Applying Uncertainty Reasoning to Model Based Object Recognition

S. A. Hutchinson, R. L. Cromwell, and A. C. Kak

ABSTRACT

In model based object recognition, the primary objective is to efficiently match features which have been extracted from sensory data to corresponding features in object models; this being done with the constraint that relations between the features in the object models are mirrored by the relations between the features extracted from the sensory data. A problem which confronts this process is the difficulty in extracting features and relations from sensed data, and precisely determining the values of their relevant attributes. Furthermore, it is often the case that the features which are visible from a single viewpoint are not sufficient to uniquely identify the object and its pose. In each of these cases, a system is needed which can formulate and associate credibilities with hypotheses about the possible identities and poses of the objects in the scene.

This paper describes an architecture for reasoning with uncertainty about the identities of objects in a scene. The main components of this architecture create and assign credibility to object hypotheses based on feature match, object, relational , and aspect consistencies. We use the Dempster-Shafer formalism for representing uncertainty, so these credibilities are expressed as belief functions which are combined using Dempster's combination rule to yield the system's aggregate belief in each object hypothesis. One of the principal objections to the use of Dempster's rule is that its worst case time complexity is exponential in the size of the hypothesis set. We will show how the structure of the hypothesis sets developed by our system allow for a polynomial time implementation of the combination rule.

1. Introduction

One of the key problems in computer vision today is that of recognizing objects from sensory data, given a set of models for the objects which might be in the scene. This model based object recognition consists of matching features extracted from sensory data to model features, subject to the constraint that the relations between the features extracted from the sensed data mirror the relations between the corresponding model features. One problem that arises during this matching process is that neither the sensed features, nor the measured relationships between those features, will be perfectly accurate. A second problem is that often the features which are visible from a single viewpoint are not sufficient to uniquely identify the object. In each of these cases, we would like for the reasoning system to be able to formulate a set of hypotheses about the possible identities of the objects in a scene, and to then associate credibilities with each of those hypotheses.

In this paper, we present a reasoning architecture which formulates, and assigns belief values to, hypotheses about the identities of objects in a scene. These belief values reflect the similarity of the matched sensed and model features, as well as the similarity of the relations between sensed features and the corresponding model relations. There are two main components in our current system: a 3D vision system (which extracts features, attributes and relations from sensed range data), and a reasoning system (which formulates, and distributes belief among, hypotheses about the identities and poses of objects). Fig. 1 shows a diagram of the vision system, and Fig. 2 shows the reasoning system.

Since the focus of this paper is the application of uncertainty reasoning to the problem of model based object recognition, we will not discuss the specifics of the vision system. We will, however, describe the part of the vision system which classifies surfaces according to their 3D shape (e.g. planar, cylindrical), and assigns confidence values to those classifications.

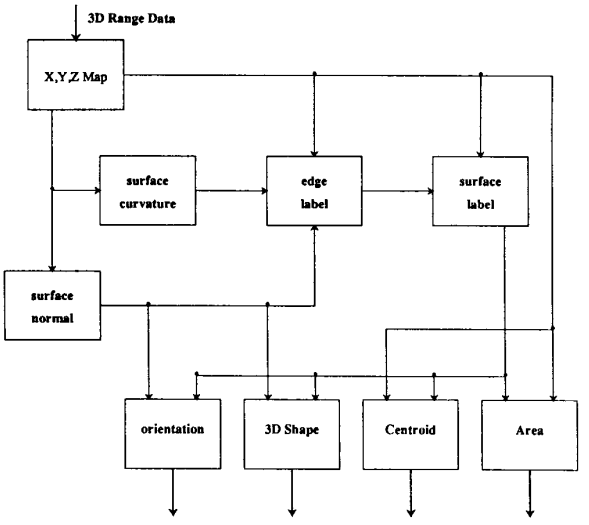

Fig. 1: Block diagram of 3D vision system.

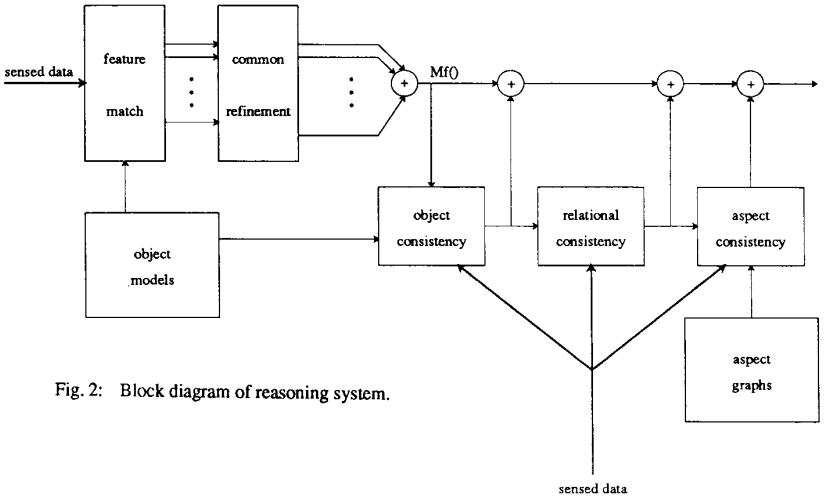

Fig. 2: Block diagram of reasoning system.

There are four main reasoning modules. The first of these creates, and assigns belief to, an initial set of hypotheses based on the goodness of matches between sensed and model features. The second module eliminates hypotheses in which sensed features are matched to model features that are not from the same object. The third module evaluates each of the initial hypotheses on the basis of relational consistency. The resulting belief values are combined with the initial belief values to create a refined hypothesis set, with corresponding belief values that reflect both the quality of the feature matches and the relational consistency of the hypotheses. This refined hypothesis set is then used as input to the final reasoning module, which assigns belief using the criterion of aspect consistency.

The system that we have developed uses the Dempster-Shafer (DS) formalism for representing uncertainty [7]. The main reason for this choice is the concept of refinement, which allows a number of distinct frames of discernment (or sets of hypotheses) to be combined into a single frame (or hypothesis set). This concept forms the basis for the method that we use to combine individual matches of sensed and model features into object hypotheses. One of the primary objections to the DS theory is that the worst case time complexity for the implementation of Dempster's combination rule is exponential in the size of the hypothesis set. Fortunately, the characteristics of the hypotheses generated by our system allow for a polynomial time implementation of the combination rule.

The remainder of the paper is organized as follows. In Section 2, we review some of the fundamental concepts of the Dempster-Shafer theory, and show how the theory is applied to our domain. In Section 3, we describe how features are extracted from range data, how a maximum likelihood classifier is used to assign types and confidences to the 3D shape attribute of surfaces, and how belief functions are derived for individual feature matches. Section 4 describes how individual feature matches are combined to produce object hypotheses, and how the belief values for the individual feature matches are combined to give belief in these hypotheses. Section 5 describes the bpa based on object consistency. In Section 6, we discuss relational consistency, and show how the system assigns belief values based on spatial relations between features. Section 7 introduces aspect consistency, and discusses how the corresponding belief values are assigned. In Section 8, we show that the structure of the hypothesis sets which are developed by our system allow for a polynomial time implementation of Dempster's combination rule. In Section 9, we present some experimental results for ambiguous range data. Finally, in sections 10 and 11, we address possible future research in this area, and conclude the paper.

2. 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 a brief 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.

2.1. 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 $1 \le i \le 6$; therefore, $\Theta = \{1,2,3,4,5,6\}$. The $2^{|\Theta|}$ subsets of Θ are called propositions and the set of all the 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

$$0.0 \le m() \le 1.0, \quad m(\varnothing) = 0 \quad \text{and} \quad \sum_{X \subset \Theta} m(X) = 1.0$$

Subsets of Θ which are assigned non-zero probability mass are said to be *focal elements* of m(). The *core* of m() is the union of its focal elements.

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

$$m(X) = K \sum_{X_1 \cap X_2 = X} m_1(X_1) m_2(X_2)$$

where

$$K^{-1} = 1 - \sum_{X_1 \cap X_2 = \varnothing} m_1(X_1) m_2(X_2)$$

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

$$m = m_1 \oplus m_2$$

to represent the combination of $m_1()$ and $m_2()$.

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 $m_1()$ and $m_2()$ 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, ..., to a common frame, Ω, is accomplished by specifying the mapping functions:

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

which must possess the following properties:

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

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

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

The first property says that any proposition that is discerned in Θ_i must also be discernible in Ω. The second property requires that the mapped propositions in Ω be disjoint. Finally, the third property specifies that if Ω is a refinement of Θ_i, then no proposition in Ω be outside the range of mappings corresponding to the different propositions in Θ_i.

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

$$m_i'(\omega_i(A)) = m_i(A)$$

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

2.2. Applying DS to Object Recognition

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:

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

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, if two edges have been found, S1 and S2, the following represent object hypotheses:

$$\theta_1 = \{S1/1, S2/2\}, \quad \theta_2 = \{S1/a, S2/b\}$$

The hypothesis θ_1 indicates that sensed edge S1 is matched to model edge 1 and that sensed edge S2 is matched to model edge 2. The hypothesis θ_2 indicates that sensed edge S1 is matched to model edge a and that sensed edge S2 is matched to model edge b.

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 the hypothesis contains a sufficient number of feature matches to estimate the pose (this will be discussed in Section 6).

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

$$\Theta = \{\theta_1, \theta_2, \cdots\}$$

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

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

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

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

To illustrate, consider the cube shown in Fig. 3a, which has two faces with large holes, and a third face with a small hole (we will assume that the remaining faces have no holes). Suppose that two sensed features, S_1 and S_2 are found, as shown in Fig. 3b. In this case, for sensed feature S_1 we might assign the bpa $m_1()$ as:

$$m_1(\{S_1/a\}) = 0.7$$

$$m_1(\{S_1/b, S_1/c\}) = 0.3$$

and for S_2, the bpa $m_2()$ as:

$$m_2(\{S_2/a\}) = 0.2$$

$$m_2(\{S_2/b, S_2/c\}) = 0.8$$

based on the similarity between the sensed and model faces.

In order to combine the evidences generated by these two measurements, we first construct a refinement of the two frames discerned by $m_1()$ and $m_2()$. The following is a valid refinement which obeys the above three properties:

$$\Omega = \{ \{S_1/a, S_2/a\}, \{S_1/a, S_2/b\}, \{S_1/a, S_2/c\},$$
$$\{S_1/b, S_2/a\}, \{S_1/b, S_2/b\}, \{S_1/b, S_2/c\},$$
$$\{S_1/c, S_2/a\}, \{S_1/c, S_2/b\}, \{S_1/c, S_2/c\} \}$$

Note that some of the elements in this frame are nonsensical, for example, those which assign distinct sensed features to the same model feature. In Section 5 we will describe how object consistency is used to eliminate such propositions from the frame (by assigning zero belief to the proposition).

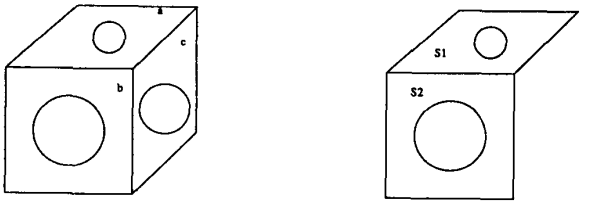

Fig. 3: Object model of cube with holes in three sides (a), and two features extracted from sensed data (b).

3. Evaluating Feature Matches

In this section of the paper, we will discuss how bpa's are associated with individual matches between sensed and model features. First we give an overview of the range data processing that is used to extract features from the sensed data. We then discuss the maximum likelihood classifier which the system uses to make 3D shape hypotheses for surfaces, and to associate confidence values with those hypotheses. Finally, we discuss how the bpa's are derived which represent the system's belief in matches between sensed and model features.

3.1. Overview of Range Interpretation Steps

The first step in assigning belief to feature matches is the extraction of features from sensed data. This proceeds as follows. First, the scene is scanned by a range sensor to obtain a range map. Window convolutions are then performed on this range map to produce a surface normal map and a curvature map, as described in [9]. Entries in the surface normal map are vectors describing the local surface normal at the corresponding points in the range map. The curvature map contains the values of the mean and Gaussian curvature for each point in the range map.

The range, surface normal, and curvature values are then examined in the neighborhood of each point to identify edge points, as described in [2]. Once edge points have been identified, the remaining points are grouped into contiguous regions, which correspond to the surfaces visible in the scene. A number of attributes are then found for each surface, including:

3D shape – The 3D shape of the surface: *planar*, *cylindrical*, *spherical*, etc.

Area – The area of the surface.

Location – The centroid of the surface.

Orientation – The normalized average of the surface normals of the surface's constituent points.

Several other attributes and relations are also found. They are described in detail, along with the algorithms used, in [2].

3.2. 3D Shape Classification by Surface Normal Distribution Analysis

Analysis of the distribution of surface normals provides a robust approach for the classification of the 3D shape of a surface. At the present time, we use this method to initially classify surfaces as *planar*, *cylindrical*, and *spherical*. A further examination of the characteristics of the surface allow the *cylindrical* surfaces to be further classified as *circularly-cylindrical*, *elliptically-cylindrical*, and *irregular*; while the *spherical* surfaces can be further classified as *spherical*, *ellipsoidal*, and *irregular*.

3.2.1. Overview of the 3D Shape Classification Method

The first step in analyzing the surface normal distribution for a particular surface is the construction of a Gaussian sphere for that surface. This is a unit sphere, with each surface normal of the surface mapped to the corresponding point on that sphere. It is well known that Gaussian spheres do not provide unique representations of nonconvex objects, however, we are only trying to classify the 3D shape of a single surface. While it is true that there is still ambiguity involved, (e.g. concave and convex cylindrical surfaces will produce the same distribution of surface normals, as will a surface with irregularity mainly in one dimension, such as a corrugated surface) an additional step is used to disambiguate these cases (this will be described shortly).

Since the surface normals are unit vectors, there is no work involved in "constructing" the Gaussian sphere; the surface normal vectors for a given surface are already in an appropriate representation. For this reason, the distribution of surface normals on the Gaussian sphere is continuous; there is no quantization (or tessellation) of the sphere.

The distribution of the surface normals on the Gaussian sphere will be distinct for each of the major 3D shapes. A *planar* surface will produce a very small distribution, a single point in the ideal case. The *cylindrical* surfaces will produce elongated distributions, which, in the ideal case, are portions of great circles. (A "great circle" is the intersection of a sphere with a plane passing through its center.) Finally, the *spherical* surfaces will produce distributions of relatively large area, but round in shape. The problem is to classify the shape of the distribution as one of these. The method that we use is to produce an equidistant-azimuthal mapping of the sphere, with the origin at the centroid of the distribution. This mapped distribution is then analyzed.

The average of the normals for a surface is a vector which corresponds to the centroid of the distribution of the normals on the surface of the Gaussian sphere. An equidistant-azimuthal (E-A) mapping, also called a great-circle mapping, is a 2D mapping of the distribution, with equally low distortion for any arbitrary origin on the sphere. The E-A mapping for each surface has its origin at the centroid of the distribution. A more detailed description of this method may be found in [2]. Let it suffice here to say that this step produces a 2D distribution of points which have the same correspondences between 2D distribution shape and 3D surface shape as discussed earlier. In this new distribution, each point corresponds to a surface normal, and is assigned the coordinates (θ_1, θ_2), where θ_1 may be thought of as *degrees-north*, and θ_2 may be thought of as *degrees-east*.

Given these 2D mappings, the following rule will produce the desired first level of 3D shape classification, when applied to the *area* and *eccentricity* of the distribution:

if the *area* is low, **then**
the surface is *planar*,
else if the *area* is high, **and** the *eccentricity* is low, **then**
the surface is *spherical*,
else if the *area* is high, **and** the *eccentricity* is high, **then**
the surface is *cylindrical*,

All that remains to be done for the initial classification is a measurement of the area and eccentricity of the distribution.

Once the initial classification has been made, a final step provides finer classification. The hypothesized cylindrical axis is used in an attempt to measure the radius of the surface. The measurements made here allow the detection of concave surfaces, as well as the further classification of *cylindrical* surfaces as *circularly-cylindrical*, *elliptically-cylindrical*, or *irregular*; and generally *spherical* surfaces as *spherical*, *ellipsoidal*, or *irregular*.

3.2.2. Associating Confidence Values With the 3D Shape Classification

The classification strategy described above can be implemented by making measurements of the mapped distribution of surface normals, and applying empirically determined thresholds. However, this approach fails to provide any information about the system's certainty in the classification. Furthermore, if the surface normals are sufficiently noisy, an incorrect classification may result. For this reason, rather than making a single classification of a surface's 3D shape, we use a statistical pattern classifier to assign confidences to the various classifications which are possible, given the distribution of surface normals. In order to do this, we measure a number of characteristic features of the distributions (which estimate the size and shape of the distribution of points), and then use a Gaussian density function to approximate the distribution of these features for each class.

3.2.2.1. Features that Describe a 2D Distribution of Points

The desired features describe the overall area and the eccentricity of a 2D distribution of points. To find these features, we first find the eigenvalues of the covariance matrix of the 2D distribution. For each of the surface normal vectors belonging to the distribution of surface i, $[\theta_{1i}, \theta_{2i}]^T$ describes the transformation of the vector into the 2D space. The mean vector is then estimated as:

$$\bar{\Theta} = \begin{bmatrix} \bar{\theta}_1 \\ \bar{\theta}_2 \end{bmatrix}$$

where

$$\bar{\theta}_i = \frac{1}{n} \sum_{j=1}^{n} \theta_{ij}$$

The covariance matrix is:

$$M = \begin{bmatrix} \hat{m}_{11} & \hat{m}_{12} \\ \hat{m}_{21} & \hat{m}_{22} \end{bmatrix}$$

where

$$\hat{m}_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (\theta_{ik} - \bar{\theta}_i)(\theta_{jk} - \bar{\theta}_j)$$

We have considered a number of features based on the eigenvalues of M, λ_1 and λ_2 (where $\lambda_1 > \lambda_2$), and the features that we have chosen to characterize the size of the distribution are the following:

$$f_1 = \lambda_1$$
$$f_2 = \sqrt{\lambda_1^2 + \lambda_2^2}$$

The feature f_1 describes the length of the major axis of the distribution. For elongated distributions, $f_2 \approx f_1$, becoming significantly larger for more compact distributions where λ_2 approaches λ_1.

For the eccentricities, we use the feature:

$$f_3 = \frac{\lambda_1 + \lambda_2}{\lambda_1 - \lambda_2}$$

The feature f_3 will be larger for more compact distribution, and show greatest variation when λ_2 approaches λ_1.

3.2.2.2. Building a Classifier To Assign Confidence Values

Given the features described above, using N training samples for each class (N = 50 in our current experiments), we constructed a classifier as follows. First, surfaces were selected from a number of scenes, and classified as being *planar*, *cylindrical*, or *spherical*. For each surface, the values of the feature vector $[f_1, f_2, f_3]^T$ were determined. Then, the mean vector and covariance matrix were estimated for each class. The mean vector for class i is:

$$U_i = \begin{bmatrix} \hat{\mu}_{i1} \\ \hat{\mu}_{i2} \\ \hat{\mu}_{i3} \end{bmatrix}$$

where the entries $\hat{\mu}_{ij}$ are estimated as:

$$\hat{\mu}_{ij} = \frac{1}{N} \sum_{m=1}^{N} f_{jm}^i$$

Here the value f_{jm}^i is the value of the j'th feature for the m'th training sample of class i, where $1 \le m \le N$. The entries of the covariance matrix Σ_i are estimated as:

$$\Sigma_i = \begin{bmatrix} \hat{\sigma}_{i11} & \hat{\sigma}_{i12} & \hat{\sigma}_{i13} \\ \hat{\sigma}_{i21} & \hat{\sigma}_{i22} & \hat{\sigma}_{i23} \\ \hat{\sigma}_{i31} & \hat{\sigma}_{i32} & \hat{\sigma}_{i33} \end{bmatrix}$$

where

$$\hat{\sigma}_{ijk} = \frac{1}{N-1} \sum_{m=1}^{N} (f_{jm}^i - \hat{\mu}_{ij})(f_{km}^i - \hat{\mu}_{ik})$$

where $\hat{\sigma}_{ijk}$ is the estimated covariance between feature j and feature k for the i^{th} class.

We make the assumption that samples of the different classes exhibit a Gaussian distribution in the feature space, with the mean vector of U_i and covariance matrix Σ_i. Given these values, and given the feature vector of the unclassified surface, denoted by X, the probability that a surface with feature vector X is of a particular class ω_i is defined by:

$$p(X \mid \omega_i) = (2\pi)^{-\frac{3}{2}} \mid \Sigma_i \mid^{-\frac{1}{2}} e^{d_i^2}$$

where

$$d_i^2 = -\tfrac{1}{2}(X - U_i)^T \Sigma_i^{-1} (X - U_i)$$

3.3. Assigning Belief Values to Individual 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, faces b and c of the block shown in Fig. 3a. Our system groups together model surfaces which appear equivalent, each such grouping corresponding to one unique model feature. Each unique model feature is given a label. 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, $m_i()$, 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

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

we define $m_i()$ as:

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

where,

$$C(F,S_i) = C_S(F,S_i)\,C_A(F,S_i)$$

The value for $C_S(F,S_i)$ is the value for $p(X \mid \omega)$ (determined by the 3D shape classifier discussed above), where ω is the 3D 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 area of the sensed surface and unique model feature, and is calculated by:

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

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 based on empirical data. Note that it is possible that $m_i()$ will assign nonzero belief to non-singleton subsets of M. When this occurs, it reflects the system's ignorance about which match for a particular sensed feature is best.

4. Generating Object Hypotheses from Feature Matches

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 2, 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, $m_i()$, which defines our belief in propositions of the form "sensed feature S_i matches model feature f." In other words, Θ_i, the frame of discernment for a particular $m_i()$ only includes propositions about the i^{th} sensed feature. In order to combine these individual bpa's, we must first combine the feature matches represented by each Θ_i

Given N sensed features, we construct Ω as follows:

$$\Omega = \{\{\theta_1, \theta_2, \cdots \theta_N\} \mid \theta_i \in \Theta_i\}$$

That is, 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 ω_i, the refining from Θ_i to Ω as:

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

for singleton subsets of Θ_i, and

$$\omega(A) = \bigcup_{\theta \in A} \omega(\{\theta\})$$

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:

$$m_f = m_1' \oplus m_2' \oplus \cdots m_n'$$

where $m_i'()$ is computed as in Section 2.

5. Object Consistency

It is quite likely that some of the hypotheses in the frame discerned by $m_f()$ 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 $m_f()$ with $m_o()$.

We want $m_o()$ 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

$$m_o(\theta) = \begin{cases} 1 & : \text{the largest } \theta \subset \Omega \text{ s.t. } \theta \text{ has no inconsistent matches} \\ 0 & : \text{otherwise} \end{cases}$$

6. 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 hypothsis 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, $m_r()$, which is a combination of a number of bpa's, one for each type of relation. For example, one component of $m_r()$ is the bpa $m_n()$ 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.

When such an $m_r()$ 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.

The first relation we use is based on the angles between surfaces. In particular, since the 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:

$$E(\phi,\psi) = |\, n_S(\phi)\cdot n_S(\psi) - n_M(\phi)\cdot n_M(\psi) \,|$$

for ϕ and ψ in θ, and $\theta \in \Omega$. Since E() 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()=0 corresponding to an exact match, and E()=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:

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

Finally, we transform C_n into a valid bpa by normalization:

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

The second bpa that we use to construct $m_r()$ is based on the fact that we can determine the correct location of a feature once a pose transformation for the object has been computed. If we have enough feature matches in a hypothesis, we can derive a pose transformation for that hypothesis, T_{obj}, as described in [5]. We can then use this transformation to measure 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 based on T_{obj}. 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 $T_{obj}L_M(\phi)$ 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:

$$c(\phi) = e^{-\tau\, |\, L_S(\phi) - T_{obj} L_M(\phi) \,|}$$

where τ is empirically determined. We combine the c()'s to obtain a confidence in the proposition θ by taking their product over the feature matches in θ.

$$C_1(\theta) = \prod_{\phi \in \theta} c(\phi)$$

We obtain $m_1()$ by normalizing $C_1()$.

7. Aspect Consistency

The final bpa which we consider in evaluating the quality of an object hypothesis is based on the idea that the system can determine which features should be observed if a pose transformation for the hypothesis has been determined. This bpa, $m_a()$, is derived by accumulating positive evidence when expected features are matched. In order to determine which features should be observed, we use the concept of the aspect graph.

The aspect graph was originally developed by Koenderink and van Doorn [6] (who referred to it as the *visual potential*) to characterize the visual stimulus produced by an object when viewed from different relative positions. They developed a function for the "sensory inflow" produced by an object, in terms of the invariant properties of the object and the relative positions of the viewer and the object. An aspect is characterized by the structure of the singularities in this function for a single view. From most all vantage points, an observer may execute small movements without affecting the aspect. However, when the observer's movement causes the structure of the singularities to be changed, 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.

In our work, we characterize aspects, not in terms of the singularities in the function which defines the visual inflow, but in terms of features which are visible to the range sensor. In particular, we define an aspect to be a set of features which can be observed simultaneously from a particular 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.

Aspect graphs for objects can be generated analytically or by an exhaustive examination of the object. Analytic techniques have been reported by Castore and Crawford [1] and Stewman and Bowyer [8]. We generate our aspect graphs by exhaustive examination. This is done by creating a CAD model of the object, centered within a tessellated viewing sphere (we currently use 60 tessellations) The geometric modeler is then used to view the object from the center point of each tessellation, and the set of visible features is recorded. Using this information, it is a simple matter to generate the aspect graph. Tessellations from which the same feature set is observed are grouped together into nodes. The arcs between nodes are generated using tessellation adjacency.

As we discussed in Section 6, it is possible to derive a position transformation for an object hypothesis if enough features have been matched in that hypothesis. Using this transformation in conjunction with the viewpoint used in the sensing process, we can determine the aspect that would be observed from that viewpoint for a particular hypothesis θ. We will use the function $A(\theta,V)$ to represent the aspect that would be observed for a given object hypothesis θ, from viewpoint V.

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 $m_a'()$, based only on the features that it extracted. These individual $m_a'()$'s are then combined to form $m_a()$.

The features associated with an aspect are given weights which reflect the likelihood that they will be found by a sensing operation. These weights are a function of how conspicuous the features are (and are currently assigned in an ad hoc fashion). We use $w_a(f,a)$ to represent the weight given to model feature f in aspect a. By using $w_a()$ in conjunction with the quality of the feature matches in an object hypothesis, we derive the aspect consistency.

We use the function $q(f,\theta,V)$ to represent the quality of the match for model feature f in the object hypothesis represented by θ, provided that feature f was observed from viewpoint V. Using the $m_i()$'s that were discussed in Section 3.3, we define:

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

Thus, to determine the value of q(), first, determine which sensed feature, S_i, is matched to model feature f in the object hypothesis represented by θ. Then, examine $m_i()$ (the bpa which assigned belief to feature matches for the the i^{th} sensed feature) and determine how much belief is placed in the proposition which includes the match of S_i to f. This is the value for q(), provided S_i was observed from viewpoint V.

Finally, given a function $F_a(x)$, which returns the set of features visible in aspect x, we define the aspect confidence in an object hypothesis to be:

$$C_a(\theta,V) = \sum_{f \in F_a(A(\theta,V))} w_a(f,A(\theta,V))\, q(f,\theta,V)$$

Essentially, this equation says that the aspect consistency is obtained by summing the product of a feature's likelihood of being extracted with the quality of the match for that feature, for each feature that we expect to find in the hypothesized aspect. We construct $m_a()$ by normalizing $C_a()$.

8. 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 polynomial 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 polynomial 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 $\subset \Theta$, if $m(A) > 0$, $m(B) > 0$, and $A \neq B$, then $A \cap B = \emptyset$.

This condition is equivalent to the statement that subsets of Θ with positive basic probability numbers 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. In particular, the system never constructs a bpa such that it assigns positive basic probability numbers to two non-disjoint 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 only if $A \cap B = C$, $m_1(A) > 0$, 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 $m(C) > 0$ and $C \neq \emptyset$, 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 non-empty. 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.

QED.

We now state our central 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 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 non-identical 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 $B_j \cap B_l=\emptyset$. Now, let us 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.

QED.

One 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 polynomial 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 polynomial 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 polynomial time and that we can find m(A) in polynomial 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. Thus, 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 polynomial time.

- In order to show that m(C) can be found in polynomial 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.

QED.

9. Experimental Results

In order to demonstrate the utility of the methods that we have described, in this section we will present the results of one experiment in which the data is somewhat ambiguous. The belief values which are assigned by the system are intuitively pleasing, given the ambiguity of the data.

A CAD model of the object which we used is shown in Fig. 4. 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 [4] 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. 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 was acquired using a single stripe structured light scanner.

In the 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. 5a, and the results of segmentation are shown in Fig. 5b. Four surfaces were found (excluding the table), and the corresponding $m_i()$'s are shown in Table 1. 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 $m_r()$ is shown in Table 2. Finally, aspect consistency was applied. In this particular experiment, aspect consistency did not provide much improvement. The resulting $m_a()$ is shown in Table 3. The final bpa is shown in Table 4. 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.

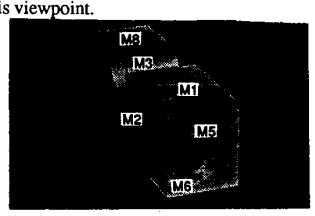

Fig. 4: Rendering of CAD model of experimental object.

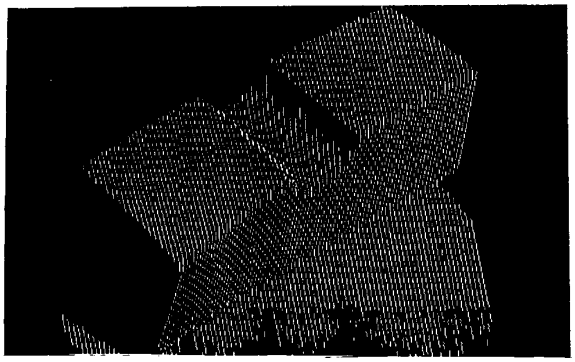

Fig. 5a: Composite light stripe image of object as positioned in first experiment.

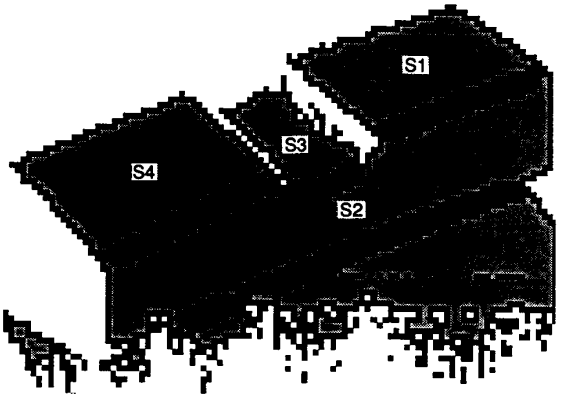

Fig. 5b: Segmented image for composite light stripe image shown in Fig. 5a.

Matches	Belief
{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 1: Bpa's $m_i(\cdot)$ for the object as shown in Fig. 5.

S1	S2	S3	S4	Belief
M1	M9	M3	M8	0.481687
M8	M2	M3	M1	0.481653
M8	M2	M3	M5	0.018331
M1	M9	M3	M7	0.0183302

Table 2: Bpa, $m_r(\cdot)$ for the object as shown in Fig. 5.

S1	S2	S3	S4	Belief
M1	M9	M3	M8	0.274548
M8	M2	M3	M1	0.274548
M8	M2	M3	M5	0.225452
M1	M9	M3	M7	0.225452

Table 3: Bpa $m_a(\cdot)$ for the object as shown in Fig. 5.

S1	S2	S3	S4	Belief
M1	M9	M3	M8	0.482252
M8	M2	M3	M1	0.482219
M8	M2	M3	M5	0.0177653
M1	M9	M3	M7	0.0177646

Table 4: Final Bpa for the object as shown in Fig. 5.

10. Future Work

Although we have dealt with the complexity issues of Dempster's combination rule, there remain complexity issues related to the size of the hypothesis sets. In particular, the formation of a common refinement of the bpa's which represent feature matches results in a hypothesis set whose size is k^N, where we have N sensed features which can, on the average, each be matched to k model features. Although this has not yet been a serious problem for our system (since both k and N have been small in our experiments) we anticipate that it will become a problem as the number of models in the library continues to grow.

It has been shown by a number of researchers (e.g. [3]) that enforcing relational consistency serves to greatly limit the number of hypotheses which a recognition system must consider. In order to take advantage of this, we must delay the combination of all m_i's into a common refinement until after we have enforced relational constraints. By initially giving the system only a small number of sensed features, we limit the size of the initial hypothesis set. We can then feed additional sensed features to the system, and the resulting m_i's are combined with the initial hypotheses in the common refinement module (via the feedback loop). In this way, relational consistency can be used to limit the number of hypotheses which are generated based on feature matches.

11. Summary

In this paper, we have presented an architecture for reasoning with uncertainty about the identities of objects in a scene. We have described the three methods our system uses to assign belief to object hypotheses based on feature matches, relational consistency and aspect consistency. For each of these, we have also described how belief functions are derived from the sensory data. The system that we have implemented uses the Dempster-Shafer formalism for dealing with uncertainty, and we have shown that the structure of the hypothesis sets which our system develops allows for a polynomial time implementation of the combination rule. Finally, we have shown experimental results which affirm the effectiveness of our method in assessing the credibility of candidate object hypotheses.

REFERENCES

[1] G. Castore, C. Crawford, "From Solid Model to Robot Vision," *Proc. of the IEEE Int'l Conf. on Robotics and Automation*, 1984, pp. 90-92.

[2] R. L. Cromwell, "Low and Intermediate Level Processing of Range Maps," Purdue University Technical Report TR-EE-87-41, 1987.

[3] W. E. L. Grimson and T. Lozano-Perez, "Model-Based Recognition and Localization from Sparse Range or Tactile Data." *The International Journal of Robotics Research*, Vol. 3, No. 3, Fall 1984, pp. 3-35.

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

[5] S. A. Hutchinson, R. L. Cromwell and A. C. Kak, "Planning Sensing Strategies in a Robot Work Cell with Multi-Sensor Capabilities," *Proc. of the IEEE Int'l Conf. on Robotics and Automation*, 1988, pp. 1068-1075.

[6] J. J. Koenderink and A. J. Van Doorn, "The Internal Representation of Solid Shape with Respect to Vision," *Biological Cybernetics*, Vol. 32, 1979, pp. 211-216.

[7] G. Shafer, *A Mathematical Theory of Evidence*, Princeton University Press, 1976, Princeton.

[8] J. H. Stewman, K. W. Bowyer, "Aspect Graphs for Convex Planar-Face Objects," *Proc. of the IEEE Workshop on Computer Vision*, Dec. 1987.

[9] H. S. Yang and A. C. Kak, "Determination of the Identity, Position, and Orientation of the Topmost Object In a Pile," *Computer Vision, Graphics, and Image Processing*, Vol. 36, 1986, pp. 229-255.