two mapping functions, $\omega_1(\cdot)$ and $\omega_2(\cdot)$, which satisfy (7)–(9). For this example, a valid refinement can be obtained by collecting every pair of matches for $S1$ and $S2$, i.e., Ω could be defined as

$$\Omega = \{\{\theta_1, \theta_2\} | \theta_1 \in \Theta_1 \text{ and } \theta_2 \in \Theta_2\} \tag{19}$$

or

$$\begin{aligned}\Omega = \{&\{S1/1, S2/1\}, \{S1/1, S2/2\}, \{S1/1, S2/3\},\\ &\{S1/1, S2/4\}, \{S1/1, S2/a\}, \cdots\\ &\{S1/2, S2/1\}, \{S1/2, S2/2\}, \{S1/2, S2/3\},\\ &\{S1/2, S2/4\}, \{S1/2, S2/a\}, \cdots\\ &\cdots \{S1/f, S2/d\}, \{S1/f, S2/e\}, \{S1/f, S2/f\}\}.\end{aligned}$$

Included in Ω are object hypotheses that would be impossible for the case of single object recognition; for example, the hypothesis $\{S1/2, S2/a\}$ is of this type since model features 2 and a do not belong to the same object. Later we will show how such unlikely hypotheses can be pruned away by enforcing appropriate constraints.

The mapping functions $\omega_1(\cdot)$ and $\omega_2(\cdot)$ that produce the common refinement of (19) would be given by

$$\omega_i(\theta) = \{\omega | \omega \in \Omega \text{ and } \theta \in \omega\} \tag{20}$$

where θ is a feature match from Θ_i and Ω is defined by (19). For example, the belief assigned by $m_1(\cdot)$ to the match $S1/3$ would be mapped to the subset of Ω

$$\begin{aligned}\omega_1(\{S1/3\}) = \{&\{S1/3, S2/1\}, \{S1/3, S2/2\},\\ &\{S1/3, S2/3\}, \{S1/3, S2/4\},\\ &\{S1/3, S2/a\}, \{S1/3, S2/b\},\\ &\{S1/3, S2/c\}, \{S1/3, S2/d\},\\ &\{S1/3, S2/e\}, \{S1/3, S2/f\}\}.\end{aligned} \tag{21}$$

That is, the belief assigned to the match $S1/3$ by $m_1(\cdot)$ in the frame Θ_1 will be mapped to the subset of Ω which contains all of the propositions that match sensed edge $S1$ to model edge 3. Similarly, the belief assigned by $m_2(\cdot)$ to the match $S2/3$ would be mapped to the subset of Ω

$$\begin{aligned}\omega_2(\{S2/3\}) = \{&\{S1/1, S2/3\}, \{S1/2, S2/3\},\\ &\{S1/3, S2/3\}, \{S1/4, S2/3\},\\ &\{S1/a, S2/3\}, \{S1/b, S2/3\},\\ &\{S1/c, S2/3\}, \{S1/d, S2/3\},\\ &\{S1/e, S2/3\}, \{S1/f, S2/3\}\}.\end{aligned} \tag{22}$$

Now the combination rule can be used to find combined belief in a proposition in Ω. For example, since

$$\omega_1(A_3) \cap \omega_2(B_3) = \{\{S1/3, S2/3\}\} \tag{23}$$

the belief in the proposition $\{\{S1/3, S2/3\}\}$ obtained by combining $m_1(\cdot)$ and $m_2(\cdot)$ is found to be

$$\begin{aligned}&m_1'(\omega(\{S1/3\}))m_2'(\omega(\{S2/3\}))\\ &\quad = m_1(\{S1/3\})m_2(\{S2/3\}) = 0.0025.\end{aligned} \tag{24}$$

Note that the normalizing constant K is unity in this example, since $\omega_1(A_i) \cap \omega_2(B_j) \neq \emptyset$ for any i, j.

B. Generating and Refining Hypothesis Sets

In our earlier work [16], hypothesis generation and subsequent refinement did not use an evidential reasoning scheme. Each sensed feature was matched to all feasible model features (where feasibility was determined by the similarities of the attributes of the sensed and model features). These matches were then pruned by enforcing object consistency, relational constraints, and aspect consistency. Object consistency ensured that all model features participating in an object hypothesis belonged to the same model object. Relational consistency was determined by examining the similarity of the relationships between the sensed features and the corresponding relationships between the matched model features. If the similarity was below some quantitative threshold, the hypothesis was discarded. Aspect consistency ensured that prominent object features were matched if they could be observed by the performed sensing operation. If they were not matched, the corresponding hypothesis was discarded.

Our current system retains feature matches, object consistency, relational consistency, and aspect consistency as the four measures of an object hypothesis' credibility, but, thresholding has been replaced by evidential reasoning. Now, a hypothesis is assigned belief which reflects how well the four criteria are satisfied. We use four bpa's: $m_f(\cdot)$, $m_o(\cdot)$, $m_r(\cdot)$, and $m_a(\cdot)$ to assign beliefs to object hypotheses based on the quality of the feature matches, object consistency, relational consistency, and aspect consistency. We combine these using Dempster's rule of combination to determine the aggregate belief in an object hypothesis. Thus, when the sensing system extracts a set of features, a set of object hypotheses is constructed by deriving $m_f(\cdot)$, $m_o(\cdot)$, $m_r(\cdot)$, and $m_a(\cdot)$, and combining them. If a set of object hypotheses already exists, the new belief function must be combined with the belief function for the existing hypothesis set to produce the revised

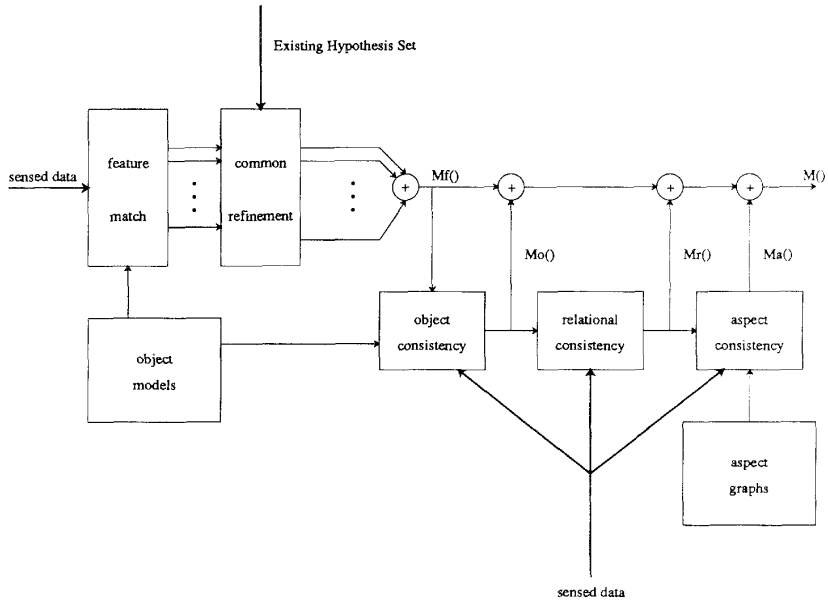


Fig. 8. Block digram of hypothesis generation/refinement system.

hypothesis set and the associated beliefs. Fig. 8 shows a block diagram of the hypothesis generation/refinement system.

1) Consistency of Feature Matches: When a set of features is extracted from sensory data, the first step in generating a hypothesis set (or refining the current hypothesis set, if it exists) is to match those sensed features to model features and derive a belief function which expresses the belief in each object hypothesis that can be derived from those matches. We do this in two steps. First, individual belief functions are derived for each sensed feature. These belief functions define the possible matches between sensed and model features and the corresponding evidence which supports those matches. These individual belief functions are then combined to form object hypotheses which represent possible combinations of the feature matches.

The process of assigning belief to individual feature matches consists of comparing attributes of the sensed features to attributes of the model features. It is possible that a number of model features will have attributes that are exactly the same. For example, edges 2, 4, b, c, d, and f of the objects shown in Fig. 1 are exactly the same. Our system groups together model features which appear equivalent, each such grouping corresponding to one unique model feature. Each unique model feature is given a label. For example, the edges 2, 4, b, c, d, and f of the objects shown in Fig. 1 could be grouped together and assigned the label M_{u1}. The set of labels for all unique model features is denoted by U. The set of all model features is denoted by M. A function, $u{:}U \rightarrow 2^M$ is used to map unique model features onto the appropriate subsets of M.

For each sensed feature S_i, we construct a bpa $m_i(\cdot)$ which represents the system's belief in the possible matches for S_i. This bpa is constructed as follows. If $u(F) = \{M_1, \cdots, M_k\}$, then for

$$\theta = \{S_i/M_1, \cdots, S_i/M_k\} \tag{25}$$

we define

$$m_i(\theta) = \frac{C(F, S_i)}{\sum_{x \in U} C(x, S_i)} \tag{26}$$

where $C(F, S_i)$ is the confidence value associated with matching unique model feature F to sensed feature S_i.

As an example, for the unique feature M_{u1}, $u(M_{u1}) = \{2, 4, b, c, d, f\}$. Then, for sensed edge $S1$, we derive one value of $m_1(\cdot)$ by

$$\theta = \{S1/2, S1/4, S1/b, S1/c, S1/d, S1/f\}$$

and

$$m_1(\theta) = \frac{C(M_{u1}, S1)}{\sum_{x \in U} C(x, S1)}.$$

Examples of the methods used to calculate the value of $C(\cdot)$ will be presented in Section VII-A.

Once bpa's have been assigned to represent belief in individual feature matches, the next step is to combine the feature matches to create object hypotheses, and to combine the bpa's to determine the belief in those object hypotheses. Unfortunately, as discussed in Section IV-A, this cannot be done by simply invoking Dempster's combination rule, since the individual bpa's do not discern the same frame. Recall, for each sensed feature S_i, we have a bpa $m_i(\cdot)$, which discerns an FOD Θ_i, whose structure is shown in (14). To remind the reader, Θ_i only includes propositions about the ith sensed feature and, therefore, $m_i(\cdot)$ only discerns propositions about the ith sensed feature. In order to combine these individual bpa's, we must create an FOD which is common to all the features sensed so far.

Given N sensed features, we construct Ω as follows:

$$\Omega = \{\{\theta_1, \theta_2, \cdots, \theta_N\} \mid \theta_i \in \Theta_i\}. \quad (27)$$

In keeping with the discussion in Section IV-A2, each element of Ω is a collection of feature matches, and each possible combination of feature matches (for the N sensed features) is represented in Ω.

We can also define $\omega_i(\cdot)$, the refining from Θ_i to Ω as

$$\omega_i(\{S_i/M_j\}) = \{\phi \mid \phi \in \Omega \text{ and } S_i/M_j \in \phi\} \quad (28)$$

for singleton subsets of Θ_i, and

$$\omega_i(A) = \bigcup_{\theta \in A} \omega_i(\{\theta\}) \quad (29)$$

for $A \subset \Theta$. In other words, $\omega_i(\{S_i/M_j\})$ is the subset of Ω that contains all object hypotheses which match sensed feature S_i to model feature M_j.

Now, we can apply Dempster's rule

$$m_f = m_1' \oplus m_2' \oplus \cdots m_N' \quad (30)$$

where $m_i'(\cdot)$ is computed as in (10).

The method used for refining an existing hypothesis set given a new sensed feature is similar to the process just described. Denote the existing hypothesis set by Ω_{k-1} (where k is the number of sensed features which have been encountered thus far), and the new sensed feature by S_k. First, we find $m_k(\cdot)$, which defines the set of beliefs in possible feature matches for sensed feature S_k. Next, we create a common refinement of Ω_{k-1} and Θ_k (where Θ_k is the frame discerned by $m_k(\cdot)$). This refinement is defined by

$$\Omega = \{\phi \cup \{\theta\} \mid \phi \in \Omega_{k-1} \text{ and } \theta \in \Theta_k\}. \quad (31)$$

Note that each element of Ω_{k-1} will be an object hypothesis which is represented as a set of feature matches. Therefore, $\phi \cup \{\theta\}$ in the equation above is a new object hypothesis, which contains the feature matches from one object hypothesis in Ω_{k-1} and one of the feature matches for S_k. The mappings from Ω_{k-1} and Θ_k to Ω are analoguous to those defined in (28) and (29). In particular

$$\omega_{\Omega_{k-1}}(\rho) = \{\phi \mid \phi \in \Omega \text{ and } \rho \subset \phi\} \quad (32)$$

and

$$\omega_k(\{S_k/M\}) = \{\phi \mid \phi \in \Omega \text{ and } S_k/M \in \phi\}. \quad (33)$$

2) Object Consistency: It is quite likely that some of the hypotheses in the frame discerned by $m_f(\cdot)$ will match sensed features to model features that are not in the same object. Since we do not currently deal with occluding objects, we do not allow such matches. (This restriction will be removed if we later allow for occlusion.) This constraint is enforced by combining the bpa $m_f(\cdot)$ with $m_o(\cdot)$.

We want $m_o(\cdot)$ to place all of its belief in the subset of hypotheses which contain only consistent matches, and no belief in any hypothesis which contains an inconsistent match. A hypothesis contains an inconsistent match if any two sensed features are matched to model features from different objects. Thus for a hypothesis set Ω, we define

$$m_o(\phi) = \begin{cases} 1, & \text{for the largest } \phi \subseteq \Omega \text{ s.t. } \phi \text{ contains no inconsistent matches} \\ 0, & \text{otherwise.} \end{cases} \quad (34)$$

3) Relational Consistency: Further pruning of the set of object hypotheses can be achieved by enforcing relational constraints. An example of a relational constraint would be the equality of the dihedral angles between planar features in the scene and the corresponding model features in an object hypothesis. Most previous approaches to robot vision have treated such constraints in a deterministic manner, meaning that a relational constraint is considered either satisfied or not satisfied depending upon whether or not the value of the relation between the scene features is within a prescribed range (which depends on the value of the relation in the model). The system presented in this paper is more general, in that the belief it associates with a given object hypothesis is made to depend on the degree of similarity between the scene relations and their corresponding model object relations.

We enforce relational constraints by constructing a new bpa, $m_r(\cdot)$, which is a combination of a number of bpa's, one for each type of relation. For example, one component of $m_r(\cdot)$ is the bpa $m_n(\cdot)$, which assigns beliefs on the basis of the similarity of the angle between the surface normals for two planar scene features and the angle between their corresponding model features. In Section, VII-A we will give examples of bpa's used to compute m_r.

When such an $m_r(\cdot)$ is combined with $m_f \oplus m_o$, the result is a weakening or elimination of object hypotheses in which the relations between sensed features do not match well with the relations between the corresponding model features.

4) Aspect Consistency: The final bpa used to evaluate the quality of an object hypothesis is based on the fact that, once a position transformation has been derived for the hypothesis, the system can determine which object features should be observed from a particular viewpoint. The bpa $m_a(\cdot)$ is derived by accumulating positive evidence when expected features are matched in the hypothesis, the exact degree of belief being a function of the quality of the feature match and the likelihood that the feature will be extracted.

As we will discuss in Section VI, it is possible to derive a position transformation for an object hypothesis which contains a sufficient number of feature matches. This transformation is used by the function $A(\theta, V)$, to determine the aspect of the object which would be observed from a certain viewpoint V for a particular hypothesis θ. The function $F_a(x)$ returns the set of features visible in aspect x. Thus $F_a(A(\theta, V))$ returns the set of features which should be visible from a viewpoint V, given the object hypothesis represented by θ. Associated with each aspect of an object is a set of weights which reflect the prominence of each feature in the aspect. The

function $w_A(x, y)$ returns the weight assigned to model feature x in aspect y.

To determine the quality of the match for model feature f in object hypothesis θ, the system first determines which sensed feature S_i is matched to model feature f in θ. Then $m_i(\cdot)$ (the bpa which assigns belief to feature matches for the ith sensed feature) is invoked to determine how much belief is placed in the proposition which includes the match of S_i to f. This is expressed by the function $q(\cdot)$

$$q(f, \theta, V) = \begin{cases} m_i(\phi), & \text{for } S_i/f \in \theta, \text{ where } S_i/f \in \phi \text{ and } S_i \text{ was observed from } V \\ 0, & \text{otherwise} \end{cases} \tag{35}$$

which returns the quality of the match for model feature f in object hypothesis θ. Note that if feature S_i were not observed from viewpoint V, or if f were not matched in θ, $q(\cdot)$ would equal 0.

We define the aspect confidence in an object hypothesis to be

$$C_a(\theta, V) = \sum_{f \in F_a(A(\theta, V))} w_A(f, A(\theta, V)) q(f, \theta, V). \tag{36}$$

This equation states that aspect consistency is evaluated by summing the product of a feature's likelihood of being extracted from sensory data with the quality of the match for that feature, for each feature that we expect to find in the hypothesized aspect. We construct $m_a(\cdot)$ by normalizing $C_a(\cdot)$. (See Note Added in Proof.)

A slight complication arises when an object hypothesis contains sensed features that were observed from different viewpoints. In such cases, each sensory operation contributes its own $m'_a(\cdot)$, based only on the features that it extracted. These individual $m'_a(\cdot)$'s are then combined to form $m_a(\cdot)$.

To illustrate the calculation of aspect consistency, consider again the example of Section II. Suppose that the current object hypothesis θ is $\{S1/a, S2/b\}$, and the viewpoint $V1$ is as shown in Fig. 3. This implies that the aspect being viewed is aspect B (see Fig. 7), i.e., $A(\theta, V1) = B$, and that $F_a(A(\theta, V1)) = \{a, b\}$. Now, if the weights for edges a and b in aspect B are set to 0.55 and 0.45, respectively, the equation for $C_a(\theta, V1)$ becomes

$$\begin{aligned} C_a(\theta, V1) &= w_A(a, B) q(a, \{S1/a, S2/b\}, V1) \\ &\quad + w_A(b, B) q(b, \{S1/a, S2/b\}, V1) \\ &= 0.55\, m_1(A_1) + 0.45\, m_2(B_2) \end{aligned}$$

where A_1 and B_2 are as defined in Section II, since $S1/a \in A_1$, $S1/a \in \theta$, $S2/b \in B_2$, $S2/b \in \theta$, and both $S1$ and $S2$ were observed from $V1$.

C. A Polynomial Time Implementation of the Combination Rule

It is well known that a brute-force implementation of Dempster's combination rule has worst case behavior that is exponential in the size of the frame of discernment (or the size of the hypothesis set), since all subsets of the frame must be examined. Fortunately, the structure of the belief functions which our system creates allows for a special implementation of Dempster's rule. In this section, we prove that our use of Dempster's rule can be achieved in polynominal time (in the size of the hypothesis set).

In order to show this, we will introduce a class of belief functions which we will call *disjoint belief functions*. All belief functions that are created by our system belong to this class (which will be evident once the definition of disjoint belief functions is stated). We will then show that the combination of two disjoint belief functions produces a disjoint belief function. Thus all belief functions that are encountered by our system, whether created directly from sensory measurements or by combining two existing belief functions, will belong to the class of disjoint belief functions. Finally, we will show that the combination of any two disjoint belief functions can be performed in polynominal time.

Def: We will say that a belief function, Bel, over the frame of discernment Θ is disjoint if its corresponding bpa is such that for all $A, B \subseteq \Theta$, if $m(A) > 0$, $m(B) > 0$, and $A \neq B$, then $A \cap B = \varnothing$.

This condition is equivalent to the statement that the subsets of Θ with positive basic probability masses form a disjoint partition of the core of Bel. (Remember that the core of Bel is the union of its focal elements, and that X is a focal element of Bel iff $m(X) > 0$.) It is clear that all belief functions derived by our system are disjoint belief functions, since the corresponding bpa's are constructed by assigning confidence values to disjoint subsets of a frame of discernment and then normalizing those confidence values. The system never constructs a bpa that assigns positive probability numbers to two nondisjoint subsets of the frame of discernment.

For convenience, we introduce one further definition.

Def: Given two belief functions with corresponding bpa's m_1 and m_2, we say that A and B form a supporting pair of C if $A \cap B = C$ and $m_1(A) > 0$, $m_2(B) > 0$.

This definition is a consequence of the fact that, in the combination rule, two subsets, A and B, contribute to the belief in exactly the subset C iff $A \cap B = C$, $m_1(A) > 0$, and $m_2(B) > 0$.

We now state and prove a lemma which will be used in the proof of our basic theorems.

Lemma: Let Bel_1 and Bel_2 be two disjoint belief functions. If their combination, Bel, has the corresponding bpa m, then if $m(C) > 0$, there is exactly one supporting pair of C.

Proof: Let A,B, and C be such that A,B is a supporting pair of C and C is nonempty. Now, also suppose that X,Y is a supporting pair of C. Since A,B is a supporting pair of C, then $A \cap B = C$ which implies that C is a subset of A. Likewise, C must also be a subset of X. But, since Bel_1 is disjoint,

$A \cap X = \emptyset$ (by the definition of disjoint belief functions, and since both A and X are focal elements of Bel). Thus since C is contained in both A and X, $C = \emptyset$, which is a contradiction since $m(C) > 0$ implies that $C \neq \emptyset$.

Q.E.D.

We now state our first theorem.

Thm: If two belief functions are disjoint, then their combination is also disjoint.

Proof: We will prove the theorem by showing that, if Bel is the combination of two disjoint belief functions with corresponding bpa m, if $m(X) > 0$ and $m(Y) > 0$ then $X \cap Y = \emptyset$.

If $m(X) > 0$, then by the lemma, there is exactly one supporting pair for X. Call this pair A_i, B_j. Similarly, if $m(Y) > 0$, Y will have exactly one supporting pair, say A_k, B_l. Now, by the definition of disjoint belief functions and supporting pairs (in particular that any two nonidentical focal elements of a disjoint belief function have a null intersection and that both elements of a supporting pair are focal elements of their respective belief functions) we can assert that either $A_i = A_k$, or $A_i \cap A_k = \emptyset$, and either $B_j = B_l$, or $B_j \cap B_l = \emptyset$. To see this, we examine the intersection of X and Y.

$$X \cap Y = (A_i \cap B_j) \cap (A_k \cap B_l).$$

Since set intersection is both associative and commutative

$$X \cap Y = (A_i \cap A_k) \cap (B_j \cap B_l).$$

If $A_i \neq A_k$, this intersection is empty since $A_i \cap A_k = \emptyset$. Similarly, if $B_j \neq B_l$, the intersection is empty. If $A_i = A_k$ and $B_j = B_l$, then $X = Y$. Thus if $m(X) > 0$ and $m(Y) > 0$ either $X = Y$ or $X \cap Y = \emptyset$, and therefore Bel is disjoint.

Q.E.D

An important consequence of this theorem is that, provided the system creates only disjoint belief functions, all belief functions which it derives by combining two existing belief functions will also be disjoint. Thus all applications of Dempster's rule in our system will be to combine disjoint belief functions. The following theorem states that such combinations can be achieved in time polynominal in the size of the frame of discernment.

Thm: If Bel_1 and Bel_2 are disjoint belief functions, then the basic probability numbers for every focal element of their combination, Bel, can be calculated in time polynominal in the size of the frame of discernment Θ.

Proof: We prove this theorem by showing that we can enumerate the focal elements of Bel in polynominal time and that we can find $m(A)$ in polynominal time, for each A which is a focal element of Bel.

- By the lemma, $m(C) > 0$ implies that there is exactly one supporting pair for C. We can find all $C = A \cap B$ with $m(C) > 0$ by examining every pair A,B such that $m_1(A) > 0$ and $m_2(B) > 0$. There are at most $|\Theta|^2$ such pairs, since Bel_1 and Bel_2 are disjoint. Therefore, the focal elements of Bel can be enumerated in polynominal time.
- In order to show that $m(C)$ can be found in polynominal time for any focal element of Bel, consider the form of the combination rule. We can evaluate the numerator by examining all pairs A,B such that $m_1(A) > 0$ and $m_2(B) > 0$ in order to find the supporting pair of C. As above, this leads to at most $|\Theta|^2$ set intersections. For the denominator, we must examine all pairs A,B such that $A \cap B = \emptyset$ and $m_1(A) > 0$ and $m_2(B) > 0$. Again, this can be accomplished in at most $|\Theta|^2$ set intersections.

Q.E.D.

V. Measuring the Ambiguity in a Set of Hypotheses

Now that we have described how sets of objects hypotheses are generated and subsequently refined to admit new evidence, we need a means of characterizing the ambiguity in a hypothesis set. In our earlier work [16], the ambiguity in a hypothesis set was trivially defined to be the number of hypotheses in the set. Of course that approach will not work once an evidential reasoning scheme is put into place. Consider, for example, a case in which no initial hypothesis is ever completely discounted although eventually a single hypothesis accrues enough evidence to emerge as the obvious choice. Clearly a more sophisticated measure of ambiguity is needed.

Before defining our measure of ambiguity, we enumerate the qualities that it should possess. If we have a set of hypotheses, with an associated bpa $m(\cdot)$, we want to characterize the amount of choice that the system would be required to exercise in order declare a single hypothesis as valid. The more choice required, the higher the amount of ambiguity. Thus our measure of ambiguity should be highest when all hypotheses are equally likely. Stated another way, given two hypothesis sets, the set whose belief function shows the greater dispersion should have more ambiguity (by dispersion, we mean the degree to which a belief function resembles a uniform distribution). Furthermore, if all hypotheses are equally likely, the ambiguity should increase with the number of hypotheses. Of course, if a hypothesis set has a single element, then its ambiguity should be zero.

Another desirable quality for a measure of ambiguity is that it be consistent across levels of a hierarchical hypothesis space. In particular, if we establish hypothesis sets in a hierarchy, then the ambiguity in a hypothesis set at one level should be equal to a weighted sum of the ambiguity in its descendants. For example, if the top level hypothesis set H_0 is the set $\{A,B\}$, with $m_0(\{A\}) = 0.3$, and $m_0(\{B\}) = 0.7$, and we split A and B to obtain two new hypothesis sets $H_1 = \{a_1, a_2\}$, and $H_2 = \{b_1, b_2\}$, then the ambiguity in H_0 should be equal to the sum of 0.3 times the ambiguity in H_1 and 0.7 times the ambiguity in H_2.

The only continuous function satisfying these requirements is of the form

$$A(\Omega) = -K \sum_{\theta \in \Omega} \Pr(\theta) \log \Pr(\theta) \tag{37}$$

where K is some positive constant, and $\Pr(\theta)$ is a measure of the certainty that θ is the correct hypothesis. A proof of this can be found in [26]. The form of $A(\cdot)$ is not totally unfamiliar. It is the form of the entropy measure from information theory. This is no mere coincidence, since information theorists use entropy to measure the freedom of choice available in selecting a message, provided that the probabilities associated with the choices are known.

Other work on characterizing the entropy in a hypothesis set has been done by Stephanou and Lu [27], Yager [29], and Higashi and Klir [13]. The measure described in [27] does not suit our purposes because it awards equal entropy to hypothesis sets with different numbers of elements in the case of total ignorance (i.e., the belief function assigns belief of 1.0 to the total frame, and no belief to any subset of the frame). The measure developed in [29] fails to meet our criteria because if any two focal elements have a nonempty intersection, the entropy is zero. Finally, the entropy measure described in [13] fails to satisfy our condition that the entropy be consistent over levels in a hierarchy of hypothesis spaces.

Note that in (37) we did not use $m(\cdot)$ to represent the likelihood that a particular hypothesis was correct. This is because there will be situations in which $m(\cdot)$ assigns belief to nonsingleton subsets of Ω, and no belief to individual hypotheses. In such cases, we must still be able assess the likelihood of the individual hypotheses. For this purpose, we calculate $\Pr(\theta)$ as follows:

$$\Pr(\theta) = \sum_{\theta \in A} \frac{m(A)}{|A|} . \tag{38}$$

In this way, when $m(\cdot)$ assigns belief to a nonsingleton subset of Ω, for the purpose of calculating ambiguity, we treat the individual elements of that subset as being equally likely.

In order to apply this measure of ambiguity to the problem of selecting a best next sensing operation, we predict the hypothesis sets which might occur if a particular sensing operation were applied. We then find the ambiguity associated with each of these possible hypothesis sets, and use the worst case value as a measure of the effectiveness of that sensing operation. We will use the symbol $A^i_{\max}$ to refer to the maximum ambiguity associated with a proposed ith sensing operation. The goal of the system is then to choose a sensing operation which minimizes the value of $A^i_{\max}$.

VI. Predicting the Possible Results of Sensing Operations

In order to determine which sensing strategy will minimize $A^i_{\max}$, we must be able to predict the possible results of candidate sensing operations. Consider Figs. 4 and 5. The ability to determine the best viewpoint depended on the ability to predict the set of edges that would be observed from the various viewpoints. This amounted to being able to predict the geometry of the line segments that would be observed from various viewpoints relative to the edges in the four object hypotheses. In the general case, predicting sensor readings depends on the ability to determine the features that will be observed by a particular sensor from a particular viewpoint for each object hypothesis in the current hypothesis set.

In order to have this predictive power, the system must be able to derive the position transformations implicitly contained in an object hypothesis (which is composed of a set of feature matches). This can be done once the position (location in 3-space and orientation) of any of the sensed features has been completely determined. Generally, if only one sensed feature has been observed, it is not possible to accurately determine both the location and orientation of that feature due to segmentation errors, occlusion of surfaces, noisy edge detection, etc. For example, if a planar surface is observed, it is possible to robustly determine the normal to the surface, but not necessarily the rotation of the surface about the normal, since this depends on the ability to robustly determine the locations of the edges of the surface. After two features have been observed, the chances of accurately determining the position transformation are much better. For example, once two adjacent planar faces have been found, their two surface normals fix the object's orientation; this formed the basis of the pose transformation procedure reported in [16] using the algorithm of [7]. It is our intention to incorporate in the system more sophisticated methods for position transformation calculations, such as those presented in [2], which are based on the principles introduced by Hebert and Faugeras [6].

Once a position transformation has been computed for a hypothesized object, it is a simple matter to determine the set of features which should be observed by a sensing operation from a particular viewpoint (provided that the object hypothesis is correct). This is done as follows. First, the position transformation T_{obj} is used to determine the tessel on the viewing sphere from which the object will be observed. The viewing tessel is the tessel intersected by the vector $T^{-1}_{\text{obj}} V$ (where V is the viewpoint in world coordinates). The set of features visible from this tessel is then intersected with the set of features which can be observed by the particular sensor. This determines the set of features visible to the sensor from the specified viewpoint.

We now turn our attention to the features themselves. In particular, we will now describe what features are used by the system for each of the sensing modalities and how those features are derived from sensory data.

VII. Observable Features

In this section, we describe the features which can be observed by each of the sensors. Currently, the sensors in our work cell include a structured light scanner to obtain 3-D information about the scene, overhead and a manipulator held cameras to obtain 2-D information about the scene, a force/torque sensor mounted on the robots's wrist, and a manipulator which can be queried to find the distance between its fingers.

A. 3-D Features

The richest set of features available to the system comes from range data. Range data are gathered for a set of points in

the scene using a single stripe, structured light range scanner which the robot manipulates. This initial data are converted to x, y, z data. Subsequent processing of these x, y, z data produces a list of surfaces, attributes of those surfaces, and relations between the surfaces. The types of attributes provided by range data processing include surface area, orientation, location, surface type, etc. Relations include adjacency, coplanarity, etc. The methods used to determine 3-D features are documented in [5], [14], [30].

Confidence in matches between model surfaces and surfaces extracted from range data are based on the similarity of the area and 3-D shape attributes of the surfaces. Specifically, we calculate the confidence in a 3-D surface match using the equation

$$C(F, S_i) = C_S(F, S_i) C_A(F, S_i). \tag{39}$$

The value used for $C_S(F, S_i)$ is the value of $p(X|\omega)$, which is determined by the 3-D shape classifier described in [14], where ω is the 3-D shape of the unique model feature F, and X is the feature vector computed for the sensed surface S_i. The value for C_A is based on the difference between the areas of the sensed surface and unique model feature, and is calculated by

$$C_A(F, S_i) = e^{-\tau(|A_F - A_{S_i}|/A_F)} \tag{40}$$

where A_F is the area of the unique model feature F, A_{S_i} is the area of sensed feature S_i, and τ is a weighting factor which is chosen empirically.

As we mentioned in Section IV-B2, for features extracted from 3-D sensory data, we use the bpa's $m_n(\cdot)$ and $m_l(\cdot)$ to construct $m_r(\cdot)$. We will now briefly describe how these are derived.

The bpa $m_n(\cdot)$ is based on the angles between surfaces. In particular, since range data processing produces an average surface normal for each surface in the sensed data, and since each surface in the object model has an associated surface normal, we can compare the angle between the surface normals of sensed surfaces to the angle between the corresponding model surfaces. In order to do this, we need to define two additional functions. For a feature match ϕ, $n_S(\phi)$ returns the surface normal of the sensed feature matched in ϕ, and $n_M(\phi)$ returns the surface normal of the model feature matched in ϕ. Note that ϕ corresponds to a feature match in a hypothesis (i.e., each element of Ω corresponds to a single object hypothesis which contains a number of matches between sensed and model features). Using these two functions, we can compute the magnitude of the difference in dot products of sensed and model surface normals as

$$E(\phi, \psi) = |n_S(\phi) \cdot n_S(\psi) - n_M(\phi) \cdot n_M(\psi)| \tag{41}$$

for ϕ and ψ in θ, and $\theta \in \Omega$. Since $E(\cdot)$ is the magnitude of the difference in two values which are in the interval [0, 1], its value will lie in the interval [0, 2], with $E(\cdot) = 0$ corresponding to an exact match, and $E(\cdot) = 2$ corresponding to the worst possible error. In order to capture the notion of conjunction, for a particular object hypothesis θ which contains N feature matches (i.e., $\theta = \{\phi_1 \cdots \phi_N\}$), we define $C_n(\theta)$ as

$$C_n(\theta) = \prod_{i=1}^{N-1} \prod_{j=i+1}^{N} (2 - |n_S(\phi_i) \cdot n_S(\phi_j) - n_M(\phi_i) \cdot n_M(\phi_j)|). \tag{42}$$

Finally, we transform $C_n(\cdot)$ into a valid bpa by normalization

$$m_n(\theta) = \frac{C_n(\theta)}{\sum_{\psi \in \Omega} C_n(\psi)}. \tag{43}$$

The bpa $m_l(\cdot)$ is derived using the fact that we can determine the correct location of a feature once a position transformation T_{obj} for the object has been computed. We use T_{obj} to derive the quality of a match between sensed and model features based on the proximity of the sensed feature to the location at which we expect to find it. For a particular feature match ϕ, the function $L_S(\phi)$ returns the location of the sensed feature matched in ϕ, and $L_M(\phi)$ returns the location of the model feature matched in ϕ. Therefore, the distance between the points $L_S(\phi)$ and $T_{\text{obj}} L_M(\phi)$ is a measure of the quality of the match expressed in ϕ. Since this distance is essentially unbounded, we place it in the exponent of a decaying exponential function to obtain

$$c(\phi) = e^{-\tau_1 |L_S(\phi) - T_{\text{obj}} L_M(\phi)|}. \tag{44}$$

We combine the $c(\cdot)$'s to obtain a confidence in the proposition θ by taking their product over the feature matches in θ

$$C_l(\theta) = \prod_{\phi \in \theta} c(\phi). \tag{45}$$

We obtain $m_l(\cdot)$ by normalizing $C_l(\cdot)$.

B. 2-D Features

The features which are visible to the 2-D camera are not nearly as robust as those visible to the range scanner. In particular, surface types typically cannot be determined from 2-D data, edge detection is not as good (since only gray-level edge detection can be used), and relationships between surfaces cannot be measured (except for adjacency). The primary advantage of 2-D vision is that it is computationally less expensive than 3-D vision. Also, since our range scanner is held by the robot, and one robot move is required for each projected light stripe, using 2-D vision reduces the number of required manipulations from the large number required to scan a scene to the much smaller number required to grasp the hand-held camera and position it at the appropriate viewpoint.

The local features (i.e., features that are confined to local areas of the object, such as a single surface or edge) that we can obtain from 2-D image processing include holes in the object, surface texture, and intensity edge information. In our current experiments, the object surfaces are all smooth, containing little or no surface texture information. Therefore, the primary 2-D features that we consider are holes and gray-level edges.

Although gray level-edge detection is not as robust as the 3-

D edge detection, it is generally much faster. Furthermore, using object hypotheses to guide the application of the edge detector, the problem is reduced from edge detection to edge verification. In particular, once we have an object hypothesis which includes a position hypothesis, we can predict the set of edges visible to the 2-D camera. If we know the camera transformation, we can predict where these edges will be found in the image plane. The image obtained from the camera can then be used to verify the presence of the edge. This edge verification is done using the Dempster–Schafer formalism applied to a binary frame of discernment (i.e., edge-present/edge-not-present) [24]. Currently, our object models do not contain edge information, and so edges are not used at this time.

In addition to using the hand-held 2-D camera to derive 2-D features, our system also uses an overhead, supervisory camera to guide the initial application of the range scanner. The supervisory camera is used in the preprocessing as follows. First, an image of the work cell is digitized. This image is subtracted from a reference image of the work cell, and the result is thresholded. This binary image is then subjected to a component labeling process. Then, the center of mass and principal axis of each of the components is computed. The center of mass is used as input to an inverse perspective transformation, which gives an approximate world location of the center of mass of the object. The inverse perspective transform is performed using the two-plane method of camera calibration [22]. Each of these operations is fairly common in the field of computer vision, therefore, we will not describe them here. The interested reader can find the details in a variety of references, including [18], [23].

C. Force/Torque Sensed Features

The last type of sensing that our system can perform is active sensing of the environment using the robot manipulator. The manipulator can be used in either of two ways. Its fingers can be closed on an object to measure its width, or, the manipulator fingers can be closed, and used as a probe. When in the latter mode, force/torque (f/t) sensing is used to execute a guarded move toward an object feature to precisely measure its height. Using these techniques, we can precisely (to within the known error of the manipulator position) measure features on the objects in the world. Like range scanning, using this type of sensing requires the active participation of the robot, thus incurring the additional overhead of planning and executing robot motions.

The utility of measuring object widths becomes evident when we have competing object hypotheses, and the difference in sizes of visible features of the two objects is less than what can be perceived by the 3-D or 2-D vision systems. Of course 2-D vision is very imprecise due to the use of the inverse perspective transform using an estimate for the world Z coordinate. Using our current range scanner, precision in 3-D data is a function of (among other things) the baseline distance between the camera and the stripe projector [5]. Furthermore, the smallest feature which can be detected using the range scanner is a function of the distance between projected light stripes. To compensate for these inaccuracies, the manipulator can be used to perform the more precise measurements only when they are required.

Measuring the height of object surfaces becomes particularly useful when those surfaces are obscured from the view of the vision systems. When cases like this arise, the vision systems are unable to observe the distinguishing features of the object. In such cases, the manipulator can be used as a probe to resolve the ambiguities. Manipulator probing can also be used to determine the existence of protrusions from object surfaces, especially when these protrusions are obscured from the view of the vision sensors (e.g., when the work piece is positioned such that it occludes the surface which has the protrusion).

Viewpoints, for both 2-D and 3-D vision sensors, lend themselves to quantization by the use of aspect graphs. However, f/t sensing is not amenable to such a representation. Furthermore, when predicting features that will be observed by the vision sensors from a particular viewpoint, the system merely uses the viewpoint in conjunction with the hypothesized pose transformation of the object to determine which aspect will be viewed. When gripping is used, the system would need to invoke a modeling system to predict the width of the object based on the manipulator's position and the hypothesized object transformation. For these reasons, we have not yet integrated the f/t sensor with the system.

VIII. Choosing the Best Sensing Strategy

In this section, we will describe the algorithms that our system uses to choose a sensing strategy. In essence, this is a search problem. The search space consists of the possible sensing operations from the possible viewpoints. Goal states are recognized using $A^i_{\max}$. Since the space of locations can be arbitrarily large (consider that the manipulator can be used anywhere in the robot's work envelope), we must devise some heuristic to guide the search through the space of possible sensor applications. In order to accomplish an efficient search of this space, we use the concept of the aspect graph.

As described in Section III, an aspect is a set of features which can be observed simultaneously from a single viewpoint. Therefore, for the purposes of observing features, for a particular object, there are only as many unique viewpoints as there are aspects. This allows us to define the space of possible sensing operations as the set of sensing operations applied from the principal viewpoints of the aspects of the possible objects.

Our algorithm for determining the best next sensing operation searches over this space of possible sensing operations as follows. First the viewpoint (expressed in the world coordinate frame) which corresponds to the principal viewpoint of some aspect of an active hypothesis is computed. Then, for each active hypothesis, we predict the set of features which would be observed from this viewpoint. For each of these, we determine the hypothesis set which would result if those features were actually found by the sensing system, and calculate the corresponding ambiguity. The maximum of these ambiguities is noted and associated with the proposed sensing operation. The viewpoint/sensor pair that minimizes the maximum ambiguity is chosen as the next sensing operation.

There are three basic components to the algorithm. First,

```
predict-ambiguity(VP,Ω_{k-1},S,sensor)
    Ω ← refine-hyp-set(Ω_{k-1},S)
    A ← 0
    foreach θ ∈ Ω
        A ← A - pr(θ) log pr(θ)
    return(A)
```

Fig. 9. The algorithm for predict-ambiguity.

```
max-ambiguity(Ω_{k-1},VP,sensor)
    max ← 0
    foreach θ ∈ Ω_{k-1}
        S ← predicted sensed values for θ, VP and sensor
        A ← predict-ambiguity(VP,Ω_{k-1},S,sensor)
        if (A > max) then
            max ← A
    return(A)
```

Fig. 10. The algorithm for max-ambiguity.

```
choose-next-view(Ω_{k-1})
    Amax ← 100
    foreach h ∈ Ω_{k-1}
        T ← compute-transform(h)
        Node-list ← get-aspect-graph-nodes(h)
        foreach S ∈ sensors
            foreach node ∈ Node-list
                VP ← node.princple-view
                W-VP ← T VP
                NAmax ← max-ambiguity(Ω_{k-1},W-VP,S)
                if (NAmax < Amax) then
                    Amax ← NAmax
                    Sensor ← S
                    V ← W-VP
    return(Amax,V,Sensor)
```

Fig. 11. The algorithm for choose-next-view.

the *predict-ambiguity* function computes the predicted ambiguity for a specified view point, hypothesis set, sensor, and set of predicted feature values. The first step in predict-ambiguity is to use the procedure described in Section IV to refine the hypothesis set using the predicted feature values. Once this is done, the ambiguity is calculated and returned. This algorithm is shown in Fig. 9.

The function *max-ambiguity* uses predict-ambiguity to find the maximum possible ambiguity for a candidate sensing operation. This is done by successively calling predict-ambiguity with S set to the set of features visible for each of the active hypotheses. The predicted feature values are computed based on the hypothesis set, sensor, and proposed viewpoint, as discussed in Section VI. The maximum of these values is returned as the maximum ambiguity. This algorithm is shown in Fig. 10.

Finally, the top level function used to determine the next sensing operation is *choose-next-view*, shown in Fig. 11. This function iterates over the nodes in the aspect graphs for each object hypothesis for each possible sensor.

IX. Experimental Results

In order to demonstrate the utility of the methods that we have described, in this section we will present the results of two experiments. For the first experiment, we describe how the system develops its initial hypothesis set, and distributes belief among the hypotheses. The purpose of this is to further illustrate the hypothesis generation system described in Section IV. After an initial set of hypotheses has been derived, the algorithms described in Section VIII are applied to choose a next sensing operation.

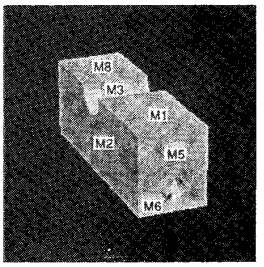


Fig. 12. CAD rendering of the experimental object, with surfaces labeled.

TABLE I
NODE INFORMATION FROM THE ASPECT GRAPH FOR THE OBJECT SHOWN IN FIG. 12

Node	Tessels	Visible Faces	Principal
Node- 1	1,46,50	8,81,9,3	46
Node- 2	2	7,9,3	2
Node- 3	3,14	9,10	3
Node- 4	4	10,9,5	4
Node- 5	5	9,5,3,6	5
Node- 6	6,7,10	5,6,9,10	6
Node- 7	8,20	5,10	8
Node- 8	9	10,5,6	9
Node- 9	11,12,15	9,10,7	11
Node-10	13	10,7	13
Node-11	16,18,31	2,10	18
Node-12	17,36,37,38,40	10,2,7	36
Node-13	19,26,27,28,30	10,2,6,5	27
Node-14	21,24,25,47	5,6,3,8,81,9	25
Node-15	22	9,5,6	22
Node-16	23	5,6	23
Node-17	29	5,6,2	29
Node-18	32	2,3,7	32
Node-19	33,58	2,3,8,81	33
Node-20	34	2,3,5,8,81	34
Node-21	35	2,3,5,6,	35
Node-22	39	2,7	39
Node-23	41,45	9,7	41
Node-24	42,43,44	8,81,3,9,7	43
Node-25	48,49	8,81,3	48
Node-26	51,52,53,54,55	8,81,3,2,5,6	55
Node-27	56,57,59,60	8,81,3,2,7	60

A CAD model of the object which was used is shown in Fig. 12. The bottom face is M_{10}, the face opposite M_2 is M_9, and the face opposite M_5 is M_7. The face M_7 has no hole, and is therefore distinct from M_5. Note that faces M_2 and M_9 are identical, and thus correspond to a single unique model feature, as do faces M_8 and M_1. Further, note that, unless one end of the object is visible (i.e., either M_5 or M_7), the pose transformation of this object cannot be uniquely determined. An aspect graph for this object was created by using the PADL2 system [12] to automatically view the CAD model of the object from each of 60 viewpoints (which correspond to the centers of the 60 tessels on a tessellated sphere), and then grouping together viewpoints which observed the same set of features. Table I shows the relevant information for each node

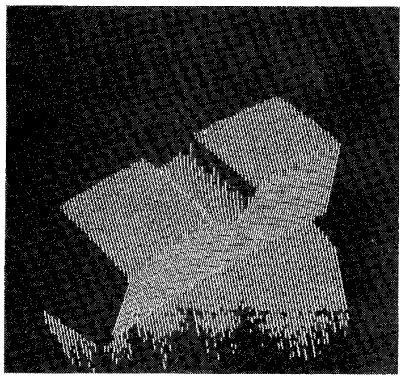


Fig. 13. Composite light-stripe image from the first experiment.

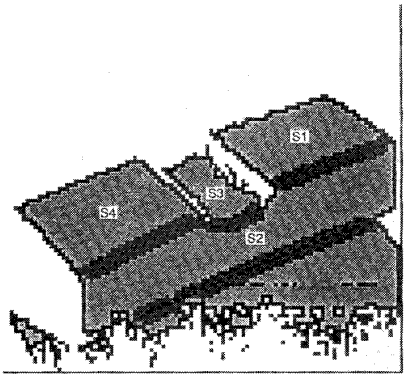


Fig. 14. Segmented image for object as shown in Fig. 13.

in the aspect graph for the object shown in Fig. 12. Feature weights were then assigned to each feature of each aspect based on the visible area of the feature in the aspect. Range data for the two experiments were acquired using a single-stripe structured light scanner.

In the first experiment, the object was placed so that the range scanner could not observe either end of the object, and therefore could not make a unique hypothesis about the pose. The corresponding composite light-stripe image is shown in Fig. 13, and the results of segmentation are shown in Fig. 14. Four surfaces were found (excluding the surface of the work table), and the corresponding $m_i(\cdot)$'s are shown in Table II. When these individual feature matches were combined, the resulting common refinement contained 686 possible hypotheses. After applying object consistency, the number was reduced to 420. This was subsequently reduced to 4 hypotheses using the location and dot product consistencies (note that in the experiments, we deleted hypotheses whose belief dropped below 1 percent of the maximum belief assigned to any hypothesis). The resulting $m_r(\cdot)$ is shown in Table III. Finally, aspect consistency was applied. In this particular experiment, aspect consistency did not provide any real improvement. The resulting $m_a(\cdot)$ is shown in Table. IV. The final bpa is shown in Table V. Note that the two hypotheses which account for better than 96 percent of the system's committed belief correspond to the two correct hypotheses which are indistinguishable from this viewpoint.

TABLE II
bpa's $m_i(\cdot)$ FOR THE OBJECT AS SHOWN IN FIGS. 13 AND 14

θ	$m_i(\theta)$
{S1/M2,S1/M9}	0.10285
{S1/M8,S1/M1}	0.232255
{S1/M5}	0.277249
{S1/M7}	0.277249
{S1/M10}	0.110399
{S2/M2,S2/M9}	0.409585
{S2/M8,S2/M1}	0.0491467
{S2/M5}	0.0285053
{S2/M7}	0.0285053
{S2/M10}	0.484258
{S3/M3}	0.957109
{S3/M6}	0.0428911
{S4/M2,S4/M9}	0.0991155
{S4/M8,S4/M1}	0.236474
{S4/M5}	0.278756
{S4/M7}	0.278756
{S4/M10}	0.1069

TABLE III
bpa $m_r(\cdot)$ FOR THE OBJECT AS SHOWN IN FIGS. 13 AND 14

θ	$m_r(\theta)$
{S1/M1,S2/M9,S3/M3,S4/M8}	0.481687
{S1/M8,S2/M2,S3/M3,S4/M1}	0.481653
{S1/M8,S2/M2,S3/M3,S4/M5}	0.018331
{S1/M1,S2/M9,S3/M3,S4/M7}	0.0183302

TABLE IV
bpa $m_a(\cdot)$ FOR THE OBJECT AS SHOWN IN FIGS. 13 AND 14

θ	$m_a(\theta)$
{S1/M1,S2/M9,S3/M3,S4/M8}	0.274548
{S1/M8,S2/M2,S3/M3,S4/M1}	0.274548
{S1/M8,S2/M2,S3/M3,S4/M5}	0.225452
{S1/M1,S2/M9,S3/M3,S4/M7}	0.225452

TABLE V
FINAL bpa FOR THE OBJECT AS SHOWN IN FIGS. 13 AND 14

θ	$m(\theta)$
{S1/M1,S2/M9,S3/M3,S4/M8}	0.482252
{S1/M8,S2/M2,S3/M3,S4/M1}	0.482219
{S1/M8,S2/M2,S3/M3,S4/M5}	0.0177653
{S1/M1,S2/M9,S3/M3,S4/M7}	0.0177646

For practical reasons, since more than 96 percent of the belief was assigned to the first two hypotheses, we only used these two hypotheses to generate a list of candidate sensing operations. However, we retained all four hypotheses in the hypothesis set which was used to predict ambiguities. Table VI shows the maximum ambiguities for the 2-D camera for the viewpoints generated using the aspect graph for the first hypothesis. Note that although the aspect graphs for the two

TABLE VI
MAXIMUM AMBIGUITIES FOR VIEWPOINTS CHOSEN USING THE FIRST HYPOTHESIS, IN THE FIRST EXPERIMENT

Viewing Location	A_{max}
14.2305, 34.3345, −4.09151	0.846614
16.1961, 26.4864, −15.2748	1.17139
14.9075, 28.3377, −23.3033	0.760302
9.45419, 34.637, −23.6519	1.04442
7.37281, 36.6786, −15.8389	0.95498
10.2462, 40.0579, −28.9101	0.728456
19.265, 47.4922, −35.6511	0.930758
14.6084, 52.6477, −31.0379	0.926031
28.277, 25.8246, −25.6415	0.872521
29.6839, 35.4568, −34.9851	2.02616
37.7903, 53.2806, −30.2827	1.23105
42.7963, 40.6989, −27.6067	1.36688
25.9384, 60.1724, −28.6844	0.842919
9.22447, 46.9162, −6.76749	0.716441
7.13906, 49.9121, −19.7573	0.638637
10.6317, 56.5486, −16.1111	0.846614
22.5702, 63.4762, −16.0431	0.846614
44.6479, 50.9364, −18.5353	1.41959
42.5666, 52.978, −10.7223	0.746736
37.1133, 59.2774, −11.0709	0.846614
35.8247, 61.1287, −19.0993	0.846614
44.8817, 37.7029, −14.6169	1.54532
29.4506, 24.1389, −18.3311	1.4336
26.0824, 27.4427, −5.68983	1.09758
19.8818, 45.0731, −1.80867	0.846614
31.3556, 59.5927, −6.13011	0.692958
42.9865, 35.5622, −7.50256	1.75241

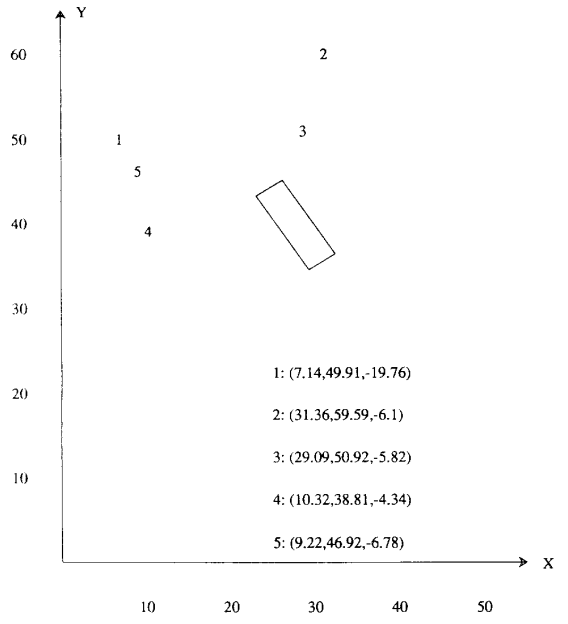


Fig. 15. Five best next viewpoints, as found for the first experiment.

hypotheses have the same structure (because they are the same object), since the two object hypotheses have different position transformations, they will generate a unique set of world viewpoints. Fig. 15 shows the five viewpoints with the smallest values for A_{max}. Note that in this figure, the viewpoints have been projected onto the X–Y plane.

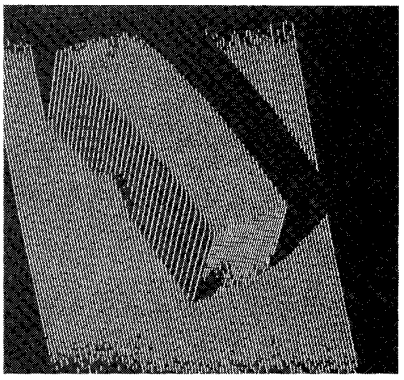


Fig. 16. Composite light-stripe image from the second experiment.

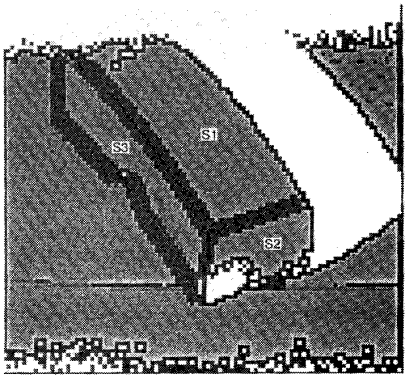


Fig. 17. Segmented image for object as shown in Fig. 16.

In the second experiment, the object was positioned so that one end was visible. The corresponding composite light-stripe image is shown in Fig. 16, and the segmented image in Fig. 17. In this experiment, three surfaces were found (excluding the surface of the work table). The common refinement of the $m_i(\cdot)$'s contained 343 hypotheses, which were reduced to 42 after the application of object and relational consistency. In this experiment, aspect consistency was more important, since one of the ends of the object was visible. In the interest of brevity, these bpa's are not shown in tabular form. They can be found in [15]. In this experiment, the hypothesis awarded the greatest belief, was the correct hypothesis. The second highest belief was given to the hypothesis which had the object turned 180° (i.e., with the other end visible). The remaining two hypotheses (after discarding all hypotheses with total belief less than 1 percent of the maximum belief for any hypothesis) each had the object on its side. This is reasonably credible, because the areas of the bottom and side surfaces are very close (within 4 in^2), and the relational constraints for these two poses are very similar to the relational constraints for the first two hypotheses (due to a high degree of object symmetry).

Again, there was a clear division between the most credible hypotheses and those which were awarded an insignificant portion of the system's belief. Again, to select the next sensing operation, only the four most credible hypotheses were used. The five best viewpoints are illustrated in Fig. 18, Again, the actual viewpoints are projected onto the X–Y plane.

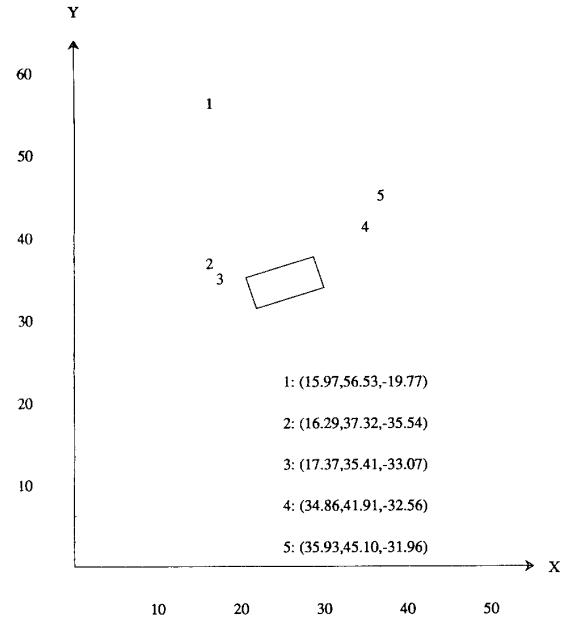


Fig. 18. Five best next viewpoints, as found for the second experiment.

X. Conclusions

In this paper, we have addressed the issue of planning sensing strategies dynamically, based on an active set of hypotheses. Our algorithm uses the aspect graphs of the hypothesized objects to propose candidate sensing operations. Then, using the pose transformation and object model which are associated with each hypothesis we predict the feature sets which would be observed upon application of the candidate sensing operation. Given these predictions, we are also able to predict the resulting set of hypotheses which would remain active. By repeating this process for different viewpoints and sensing operations, we are able to choose the sensing operation which minimizes the maximum ambiguity in these sets, thereby minimizing the amount of ambiguity which can remain after the next sensing operation is applied.

Note Added in Proof (see Section IV-B4)

Since the individual sensory measurement bpa's m_i enter the calculation of both the composite feature match bpa m_f and the aspect consistency bpa m_a, the reader might wonder if the latter two are indeed independent. If by independence is meant lack of predictability, we believe m_f and m_a as defined are independent owing to the multiplication by weight factors w_A's in the calculation of m_a. If we use the metaphor that m_i's are supplied to us by a "geometry expert," whose sole capability lies in being able to tell us how similar a scene feature is to each of the unique model features, then the determination of the weight factors is outside the purview of this expert. In other words, the criteria for judging the "detectability" of a model feature from a given viewpoint are independent of the criteria that tell us how similar or dissimilar a model feature is vis a vis a scene feature.

Acknowledgment

The authors wish to thank B. Cromwell, who provided all of the range data, segmented range images, and symbolic descriptions of range scenes which were used to derive the experimental results; M. Carroll, who provided assistance with the hardware aspects of this project; and J. Lewis, who provided software support, including a number of modifications to the PADL2 system, which enabled the construction of aspect graphs.

References

[1] G. Castore and C. Crawford, "From solid model to robot vision," in *Proc. IEEE Int. Conf. on Robotics and Automation,* pp. 90–92, 1984.

[2] C. H. Chen and A. C. Kak, "3D-POLY: A robot vision system for recognizing objects in occluded environments," Tech. Rep. TR-EE-88-48, School of Electrical Engineering, Purdue University, Lafayette, IN.

[3] C. I. Connolly, "The determination of next best views," in *Proc. IEEE Int. Conf. on Robotics and Automation,* pp. 432–435, 1985.

[4] G. K. Cowan and P. D. Kovesi "Automatic sensor placement from vision task requirements," *IEEE Trans. Pattern Anal. Machine Intell.,* vol. 10, no. 3, pp. 407–416, May 1988.

[5] R. L. Cromwell, "Low and intermediate level processing of range maps," Purdue University Tech. Rep. TR-EE-87-41, 1987.

[6] O. D. Faugeras and M. Hebert, "A 3-D recognition and positioning algorithm using geometrical matching between primitive surfaces," in *Proc. 8th IJCAI,* pp. 996–1002, 1983.

[7] J. D. Foley and A. Van Dam, *Fundamentals of Interactive Computer Graphics.* Reading, MA: Addison Wesley, 1984.

[8] Z. Gigus and H. Malik, "Computing the aspect graph for line drawings of polyhedral objects" in *Proc. Comput. Soc. Conf. on Computer Vision and Pattern Recognition,* IEEE Comput. Soc. Press, 1988, pp. 654–651.

[9] W. E. L. Grimson and T. Lozano-Perez, Model-based recognition and localization from sparse range or tactile data," *Int. J. Robotics Res.,* vol. 3, no. 3, pp. 3–35, Fall 1984.

[10] G. Hager and M. Mintz, "Searching for information," in *Proc. AAAI Workshop on Spatial Reasoning and Multi-sensor Fusion.* Los Altos, CA: Morgan Kaufmann Pub., 1987, pp. 313–322.

[11] C. Hansen and T. Henderson, "CAGD-based computer vision," in *Proc. Workshop on Computer Vision,* pp. 100–105, Dec. 1987.

[12] E. E. Hartquist and H. A. Marisa, *PADL-2 User's Manual,* Production Automation Project, University of Rochester, Rochester, NY, 1985.

[13] M. Higashi and G. J. Klir, "Measures of uncertainty and information based on possibility distributions," *Int. General Syst.* vol. 9, pp. 43–58, 1982.

[14] S. A. Hutchinson, R. L. Cromwell, and A. C. Kak, "Applying uncertainty reasoning to model based object recognition" in *Proc. IEEE Int. Conf. on Computer Vision and Patter Recognition,* 1989.

[15] S. A. Hutchinson, "Sensor and task planning in robotic assembly," Ph.D. dissertation, Purdue University, School of Electrical Engineering, W. Lafayette, IN, Dec. 1988.

[16] S. A. Hutchinson, R. L. Cromwell, and A. C. Kak, "Planning sensing strategies in a robot work cell with multi-sensor capabilities," in *Proc. IEEE Int. Conf. on Robotics and Automation,* pp. 1068–1075, 1988.

[17] K. Ikeuchi, "Generating an interpretation tree from a CAD model for 3D-object recognition in bin picking tasks," *Int. J. Comput. Vision,* pp. 145–165, 1987.

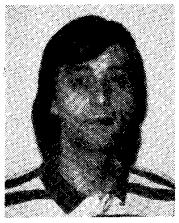
[18] A. C. Kak, "Depth perception for robots," in *Handbook of Industrial Robotics,* S. Nof, Ed. New York, NY: Wiley, 1986, pp. 272–319.

[19] H. S. Kim, R. C. Jain, and R. A. Volz, "Object recognition using multiple views," in *Proc. IEEE Int. Conf. on Robotics and Automation,* pp. 28–33, 1985.

[20] J. J. Koenderink and A. J. Van Doorn, "The internal representation of solid shape with respect to vision," *Biol. Cybern.,* vol. 32, pp. 211–216, 1979.

[21] M. Magee and M. Nathan, "Spatial reasoning, sensor repositioning, and disambiguation in 3D model based recognition," in *Proc. AAAI Workshop on Spatial Reasoning and Multi-sensor Fusion.* Los Altos, CA: Morgan Kaufmann Pub., 1987, pp. 262–271.

[22] H. A. Martins, J. R. Birk, and R. B. Kelly, "Camera models based on data from two calibration planes," *Comput. Vision, Graphics Image Process.*, vol. 17, pp. 173–180, 1981.
[23] A. Rosenfeld and A. C. Kak, *Digital Picture Processing*. New York, NY: Academic Press, 1982.
[24] R. J. Safranek, S. N. Gottschlich, and A. C. Kak, "Evidence accumulation using binary frames of discernment for verification vision," submitted for publication in *IEEE Trans. Robotics Automat.*
[25] G. Shafer, *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton Univ. Press, 1976.
[26] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* Urbana, IL: Univ. of Illinois Press, 1963.
[27] H. E. Stephanou and S. Y. Lu, "Measuring consensus effectiveness by a generalized entropy criterion," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 10, no. 4, pp. 544–554, July 1988.
[28] J. H. Stewman and K. W. Bowyer, "Aspect graphs for convex planar-face objects," in *Proc. IEEE Workshop on Computer Vision*, Dec. 1987.
[29] R. R. Yager, "Entropy and specificity in a mathematical theory of evidence," *Int. J. General Sys.*, vol. 9, pp. 249–260, 1983.
[30] H. S. Yang and A. C. Kak, "Determination of the identity, position, and orientation of the topmost object in a pile," *Comput. Vision, Graphics, and Image Process.*, vol. 36, pp. 229–255, 1986.



Seth A. Hutchinson (S'85–M'86–S'87–M'88) received the B.S.E.E., M.S.E.E., and Ph.D degrees from Purdue University, Lafayette, IN, in 1983, 1984, and 1988, respectively.

He is currently a Visiting Assistant Professor of Electrical Engineering at Purdue. His current research interests include dynamic planning of sensing strategies, constraint based reasoning, task planning in automated assembly, evidential reasoning applied to model-based object recognition, and sensor integration.

Avinash C. Kak (M'71) is a professor in the Department of Electrical Engineering, Purdue University, Lafayette, IN. His current research interests are in reasoning architectures for solving spatial problems, sensor-based robotics, and computer vision. He has co-authored the books *Digital Picture Processing* (New York: Academic Press) and *Principles of Computerized Tomographic Imaging* (New York: IEEE Press).