24

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A. I. Memo 763A **March, 1986**

THE COMBINATORICS OF LOCAL CONSTRAINTS IN MODEL-BASED RECOGNITION AND LOCALIZATION FROM SPARSE DATA

W. Eric L. Grimson

Abstract. The problem of recognizing *what* objects are *where* in the workspace of a robot can be cast as one of searching for a consistent matching between sensory data elements and equivalent model elements. In principle, this search space is enormous and to control the potential combinatorial explosion, constraints between the data and model elements are needed. We derive a set of constraints for sparse sensory data that are applicable to a wide variety of sensors and examine their characteristics. We then use known bounds on the complexity of constraint satisfaction problems together with explicit estimates of the effectiveness of the constraints derived for the case of sparse, noisy three-dimensional sensory data to obtain general theoretical bounds on the number of interpretations expected to be consistent with the data. We show that these bounds are consistent with empirical results reported previously. The results are used to demonstrate the graceful degradation of the recognition technique with the presence of noise in the data, and to predict the number of data points needed in general to uniquely determine the object being sensed.

Acknowledgments. This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the Laboratory's Artificial Intelligence research is provided in part by a grant from the System Development Foundation, and in part by the Advanced Research Projects Agency under Office of Naval Research contracts N00014-80-C-0505 and N00014-82-K-0334.

1. The Recognition and Localization Problem

A central characteristic of advanced applications in robotics is the presence of significant uncertainty about the identities and positions of objects in the workspace of the robot. In simplest terms, if a robot is to interact intelligently with its environment, it must know *what* objects are *where*. This normally necessitates sensing of the external environment as a means of obtaining the information needed to solve the recognition and localization problem. The process of sensing can be loosely divided into two stages: first, the measurement of properties of the objects in the environment, and second, the interpretation of those measurements. Since the sensory information could come from a variety of very different sources, for example, tactile, ranging, sonar or vision, both binary and grey-level, it is important to derive recognition and localization techniques that solve the interpretation stage of the sensing process with very few assumptions on the sensory measurements themselves. In this article, we assume only that the sensory data is characterized as sparse, noisy measurements of the local geometry of a small patch of the object's surface, for example, the position and orientation of a small planar patch of the surface in some coordinate frame defined relative to the sensor.

Given these simple data elements derived from the sensory data, the problem of model-based recognition and localization essentially can be considered as one of searching for a consistent matching between these data elements on the one hand, and model elements representing known objects on the other hand. That is, we need to assign each sensory data point to a corresponding part of an object model in a manner consistent with the assignment of the other sensory data. Since the data elements are assumed to approximate the local geometry of a small planar patch of the surface, initially we assume that the objects can be modeled as polyhedra which are defined relative to some local coordinate frame. The sensed objects corresponding to the models are assumed to have up to six degrees of positional freedom (three translational and three rotational). In other words, the coordinate frame transformation required to rigidly transform the model from its own local coordinate frame into the corresponding object being sensed in a coordinate frame defined relative to the sensor can have up to six degrees of freedom, three rotational and three translational. Note that three rotational degrees of freedom are required, since one needs two to specify the unit normal which defines the direction of rotation, and an additional degree of freedom is needed to specify the rotation about that unit normal. The goal is then to define a matching process whereby the space of possible interpretations of the sensory data can be searched for a consistent matching of sensory data to model elements. If such a matching can be found, it will then determine both the identity of the object being sensed and its position and orientation relative to the sensor.

Since, in general, the search space is far too large to explore explicitly, the key to the problem is to derive constraints based on the data that will efficiently restrict the portions of the search space that must be explored. In this paper, we present a set of general criteria on these constraints and derive a specific set of such constraints for the case of geometric sensor measurements. These constraints have previously been used in an empirical investigation of an object recognition system [Grimson and Lozano-Pérez 84, 85a, 85b]. Indeed, the remarkable efficiency of this recognition and localization technique on a broad range of real and simulated data was the original motivation for the work presented here. In this paper, we provide theoretical support for the performance of the recognition technique. In particular, we show that these constraints are complete both for the simpler case of three degrees of positional freedom, (two translational and one rotational) corresponding to the situation of isolated objects in stable positions, and for the general case of six degrees of positional freedom (three translational and three rotational).

The main result is establishing theoretical bounds on the effectiveness of these local constraints in controlling the combinatorics of the search process. We use known results on the complexity of constraint satisfaction problems together with explicit estimates of the effectiveness of the specific constraints to derive general bounds on the number of interpretations expected to be consistent with the data. These results are compared with empirical results reported earlier in [Grimson and Lozano-Pèrez 1984]. Finally, several predictions of the theory are discussed, including the degradation of the technique with increased error, and the number of sensory points generally needed to guarantee a unique interpretation of the data.

1.1 The Basic Problem

The recognition and localization of objects from sensory data is a central problem of most advanced robotics situations. It is usually convenient to pose the problem as one of search, that is, given a set of known models, we identify and locate the particular object that we are sensing by searching a large space of possible solutions until we find one (or all solutions) that matches the information available to us from the sensors. One of the main difficulties with the problem, true of most search problems, is that the space of possible solutions is usually enormous, and one seeks methods that will effectively reduce the portions of the search space that must be explicitly explored. The problem is further compounded in the case considered here by the fact that the sensory data against which a match is sought are typically inaccurate, so that the matching process must be tolerant to errors in the data.

In tackling the problem of object recognition, it is frequently convenient to concentrate on the subproblem of localization. If there exists a coordinate frame transformation that rotates and translates an object model in a manner that is consistent with the sensory data, then the existence of this transformation implicitly solves the recognition problem. Thus, if we have a technique for solving the localization problem, we can use it to solve the recognition problem as well. For example, we can use sequential search of a library of object models to determine which object is being sensed, as well as where it is in the coordinate frame of the sensor. Of course, this concentration on the localization portion of the general recognition problem ignores some important issues, especially when large libraries of objects are being considered, but it forms a sufficient starting point for investigating object recognition techniques. In the context of this paper, we will assume that the object recognition problem can be basically considered as a problem of object localization. The only real issue we are ignoring is the efficiency of recognizing an object from among a large set of objects.

There have been a wide variety of techniques applied to the recognition problem, all attempting in some manner to control the potential combinatorial explosion of the search. Much of the variation between existing recognition schemes can be accounted for by the choice of descriptive tokens to match. Some methods rely on computing a few very distinctive descriptors (*features*) that sharply constrain the identity and/or location of the object. Others use less distinctive descriptors and rely more on the relationships between them to effect recognition and localization.

The use of a few distinctive features sharply constrains the size of the search space, and hence the search process can be very efficient. As an extreme example, consider the problem of recognizing a soft drink can from visual data. One method would be to process the image to obtain the UPC bar code, which would uniquely identify the type of can. Moreover, knowing the position of the UPC code on the can and in the image would allow us to determine the position and attitude of the can in the scene. Of course, not all features will be as distinctive as a UPC code. Simpler examples might include corners, holes, notches and other local features. The idea, however, is that very few such distinctive features should be needed to identify the object, and the search space can be effectively collapsed. Examples of techniques in this vein include the use of a few extended features [Perkins 78, Ballard 81], or the use of one feature as a focus, with the search restricted to a few nearby features [Tsuji and Nakamura 75, Holland 76, Sugihara 79, Bolles and Cain 82, Bolles, Horaud and Hannah 84].

The approach of using a few distinctive descriptors is also common to many commercial systems (see, for example, [Bausch and Lomb 76, Gleason and Agin 79, Machine Intelligence Corporation 80, Reinhold and Vanderbrug 80]). These systems characterize both the measurements and the object models by a vector of global properties of binary images, such as area, perimeter, elongation and Euler number. The matching process between such characteristic vectors is then straightforward. Because of their global support, however, these descriptions do not extend well to overlapping or occluded parts.

Another type of recognition method relies on building elaborate high-level descriptions of the measured data before matching to the model. These approaches also rely on reducing the size of the search space by matching only a few distinctive descriptors. Examples of this approach include [Nevatia 74, Nevatia and Binford 77, Marr and Nishihara 78, Brooks 81, Brady 82].

Approaches that rely on a few distinctive features have some weaknesses. First, the cost of the search has been greatly reduced, but at the expense of global pre-processing of the sensory data. Sensors do not provide distinctive descriptors directly, the descriptors must be computed from the mass of local data provided by the sensor. In some sensing modalities, such as tactile sensing, searching for data to build distinctive descriptors can be very time consuming. Second, heavy reliance on a few features can make recognition susceptible to measurement noise. If the imaging device is out of focus, for example, so that the image of the UPC bar code is blurred significantly, recognition may be altogether impossible. In this case, degradation in the presence of error is not graceful, that is, the recognition process may fail badly when small amounts of error are present in the sensory data. Third, useful features are by definition sparse or they cease to be distinctive; this sparsity may be a problem when dealing with occlusion. Note that in this context, the term distinctive refers to the uniqueness of the features relative to the entire set of features of an object, not to their spatial distinctiveness. In our UPC bar code example, if some other object occludes the UPC code from the sensor, we will not be able to recognize the can. This may occur even though virtually all of the rest of the can is available to the sensor.

An alternative approach to recognition relies more on the geometric relationships between simpler descriptors, rather than on a few distinctive features. Such descriptors are densely distributed and not particularly distinctive taken individually; for example, surface normals fit into this category. In these circumstances, the search space is large and constraints to prune it are critical. While the size of the search space explored by these methods will be larger than in the feature–based methods, the expectation is that the individual tests are very efficient and if the constraints have sufficient power, the simplicity of the tests will offset the larger search space. Representative examples of such schemes include [Brou 83, Horn 83, Horn and Ikeuchi 83, Ikeuchi 83, Faugeras and Hebert 83, Gaston and Lozano-Pérez 84, Grimson and Lozano-Pérez 84, Stockman and Esteva 84].

The key difference between matching on these low–level descriptors and on distinctive features lies in the availability of descriptors. The simpler sensor measurements are likely to be dense over the object. As a consequence, recognition schemes based on such simple measurements should be applicable to sparse sensors, and should be less sensitive to problems of occlusion and sensor error, since an input description can always be obtained and matched to the model. In this paper, we explore some theoretical aspects of a recognition scheme that uses very simple sensor primitives that can be computed over the entire object [Grimson and Lozano-Pérez 84]. We rely on the power of geometric constraints to keep the combinatorics of the search process reasonably controlled.

1.2 Assumptions and Approach

As a consequence of this discussion, we will assume that the basic available sensory data consists of local estimates of three-dimensional position and orientation of small patches of the object surface. In this case, we can make very simple assumptions about the elements of the object models needed for matching. In particular, since the data elements are measuring the local geometry of small patches of the object surface, we assume that the object models are also constructed of small local patches. Thus, our two assumptions about the elements to be matched between sensory input and object models are:

- The objects are all modeled as polyhedra. The objects have six degrees of positional freedom, when transformed from a model based coordinate system to a sensor based coordinate system. In the simpler case of flat objects lying on a plane, the objects are modeled as polygons, which have three degrees of positional freedom.
- The sensory data available to the process include positions of points on the object, to within some known volume of error, and surface orientations at those points, to within some known cone of error.

The basic approach to the problem is to determine the set of positions and orientations of an object that are consistent with this sensed data. If there are no consistent positions and orientations, then the object can be excluded from the set of possible objects. The elements to be matched are thus simple local patches of a surface.

The technique proceeds in two steps:

- *Generate Feasible Interpretations:* A set of feasible interpretations of the sense data is constructed. Interpretations consist of pairings of each sensed point with some object surface on one of the known objects. Interpretations inconsistent with local constraints, derived from the model, on the sense data are discarded.
- *Model Test:* The feasible interpretations are tested for consistency with surface equations obtained from the object models. An interpretation is legal if it is possible to solve for a rotation and translation that would place each sense point on an object surface. The sensed point must lie *inside* the object face, not just on the surface defined by the plane equation.

There are several possible methods of actually searching for consistent matches. For example, Grimson and Lozano-Pérez [1984] chose to structure the search as the generation and exploration of an interpretation tree, more commonly known as backtracking. Thus, starting at a root node, we construct a tree in a depth first fashion, assigning data points to model faces. At the first level of the tree, we consider assigning the first data point to all possible faces, at the next level, we assign the second data point to all possible faces, and so on.

Clearly, the first of the two steps is the key to this process. The number of possible interpretations given s sensed points and m surfaces is m^s. Therefore, it is not feasible to explore the entire search space in order to apply a model test to all possible interpretations. Moreover, since the computation of coordinate frame transformations tends to be expensive, we want to apply this part of the technique only as needed. The goal of the recognition algorithm is thus to exploit local constraints on the sensed data so as to minimize the number of interpretations that need testing, while keeping the computational cost of each constraint small. In the case of the interpretation tree, we need constraints between the data elements and the model elements that will allow us to remove entire subtrees from consideration without explicitly having to search those subtrees.

In searching for appropriate constraints to apply to the generation stage, several criteria are appropriate.

- The constraints should be coordinate frame independent. That is, we would like to derive constraints that remove large portions of the search space, independent of the particular orientation of the object. This suggests that the constraints should embody restrictions due to the shape of the object, and not due to the specifics of the sensing geometry.
- The constraints should be simple and have low computational cost.
- The constraints should, at the same time, be as powerful as possible in the sense of removing large portions of the overall search space.
- The constraints should degrade gracefully in the presence of error in the sensory measurements.
- The constraints should be independent of the specifics of the sensor from which the data came, so that they will apply equally to different sensing modalities.

These criteria are very similar to those suggested by Marr and Nishihara [1978] (see also Marr [1982]).

We can illustrate the role of constrained search in solving the recognition and localization problem with a simple example. Consider the primitive two dimensional object illustrated in Figure 1, and the set of three sensory points, P_1, P_2, P_3, which are known

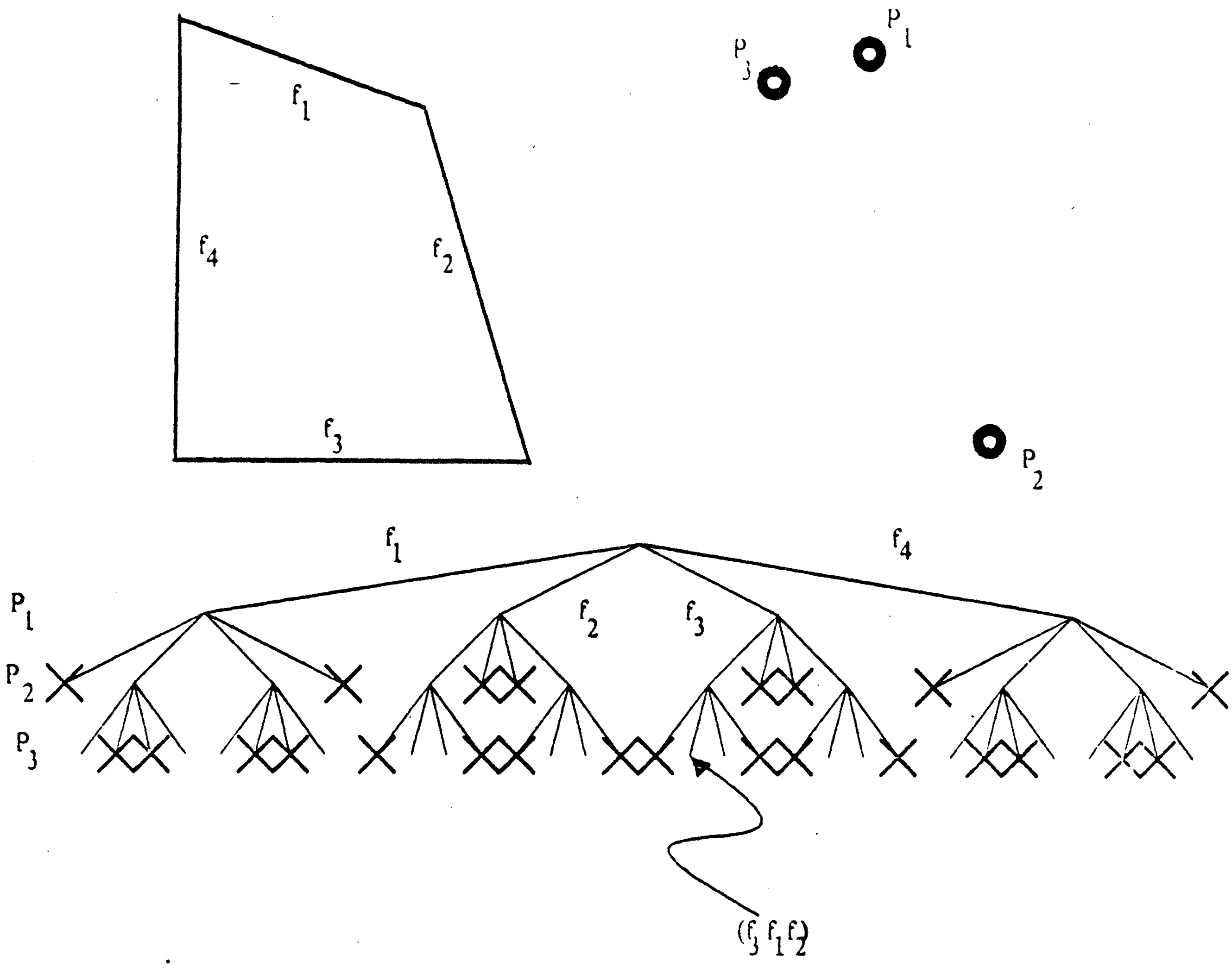

Figure 1. A simple example of constrained search. We want to find consistent matchings of the three data points to the edges of the indicated quadrilateral. If we only use distance between data points, then the table of possible ranges between the edges of the object is given by

	1	2	3	4
1	[0, 1.5]	[0, 3.25]	[2, 3.25]	[0, 2.5]
2	[0, 3.25]	[0, 2]	[0, 2.5]	[1.4, 3.25]
3	[2, 3.25]	[0, 2.5]	[0, 2]	[0, 3.25]
4	[0, 2.5]	[1.4, 3.25]	[0, 3.25]	[0, 2.5]

The tree indicates the set of possible assignments of data points to object edges, given distance as the only constraint. One can see that only 16 out of 64 possible interpretations are consistent with this very simple constraint.

to lie on the edges of the object. To find the possible interpretations of the data, we need to determine the set of possible assignments of data points to model edges. For the sake of illustration, we will further assume that the only constraint available to us is that of distance, i.e. we can measure the distance between the sensory data points, and that distance must be consistent with the object model. The table in Figure 1 indicates the ranges of possible distances between pairs of edges of the model. That is, for each pair of edges in the model, we determine the range of possible distances that can arise between two points, one on each edge, as the points are taken through all possible positions on their respective edges.

Now, suppose that the measured distances between the three sensed points are $dist(P_1, P_2) = 2.6$, $dist(P_1, P_3) = 0.6$, and $dist(P_2, P_3) = 2.7$. From this we can see, for example, that any interpretation of the sensed points that assigns P_1 and P_2 both to edge f_1 is inconsistent with the model, since the feasible range is only from 0.0 to 1.5. The tree structure of Figure 1 indicates those branches of the interpretation tree that correspond to feasible interpretations, as well as those that can be terminated due to inconsistencies. For example, the indicated node of the tree corresponds to assigning sensory point P_1 to edge f_3, point P_2 to edge f_1 and point P_3 to edge f_2.

One can see that only 16 of the 64 possible interpretations are consistent with this very simple constraint of distances between points. Of course, even fewer interpretations are completely consistent. For example, if we were also to include angles between the surface normals at the sensed points, this additional constraint would further reduce the number of consistent interpretations. In the following sections, we look at ways to extend this simple example in a coherent manner.

2. A Specific Set of Local Constraints

The example of the previous section indicated the kind of constraint of interest to us in solving the recognition and localization problem. We begin our analysis by deriving a more detailed set of coordinate-frame-independent constraints, which were first presented in [Grimson and Lozano-Pérez 1984]. To do this, we first consider what types of coordinate-frame-independent constraints are possible, given that the sensory data is characterized as sparse points, each consisting of a position measurement and a unit surface normal (for example, the pair $(\mathbf{p}_1, \mathbf{n}_1)$ in Figure 2). Clearly a single data point provides no constraint on the possible faces from the model that could consistently be assigned to it. Thus, we look at pairs of sensory points. The basic information available from these points consists of a pair of unit normals at those sensed points, as well as the vector separating their bases, as shown in Figure 2.

One way to get coordinate-frame-independent constraints is to construct a local coordinate frame relative to the configuration itself. Thus, we can use each of the two unit normals as axes of the local coordinate frame. In two dimensions, these define a local system, except in the degenerate case of the unit normals being (anti-)parallel. In three dimensions, the third component of the local coordinate frame can be taken as the unit vector in the direction of the cross product of the first two basis vectors. Given such a local basis, clearly one set of coordinate-frame-independent measurements provided by the configuration of Figure 2 is the components of the separation vector along each of

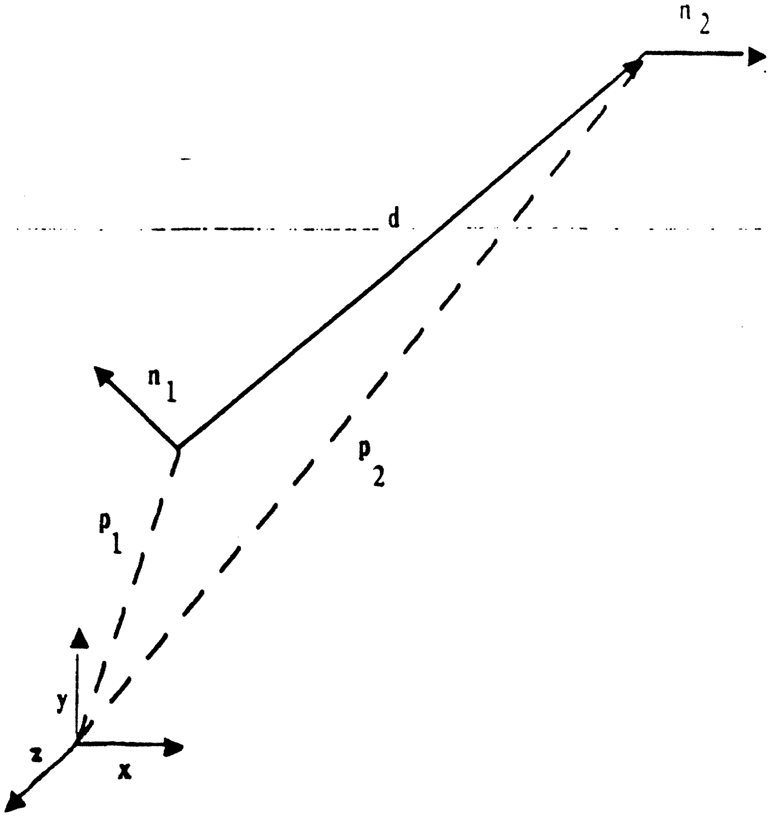

Figure 2. The constraints between pairs of sensory points. The sensed points are indicated by the endpoints of their local unit normals, $\mathbf{n}_1$ and $\mathbf{n}_2$, and are separated by a vector $\mathbf{d}$. This whole configuration is placed in some global coordinate frame, specified by the orientation and position of the coordinate frame $\mathbf{x}, \mathbf{y}, \mathbf{z}$.

the basis directions. (Note that the use of the distance and two of the components is equivalent, up to a possible sign ambiguity, to using the three components of the vector.) Additionally, the angle between the two basis vectors is also specific only to the local coordinate frame. More formally, if the unit normal vectors are denoted by $\mathbf{n}_1, \mathbf{n}_2$ and the vector separating the two points is $\mathbf{d}$, then one set of coordinate-frame-independent measurements of this configuration is

$$\begin{gathered}\mathbf{n}_1 \cdot \mathbf{n}_2 \\ \mathbf{d} \cdot \mathbf{n}_1 \\ \mathbf{d} \cdot \mathbf{n}_2 \\ \mathbf{d} \cdot \mathbf{n}_{12}\end{gathered}$$

where $\mathbf{n}_{12}$ is a unit vector in the direction of $\mathbf{n}_1 \times \mathbf{n}_2$. (Note that these measurements are isomorphic to the set used in [Grimson and Lozano-Pérez 1984].)

These are coordinate-frame-independent measurements on the configuration defined by a pair of sensory points. To turn them into constraints on the search process, we must map them into equivalent measurements on the model elements. Since each object is modeled as a complex polyhedra, the mapping is fairly straightforward. For example, consider the first measurement, $\mathbf{n}_1 \cdot \mathbf{n}_2$. In order for these two sensory points to be consistent with a pair of faces on an object, the dot product of the corresponding face normals must agree with this measurement. Thus, by precomputing the angles between all pairs of faces on an object, this sensory measurement can be used to constrain the search for a consistent interpretation. In particular, if a pair of points is inconsistent with a particular assignment of faces to those points, the entire subtree of the interpretation tree lying below the node corresponding to that assignment can be ignored, thereby reducing the amount of searching required. Similar constraints can be derived for the components of the separation vector in the directions of the unit normals. That is, for each pair of faces in the model, one can precompute the range of values of the component of a vector in the direction of one of the face normals as that vector assumes all possible positions having one endpoint on the first face and the other endpoint on the second. Again, for an assignment of sensory points to faces to be consistent, it must be the case that the coordinate-frame-independent sensory measurement must agree with the precomputed model values. Thus, we have defined one possible matching process between sensory measurements and models, similar to that presented in [Grimson and Lozano-Pérez 84].

As stated, such constraints ignore the possibility of noise in the sensory data. Consider, for example, the dot product between two sensed surface normals. We have stated that for these sensory measurements to be consistent with an assignment to a pair of faces from an object model, the dot product between them must be the same as the dot product between the model faces. If there is error in the sensory measurements, however, we must account for it. Given bounds on the amount of error in the measurements, it is possible [see Grimson and Lozano-Pérez 84] to compute ranges of possible values for the true sensory measurement, given the recorded sensory measurement. In this case, rather than requiring exact agreement between the recorded dot product between the sensed normals and the dot product of the model normals, the constraint simply requires that the range of possible dot products, given the recorded dot product, include the dot product between model normals. Similarly, the constraints based on the components of the vector separating the sensed positions in the directions of the sensed normals can be relaxed to allow for bounded amounts of error. For a geometric derivation of these constraints, see [Grimson and Lozano-Pérez 84]. For the purposes of this article is suffices to note that such simple geometric measurements can be modified to account for error in the original sensory measurements.

3. A Practical Context for the Constraints

Before we proceed to the heart of the paper, which is a theoretical investigation of the use of such constraints in object recognition and localization, it is important to briefly outline the practical context in which such constraints can be used. This is especially true since several simplifying assumptions will be made to ease the theoretical analysis, and this can potentially obscure the practical importance of recognition algorithms based on the constraints outlined above.

We have outlined a set of coordinate-frame-independent constraints based on the geometric relationship between pairs of sensed points, each of which is defined as the measurement of the position of a point of contact on an object and the local surface orientation at that point of contact. We can use these constraints in a straightforward manner to search for consistent matchings between the sensed points on the one hand and faces of the object model on the other [Gaston and Lozano-Pérez 84, Grimson and Lozano-Pérez 84]. For example, we can use backtracking, or depth first search, to explore a tree of interpretations, as in Figure 1, using the constraints to remove entire subtrees of the search space from consideration when one of the constraints is violated. In this manner, any leaves of the interpretation tree that also pass the model test are legitimate interpretations of the sensory data.

There are a number of assumptions inherent in this recognition scheme, however. For example, we have developed constraints on the matching process based on pairs of points. One could naturally ask whether constraints based on triples or higher order collections of points are possible. While the answer to this is undoubtedly yes, we have chosen to concentrate solely on pairwise constraints for practical reasons. In particular, given constraints based on pairs of model elements, we only require a set of m^2 tables to represent each object, where m is the number of faces in the model. For constraints based on triples of elements, we would require a set of m^3 tables, and so on. Hence, if the simple constraints based on pairs of points are sufficiently powerful in solving the recognition problem, the reduced overhead in representing object models is to be preferred.

We have also assumed that all of the sensory data comes from the same object. In situations of objects in isolation, this may be reasonable, but for most robotics applications such an assumption is overly restrictive. We need recognition techniques that can also deal with occluded objects. In this case, the sensory data may come from several objects, and there will be no single polyhedral interpretation that can consistently account for all the data points. There are ways to extend the constraint satisfaction technique to find interpretations of the data in this case. One is described in [Grimson and Lozano-Pérez 85a, 85b] together with empirical evidence in support of it.

We are also assuming that the sensory data is metrically accurate, that is, the object model is of a fixed size. This assumption can be relaxed to allow for scale factors in the sensory data. In particular, [Grimson and Lozano-Pérez 85a, 85b] report on a natural extension to the technique that allows the sensed object to be an arbitrarily scaled version of the model, and automatically solves for the magnification, or scale, factor as well as the location of the object.

For the purposes of the analysis of this paper, we will restrict our attention to isolated, metrically accurate objects. Extensions of the results to the more general case will depend on the specific methods used to deal with occluded objects.

4. The Constraints Satisfy Our Criteria

We begin our theoretical investigation by examining the characteristics of the constraints derived above. We will then consider their combinatorial power and their degradation with error.

4.1 Coordinate-Frame-Independence

By the specifics of the derivation, the constraints are coordinate-frame-independent. Moreover, they also satisfy our requirement of simplicity. Obtaining the sensory half of the constraint is straightforward, and the model half of the constraints can be precomputed directly from the model. The matching process itself then becomes a simple table-lookup process, which also satisfies the notion of simplicity and computational ease.

4.2 Completeness of the Constraints

While the constraints derived above meet our basic criterion, it is important to also demonstrate that they form a complete set, that is, there are no other independent coordinate-frame-free constraints between pairs of sensory points that are not already incorporated in these particular constraints. This can be easily established by the following equation counting argument. Consider the configuration illustrated in Figure 2. Suppose we construct some arbitrary global coordinate frame, having three rotational degrees of freedom ϕ, θ, ψ and three translational degrees of freedom, given by the components of the vector $\mathbf{p}_1$, relative to the described configuration. To completely describe the configuration shown will require a total of 10 equations, three to specify the base position $\mathbf{p}_1$, three more to specify the separation vector $\mathbf{d}$, and two each to specify the relative attitudes of $\mathbf{n}_1, \mathbf{n}_2$. These ten equations are explicit functions of several parameters, including the three angular and three translational degrees of freedom described above. Thus, to reduce this set of 10 constraints to a set of coordinate-frame-independent constraints, we must resolve the set of equations. In other words, we must remove the explicit dependence on the parameters related to the specific choice of global coordinate frame. Clearly this will require at least six equations, and thus there are at most four coordinate-frame-independent constraints given by this particular pairwise configuration of sensory points.

In the simpler case of two dimensions, we find that there are three coordinate-frame-independent constraints. Thus, we see that in both cases, the constraints outlined above are complete.

5. A General Theoretical Basis

The key theoretical issue still to be settled is the combinatorial power of the local constraints in reducing the number of consistent hypotheses. By its formulation as a search process, our problem of object recognition can be considered as one of constraint satisfaction, or consistent labeling. Hence, there are some general results available to us [for example, Nudel 83, Haralick and Elliot 80, Gaschnig 79]. In particular, general bounds on the expected number of solutions, on the expected number of consistency checks performed at each level of the search tree and on the expected number of consistent nodes at each level of the tree are known. Our goal is to derive explicit values for these bounds in the case of our particular constraints and show as a consequence that our constraint satisfaction technique for object recognition will generally be extremely efficient.

5.1 General Bounds

To establish bounds on the effectiveness of our constraints in handling the problem of object recognition and localization, we will rely on known general results. To do this, we quickly review the results reported in [Nudel 83] and using the notation developed there, (see the bibliography in [Nudel 83] for additional references.)

Let $\mathbf{V} = (v_1, \ldots, v_s)$ be a list of variables, which in our case will represent the sensed data points. Each variable has an associated domain $D_i = (v_{i1}, \ldots, v_{iM_i})$ from which it can take any of M_i values or labels. In our case, the domains are all of the same size, $D_i = D = (1, \ldots, m)$ – the labels of the faces of the model. In general, a set of constraints $\mathbf{R} = \{R_1, \ldots, R_k\}$ consists of relations that specify which values are mutually compatible for various subsets of the variables,

$$R_j \subseteq D_{i_1^j} \times D_{i_2^j} \times \ldots \times D_{i_{r_j}^j}.$$

Our case is particularly simple since all of the constraints are binary $(r_j = 2)$.

Binary constraint labeling problems can be represented by their relations matrix $[T_{kl}^{ij}]$ (undefined for $i = j$), a bit matrix such that $T_{kl}^{ij} = 1$ if and only if the k^{th} value for variable i is consistent with the l^{th} value for variable j. In our case, this translates as saying that the i^{th} data point can be assigned to the k^{th} face of the model and the j^{th} data point to the l^{th} face in a consistent manner.

If we assume that

$$\text{Prob}\left(T_{kl}^{ij} = 1\right) = p_{ij}$$

then one can show [Haralick and Elliot 80, Nudel 83] for backtracking that the expected number of solutions is

$$m^s \left(\prod_{i<j} p_{ij} \right)$$

where m is the number of faces in the model, and s is the number of sensed points. If $p_{ij} = p$ for all i and j, this reduces to

$$I_{\text{exp}} = m^s p^{\binom{s}{2}}. \tag{1}$$

At the k^{th} level of the search tree, the expected number of checks performed is

$$m^k p^{\binom{k-1}{2}} \frac{1 - p^{k-1}}{1 - p}.$$

The expected number of nodes at the k^{th} level is simply

$$m^k p^{\binom{k-1}{2}}.$$

As a consequence, the total number of expected nodes in the search tree is

$$\sum_{k=1}^{s} m^k p^{\binom{k-1}{2}}$$

and the total number of checks expected to be performed is

$$\sum_{k=1}^{s} m^k p^{\binom{k-1}{2}} \frac{1 - p^{k-1}}{1 - p}.$$

Even without explicit values for the probability p these results provide some useful information. For example, one can determine when the expected number of interpretations is a maximum as a function of the number of sensed points. A straightforward application of the calculus shows that this occurs for

$$s = \frac{1}{2} - \frac{\log m}{\log p}. \tag{2}$$

Secondly, one can determine the number of data points needed to reduce the expected number of interpretations to at most one. Again, a straightforward application of the calculus shows that this occurs for

$$s \geq s_{\text{min}} = 1 - \frac{2 \log m}{\log p}. \tag{3}$$

We will use equation 1 and 3 in Tables 1, 2 and 3 when we compare the predictions of the theoretical results with empirical evidence obtained by apply a recognition algorithm to real data.

These are general results, true of arbitrary consistent labeling problems. To apply them to our problem of object recognition and localization, we need to estimate the probability p of consistency between pairs of data points. We will do this by applying the useful tool of a Relative Configuration (RC) Space.

5.2 Relative Configuration Space (RC-space)

We need to provide estimates for the probability that two faces can be assigned consistently to two data points of the model. To do this, we need a way of mapping both the sensory data and the object models into some common space in which they can be compared.

The space we will use for doing this is a relative configuration space (RC-space). Throughout the analysis we will consider two separate cases. The first is one in which the objects are non-overlapping and lie in stable positions. This case is essentially a two dimensional one, and the objects have three degrees of positional freedom relative to their models, two translational and one rotational. The second case is one in which the objects are arbitrarily oriented in space. This is a three dimensional problem and the objects here have six degrees of positional freedom relative to their models, three translational and three rotational.

In the two dimensional case, we define an RC-space in the following manner. Each face of an object is defined by a two-dimensional position (given by the position of the midpoint of the face) and an orientation (given by the orientation of the normal of the face). To ease the analysis, we will assume that all of the edges of the object (or polygon) have equal length ℓ. For a three dimensional object, we assume that all of the faces are squares of side ℓ.

Given a face f_i assigned to the first point P_1, we can define a three dimensional coordinate system relative to this face, consisting of two coordinates of spatial extent and one of angular extent. The origin of the two spatial dimensions is set at the midpoint of the base face (f_i) (with the face extending along the x axis and with the normal of the face pointing in the $-y$ direction) and the origin of the angular dimension is set to the orientation of the normal of the base face. Thus we have defined a configuration space relative to the orientation of a particular face. The position of another face is completely specified by the position and orientation of its midpoint in this RC-space.

In the three dimensional case, we construct a 5 dimensional relative configuration space, based on the first face f_i. This RC-space has three positional dimensions and two orientation dimensions (since only two angles are needed to specify the orientation of a unit vector). Note that in this case, we cannot determine the rotation about the normal of the first face. This becomes a free parameter in our analysis, and hence we need only consider a five degree of freedom relative configuration space.

In this case, the origin of the spatial components is defined to be the midpoint of the base face, with the surface normal pointing in the $-z$ direction. The edges of the face are aligned with the x and y axes. Thus, the position of any other face is given by the position of its midpoint in this RC-space. The rotational components of a face are defined by two different angles, $-\pi < \omega \leq \pi$ and $0 \leq \phi < \pi$, where ϕ describes the elevation of the unit normal relative to the $-z$ axis, and ω describes the orientation of that unit normal in the $x-y$ plane.

To obtain bounds on the effectiveness of local constraints in pruning the search space of feasible interpretations, we need to map both the models and the sensory information into this RC-space. By enumerating the intersection of these two mappings, we will be able to analyze the combinatorial efficiency of the constraints, and determine the probability p of equation 1.

Clearly, given a base face for the RC-space, each additional face of the model is represented by the position of its midpoint in this space, and that is given by the spatial offset of its midpoint relative to the midpoint of the base face, and the angular variation of its normal, relative to the normal of the base face. Thus, each face projects to a point in this RC-space, and the entire object is given by a scattered collection of such points.

Consider now what the constraints on the sensory measurements tell us about the positions of feasible faces in this RC-space. With no constraint, the faces of the object corresponding to a sensory data point could lie anywhere within the RC-space. The constraints on distance and relative angle, however, will restrict the set of feasible faces that can be assigned to a data point to lie within a reduced volume in RC-space. Thus, given the assignment of the first point P_1 to face f_i, the face corresponding to the second data point, P_2, can be restricted to lie within some finite volume of RC-space. Clearly, this should generally place a restriction on the number of faces consistent with the sensory data, and our goal is to obtain bounds on this restriction, by considering the characteristics of this restricted volume in RC-space.

In more detail, we consider what the effects of the constraints are in the case of three degrees of positional freedom. In particular, given a second sense point P_2, we know the following facts.

- **Distance.** The measured distance between the two sensed points clearly restricts the position components of the second face to lie within a restricted area. This follows from the observation that if the base of the vector connecting the two sensed points must lie somewhere on the first face, and the length of the vector is given by a measured distance d, which is accurate to within some range ϵ, then the set of positions in which the midpoint of the second face can lie is restricted to lie within an area specified by d and ϵ. (We will see later one method for analytically specifying this area.)
- **Angle of the vector.** The measured components of the vector between sensed points relative to the measured surface normals define the angle, θ, between those vectors. This angle is subject to a range of values, which in this case is a function of d, ϵ and γ, the bound on errors in measuring surface normals.

- **Angle of the normal.** The angle between the surface normals at the first and second faces can be restricted to some measured value ψ plus an error range defined by γ.

As a consequence of these restrictions, we see that if the first point P_1 lies somewhere on the base face f_i, then the second point P_2 must lie within some volume

$$\mathbf{V}(d, \epsilon, \ell, \theta, \psi, \gamma)$$

of RC-space. Note that three of the parameters defining this volume are measured values and three are global parameters set by the object and the error sensitivities of the measuring device. To reflect this, we rewrite this volume of RC-space as

$$\mathbf{V}_{\epsilon,\ell,\gamma}(d, \theta, \psi).$$

Note that this symbol represents a region in a three-dimensional RC-space. We will use

$$v_{\epsilon,\ell,\gamma}(d, \theta, \psi)$$

to denote the magnitude of this volume.

Given that face f_i is initially assigned to the first sensed point, we let

$$\rho_i(x, y, \phi)$$

denote the distribution of faces relative to this assignment. Thus, ρ_i is a sum of m delta functions, each of which reflects the configuration of a face relative to f_i, and

$$\int_x \int_y \int_{\phi=0}^{2\pi} \rho_i(x, y, \phi)\, dx\, dy\, d\phi = m.$$

If we let $d_{ij}, \psi_{ij}, \theta_{ij}$ denote the sensed measurements between points P_i and P_j, then the number of faces that can be consistently assigned to a second data point, P_2, given face f_i is assigned to P_1 is bounded by

$$\int\int\int_{\mathbf{V}_{\epsilon,\ell,\gamma}(d_{12},\theta_{12},\psi_{12})} \rho_i(x, y, \phi)\, dx\, dy\, d\phi.$$

We could use this approach directly to estimate the expected number of interpretations (see, for example, [Grimson 84]). However, we really want to estimate p, the probability that

$$\int\int\int_{\mathbf{V}_{\epsilon,\ell,\gamma}(d_{12},\theta_{12},\psi_{12})} \rho_i(x, y, \phi)\, dx\, dy\, d\phi \geq 1$$

for all i, and for any pair of data points P_1, P_2.

There are several ways we could do this. We choose to seek expected bounds, that is, bounds that will hold in general, so that the class of objects for which the bounds are exceeded is small, and hopefully, has measure zero, yielding an almost everywhere bound. This relies in some sense on a type of limiting behavior assumption as follows. Suppose we consider k different polygons, each with m faces and diameter D, where the diameter of an object is defined as the maximum distance between any two points lying on the object. Then in the limit as k tends to ∞, the distribution of the mk faces of the ensemble of objects tends towards a uniform distribution. If we then average over k, we can assert that the limiting distribution of faces is also uniform, over an area defined by the diameter of the object, D (which is defined as the maximum distance between any pair of points on the object).

Even for single objects, this assumption of uniform distribution may not be badly violated. Consider the simple, two-dimensional object illustrated in Figure 3(a). The

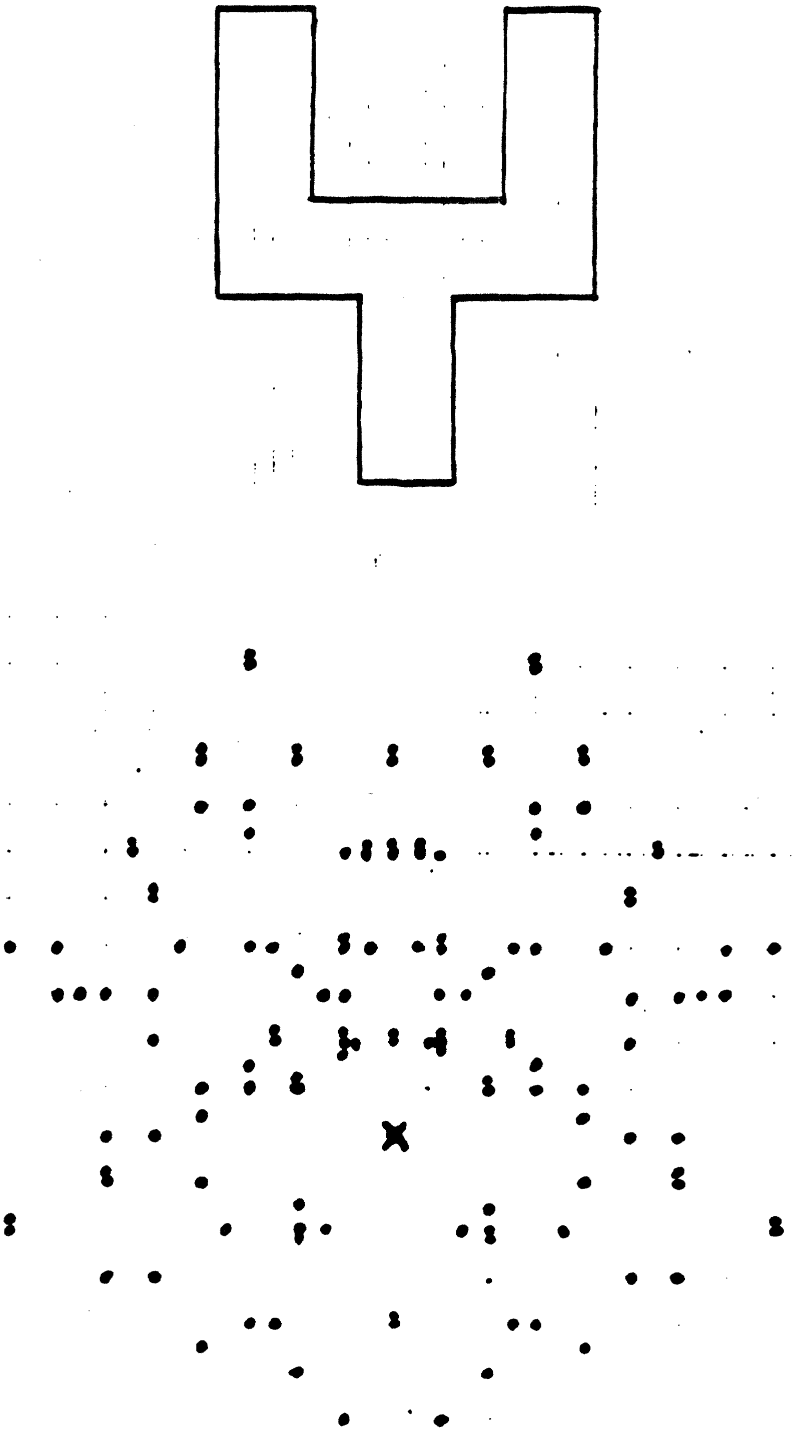

Figure 3. A sample object and its distribution of faces in relative configuration space. Each point in the bottom figure corresponds to the position of the midpoint of an edge of the object. The entire collection of points was obtained by assigning the midpoint of each edge to the origin of the RC-space, accumulating the midpoints of all the edges relative to that edge, and accumulating these collections of points as all the edges were assigned to the origin of the space. One can see that the distribution of faces roughly fills a disk about the origin of RC-space.

spatial distribution of midpoints of edges of this object in relative configuration space is illustrated in Figure 3(b). This distribution was obtained in the following manner. First, the midpoint of each edge was marked. Then the midpoint of the first edge was aligned with the origin of the space, with the normal to the first edge pointing along the $-y$ axis. Having aligned the object in this way, the positions of all of the other midpoints were marked in the space. This was repeated, aligning the midpoint of the second edge at the origin, and marking the positions of the other midpoints, and so on.

Several things are noticeable in this diagram. The first is that the relative midpoints of the faces do tend to uniformly fill the disk of possible positions. There are some clusters, however, and these generally correspond to partial symmetries in the object. This suggests that one set of objects that will badly violate our assumption of uniform distribution is regular polyhedra. For example, a regular n-gon, because of its n-fold symmetry, will have a face distribution consisting of n clusters of n indistinguishable points, and thus the number of consistent hypothesis should be higher than predicted by the theoretical analysis. This is in fact observed in simulations reported in [Grimson and Lozano-Pérez 84].

Note also that while the spatial components of the RC-space are reasonably uniformly distributed for the object in Figure 3, since all the angles between faces are multiples of $\frac{\pi}{4}$, the rotational component of RC-space will only contain points clustered at $0, \pm\frac{\pi}{4}, \pm\pi$. That is, the spatial components of RC-space basically meet the assumption of uniformity while the rotational components do not. On the other hand, a regular polygon would have rotational RC-space points that are uniformly distributed, while the spatial components are strongly clustered. This suggests that in cases where only some of the components of RC-space satisfy the criteria of uniform distribution, the expected number of consistent hypotheses will be higher than that predicted.

The key point of this assumption of uniform distribution is that the integral of the distribution function will depend only on the magnitude of the swept volume, and not on its specific position in RC-space. Suppose we let v_T denote the magnitude of the total volume of RC-space that could possible contain a point corresponding to an object face. This is given by

$$v_T = \left[\pi D^2\right] 2\pi$$

and is found by considering the area swept out by rotating the object about one end point of its diameter, which denotes the total possible range of the translation components (πD^2), and the total range of the rotation component (2π).

As a consequence, the expected number of faces consistent with data points P_1 and P_2 is

$$E\left(\int\int\int_{\mathbf{V}_{\epsilon,\ell,\gamma}(d_{12},\theta_{12},\psi_{12})} \rho_i(x,y,\phi)\, dx\, dy\, d\phi\right) = m\,\frac{v_{\epsilon,\ell,\gamma}(d_{12},\theta_{12},\psi_{12})}{v_T}.$$

Moreover, the assumption of a uniform distribution implies that this value is not dependent on the choice of the first face f_i. Thus, if we can bound the magnitude of the volume

$$v_{\epsilon,\ell,\gamma}(d_{12},\theta_{12},\psi_{12})$$

independent of the specific sensory measurements, say,

$$v'_{\epsilon,\ell,\gamma} = \sup_{d,\theta,\psi}\left[v_{\epsilon,\ell,\gamma}(d,\theta,\psi)\right],$$

then under the assumption of uniform distribution, the probability of consistency is simply given by

$$p = \left[\frac{v'_{\epsilon,\ell,\gamma}}{v_T} \right]. \tag{4}$$

In other words, by assuming that the faces of the object are uniformly distributed, we can assume that their corresponding representation in RC-space consists of a uniform distribution of points in the total volume of the space. Thus, the probability that two faces will be consistent with assignments to a pair of sensed points is simply given by the ratio of the two volumes, as indicated in equation 4.

6. Specific Bounds

We now turn to a careful consideration of the specific constraints that apply to our particular case, and derive explicit expected bounds on the number of interpretations consistent with s sensory points, by estimating the volume parameter

$$v_{\epsilon,\ell,\gamma} = |\mathbf{V}_{\epsilon,\ell,\gamma}|.$$

6.1 The Two Dimensional Case

We will first consider the two-dimensional case of objects with three degrees of positional freedom, and will then extend this to the general three dimensional case of objects having six degrees of positional freedom.

First, we will make the following assumptions about the object being sensed.

- All the edges of the object have the same length ℓ.
- The diameter of the object, that is the maximum distance between any pair of points on the object, is denoted by the constant D.

We make the following assumptions about the sensory information.

- The positions of the sensed data points are known to within a circle of radius ϵ.
- The normals to the sensed edges are known to within an angular error of γ.

Now suppose that we arbitrarily choose some face on which to place the first sensed point (obviously we have no constraints on this anyway). Given a second sensed point, we want to consider the probability that some other face is consistent with the constraints between the two sensory points. To answer this, we first review what constraints are available from two sensory points.

- We can measure (to within some error) the angle ψ_0 between the normal of the first face and the normal of the second face. This is given by the combination of knowing the dot product between the two normals and the cross product of the two normals. In fact, this measurement is accurate to within a range of $\pm\gamma$, by our assumption above.
- We can measure the length d of the contact vector $\mathbf{d}$ between the two sensed points (with a possible error of $\pm\epsilon$).

- Finally, we can determine the angle between the contact vector **d** and the first face normal. To determine the range of possible error, we note that each endpoint of **d** lies within a circle of radius ϵ. It is straightforward to show, either algebraically or geometrically, that the maximum angular deviation of the true contact vector from its measured position is given by

$$\tan^{-1}\left(\frac{2\epsilon}{d}\right).$$

Our general technique in estimating bounds on the combinatorial efficiency of these constraints will be the following:

- Construct a relative configuration space (RC-space) centered at the midpoint of the first face.
- Bound the volume of this space in which the second face must lie.
- Normalize this volume relative to the entire usable volume of RC-space in order to obtain a bound on the probability of consistency p.

We arbitrarily assign a face to the first sensed point, and construct a relative configuration space about this face. Recall that the origin of the spatial dimensions of the RC-space lies at the midpoint of the face and the normal of this face defines the origin of the angular dimension of the RC-space. Thus, a face is represented in RC-space by the position of its midpoint relative to the midpoint of the base face, and by the angle of its normal relative to the normal of the base face.

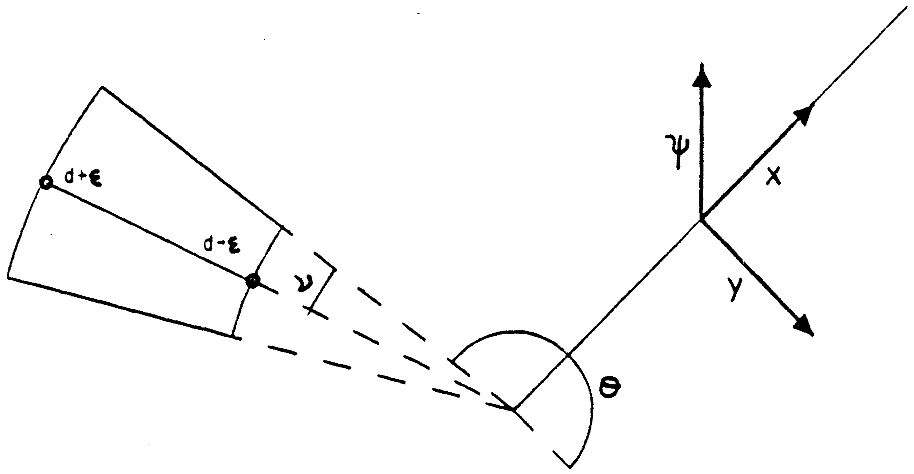

Figure 4. The range of possible positions for a second sensory point, given that the first sensory point lies at one end of the base face.

Suppose that the base of the contact vector **d** lies at one of the endpoints of the first face. What positions could the endpoint of **d** take in RC-space? First, we know that the angle of this vector relative to the sensed normal to within a range $\tan^{-1}(2\epsilon/d)$ and we know that the variation in the sensed normal relative to the face normal is given by γ. Thus, the angle θ made by the contact vector relative to the face normal is bounded by the range

$$\nu = \tan^{-1}\left(\frac{2\epsilon}{d}\right) + \gamma.$$

Thus the area of RC-space swept out by the endpoint of **d** is given by

$$\int_{\rho=d-\epsilon}^{d+\epsilon}\int_{\mu=\theta-\gamma-\tan^{-1}\frac{2\epsilon}{d}}^{\theta+\gamma+\tan^{-1}\frac{2\epsilon}{d}} \rho\, d\rho\, d\mu$$

which evaluates to

$$4\left(\gamma + \tan^{-1}\frac{2\epsilon}{d}\right)\epsilon d = 4\nu\epsilon d.$$

This is illustrated in Figure 4.

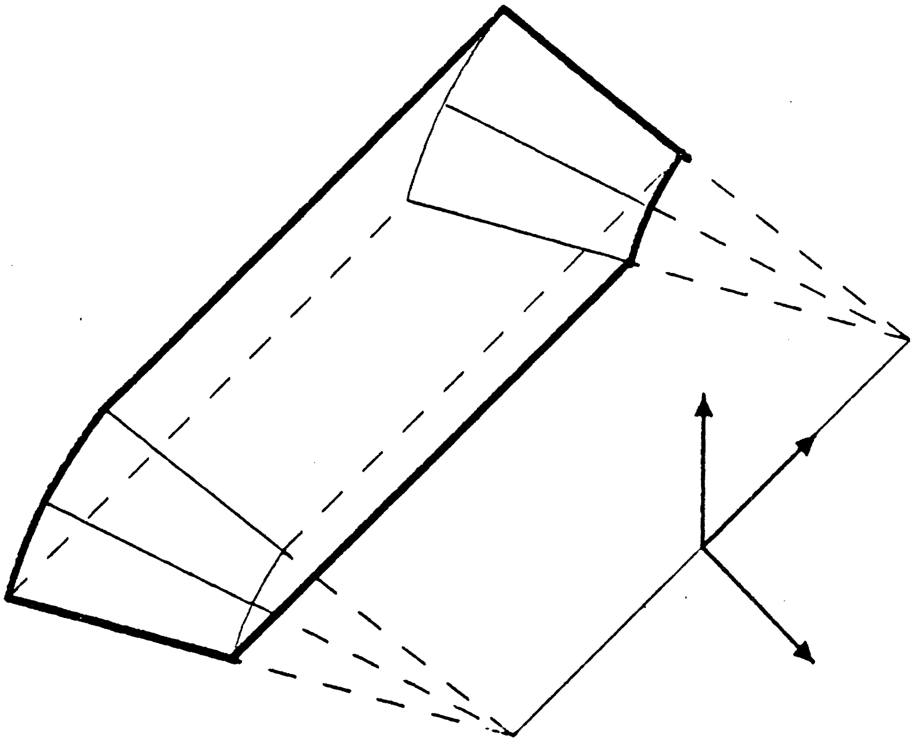

Figure 5. The range of possible positions for a second sensory point, given that the first sensory point lies anywhere on the base face.

But the first sensory point, P_1, could lie anywhere on the first face. The total area of RC-space swept out as the base point of **d** is moved across the base face is diagrammed in Figure 5.

We could simply integrate the area swept out by this process to obtain an expression for the probability of consistency. It is somewhat easier, however, to simply consider the problem of enscribing a rectangle, with sides aligned along the coordinate axes of the RC-space, about this volume. For example, the bounding rectangle for the swept area of Figure 5 is indicated in Figure 6.

We will assume that $\gamma \leq \frac{\pi}{4}$ and that we only consider contact points such that $d \geq 2\epsilon$. In this case, $0 \leq \nu \leq \frac{\pi}{2}$. We may assume, without loss of generality, that θ lies

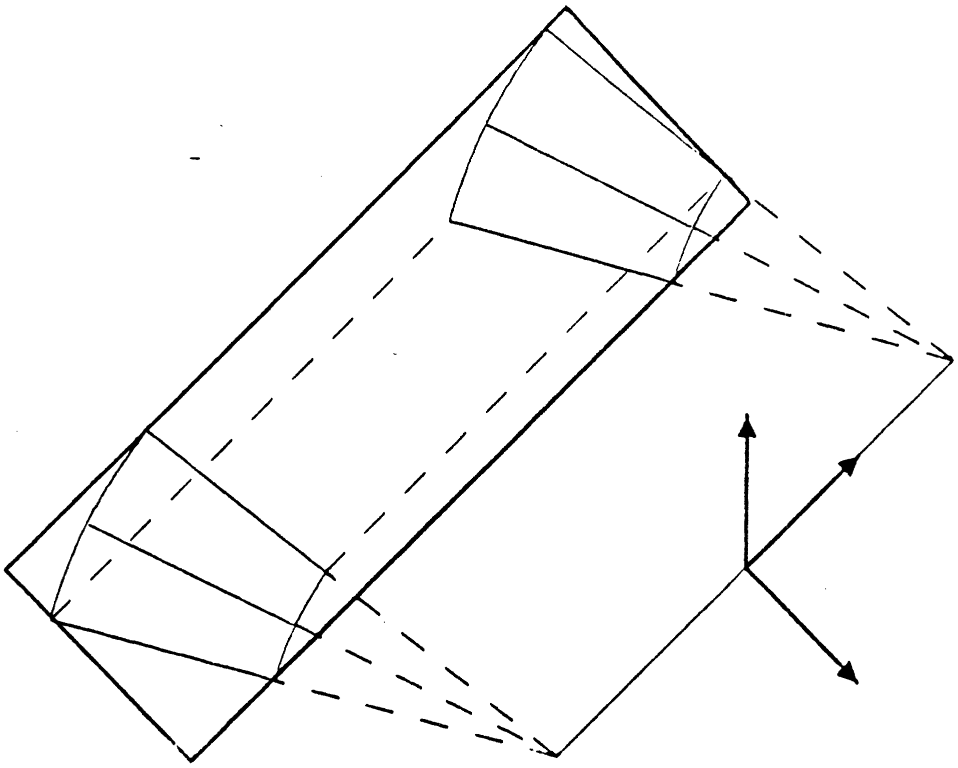

Figure 6. A bounding rectangle for the range of possible positions for a second sensory point, given that the first sensory point lies anywhere on the base face. This particular bounding rectangle corresponds to the case of the swept area of Figure 5.

in the range $[0, \frac{\pi}{2}]$. In order to determine bounds on the dimensions of the enscribing rectangle, we consider three case.

Consider first the case of $\theta > \frac{\pi}{2} + \nu$. In this case, one can show that

$$\begin{aligned}
y_{\mathrm{min}} &= (d - \epsilon)\sin(\theta + \nu) \\
y_{\mathrm{max}} &= (d + \epsilon)\sin(\theta - \nu) \\
x_{\mathrm{min}} &= (d + \epsilon)\cos(\theta + \nu) \\
x_{\mathrm{max}} &= (d - \epsilon)\cos(\theta - \nu) + \ell.
\end{aligned}$$

Similarly, if $\theta < \frac{\pi}{2} - \nu$, then

$$\begin{aligned}
y_{\mathrm{min}} &= (d - \epsilon)\sin(\theta - \nu) \\
y_{\mathrm{max}} &= (d + \epsilon)\sin(\theta + \nu) \\
x_{\mathrm{min}} &= (d - \epsilon)\cos(\theta + \nu) \\
x_{\mathrm{max}} &= (d + \epsilon)\cos(\theta - \nu) + \ell,
\end{aligned}$$

and if $\frac{\pi}{2} - \nu < \theta < \frac{\pi}{2} - \nu$, then

$$\begin{aligned} y_{\min} &= (d - \epsilon) \min \{\sin (\theta - \nu), \sin (\theta + \nu)\} \\ y_{\max} &= (d + \epsilon) \\ x_{\min} &= (d - \epsilon) \cos (\theta + \nu) \\ x_{\max} &= (d + \epsilon) \cos (\theta - \nu) + \ell. \end{aligned}$$

Thus, for example, the range in y is given in these three cases by

$$\begin{aligned} \Delta y &= -2d \cos\theta \sin\nu + 2\epsilon \sin\theta \cos\nu && \text{Case 1} \\ &= 2d \cos\theta \sin\nu + 2\epsilon \sin\theta \cos\nu && \text{Case 2} \\ &= (d + \epsilon) - (d - \epsilon) \min \{\sin (\theta - \nu), \sin (\theta + \nu)\}. && \text{Case 3} \end{aligned}$$

We want to consider the maximum extent of this range. Applying standard minimization techniques, we find that the maximum value for both Case 1 and Case 2 are given by

$$\Delta y = 2\sqrt{d^2 \sin^2 \nu + \epsilon^2 \cos^2 \nu}$$

and for Case 3, the maximum occurs at

$$2\left(d \sin^2 \nu + \epsilon \cos^2 \nu\right).$$

Moreover, the extremum observed in Case 1 and Case 2 is always greater than that of Case 3, so we can bound the y component of the enscribing rectangle by

$$\Delta y = 2\sqrt{d^2 \sin^2 \nu + \epsilon^2 \cos^2 \nu}.$$

In a similar manner, we find that we can bound the x component of the enscribing rectangle by

$$\Delta x = \ell + 2\sqrt{d^2 \sin^2 \nu + \epsilon^2 \cos^2 \nu}.$$

This gives us a rectangular bound on the area of RC-space in which the second sensory point can lie, given that the first sensory point lies somewhere on the base face. Since a face is described in RC-space by the position and orientation of its midpoint, relative to the base face, we need to expand this volume to cover the possible positions of the midpoints of the face. Clearly the most straightforward way to do this is to expand this rectangle by $\frac{\ell}{2}$ on all sides, since a sensed point must lie within that distance of the midpoint of the face. This yields an area in RC-space of dimensions

$$2\ell + 2\sqrt{d^2 \sin^2 \nu + \epsilon^2 \cos^2 \nu}$$

by

$$\ell + 2\sqrt{d^2 \sin^2 \nu + \epsilon^2 \cos^2 \nu}$$

in which the midpoint of a face must lie in order to be consistent with the sensory information.

Finally, we must account for the orientation of the face. Clearly, the angular component of RC-space has extent 2π and the range of possible values for the orientation of the second face relative to the first is given by 2γ. Thus, the swept volume of RC-space corresponding to possible positions for the second sensory point is given by sweeping the area illustrated in Figure 5 through the appropriate range of values along the orientation dimension, as illustrated in Figure 7.

This defines a volume in RC-space in which a face must lie in order to be consistent with the sensory information. In order to make useful estimates concerning this volume, we need to normalize this volume. To do so, we have made the assumption that the faces

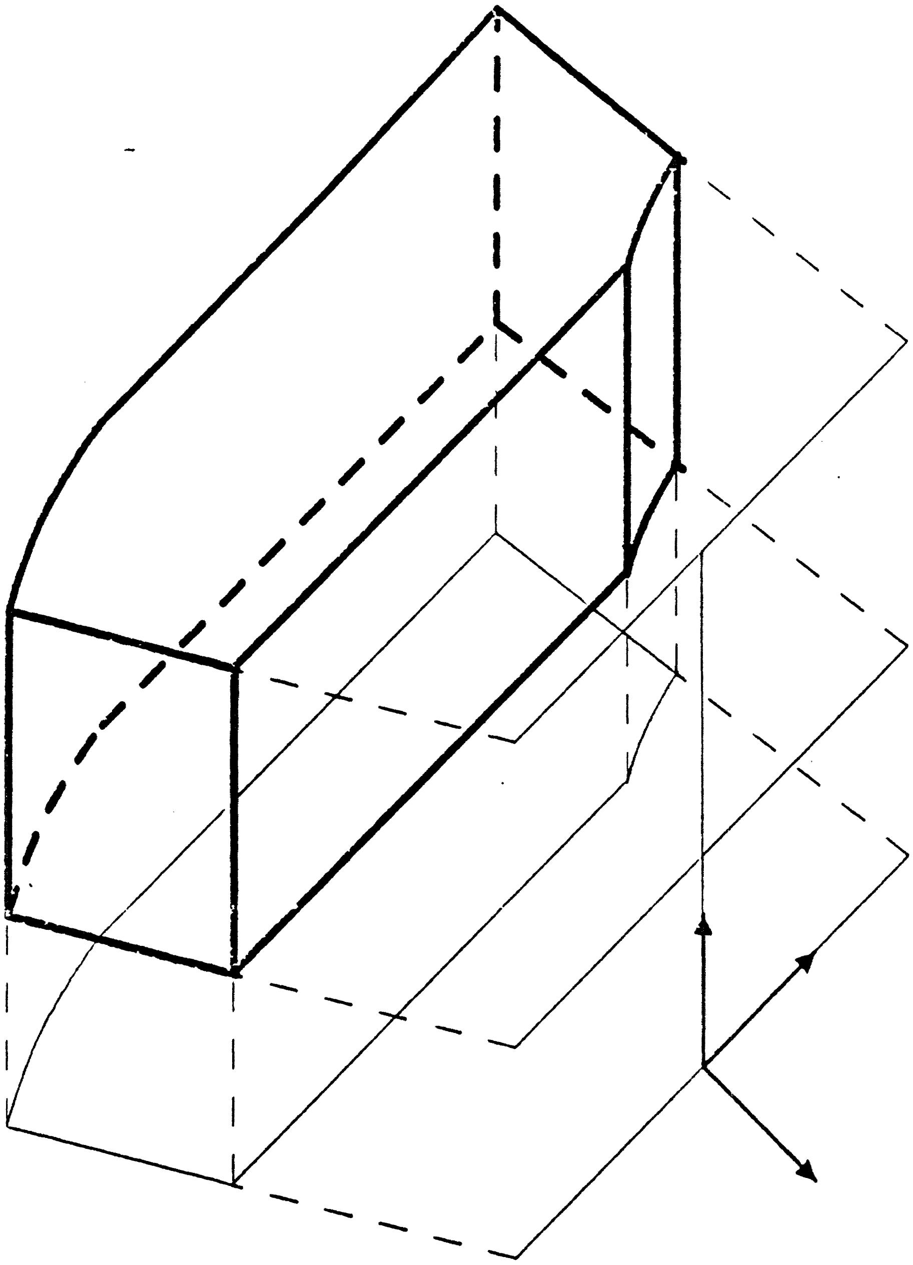

Figure 7. The range of possible positions for a second sensory point, given that the first sensory point lies anywhere on the base face.

have equal probability of lying anywhere within this RC-space (which spans an area πD^2 in the spatial dimensions and 2π in the angular dimension). Clearly this assumption is not necessarily valid for any particular object, although when averaged over all possible orientations of the object, it becomes more accurate. In this case, the normalized volume of RC-space in which a face must lie is bounded above by

$$V_2 = \left[\frac{v'_{\epsilon,\ell,\gamma}}{v_T}\right] \leq \frac{\gamma}{\pi} s_1 \, s_2$$

where

$$s_1 = \frac{1}{\sqrt{\pi}}\left[\frac{2\ell}{D} + 2\sqrt{\sin^2\nu - \left(\frac{\epsilon}{D}\right)^2\cos^2\nu}\right]$$

$$s_2 = \frac{1}{\sqrt{\pi}}\left[\frac{\ell}{D} + 2\sqrt{\sin^2\nu + \left(\frac{\epsilon}{D}\right)^2\cos^2\nu}\right].$$

Note that, as expected, V_2 is a dimensionless quantity, depending only of ratios of parameters of the polygonal object. Also, we can restrict our computation to cases where $d \geq 2\epsilon$ so that ν can be bounded above by $\gamma + \frac{\pi}{4}$.

Since V_2 is an upper bound on the probability of consistency p, by equation 1, this normalized volume leads to a straightforward bound on the expected number of interpretations consistent with s sensed points,

$$I_s = m^s p_s^{\binom{s}{2}} \tag{5}$$

where

$$p \leq p_s = \frac{\gamma}{\pi^2} s_1 s_2.$$

6.2 The Three Dimensional Case

We can expand the derivation of the previous section to deal with the three dimensional case. Here, the faces are squares of side ℓ, rather than one dimensional edges. It is straightforward to show that the positional aspects of the derivation can be decomposed into x and y components, both yielding normalized ranges of size s_1, while the z component has a normalized range given by s_2. Only one of the two rotational parameters can be constrained by the measurements between the two unit surface normals, so that the normalized volume in this case is given by

$$V_3 = \left[\frac{v'_{\epsilon,\ell,\gamma}}{v_T}\right] \leq \frac{\gamma}{\pi} s_1^2 s_2.$$

As in the two dimensional case, this is an upper bound on the probability p, and allows us to define an upper bound on the expected number of interpretations, by using equation 1.

6.3 Computable Bounds

To use these results for actual computations, we make two final assumptions. In particular, we assume that we can filter our sensory data so as to ensure that

$$d \geq 2\epsilon.$$

This is straightforward to do with any of the sensing modalities mentioned, and furthermore makes intuitive sense. Clearly, allowing closely spaced sensory points (where closeness here is defined by the sensitivity ϵ of the sensing device) is unlikely to provide much additional constraint on the recognition problem. This assumption then implies that

$$\gamma \leq \nu \leq \frac{\pi}{4} + \gamma.$$

We further assume that $\gamma \leq \pi/4$, since errors in sensing the surface orientation beyond that range imply that the sensor cannot tell which side of a plane the sensor lies on.

Given these simple assumptions on the quality of the sensory data, if we let $\ell' = \ell/D$ denote the normalized, dimensionless edge length, and $\epsilon' = \epsilon/D$ denote the normalized, dimensionless sensing error, then we have

$$\begin{aligned} s_2 &\leq \frac{1}{\sqrt{\pi}} \left[\ell' + \sqrt{2\left(\sin\gamma + \cos\gamma\right)^2 + 4\left(\epsilon'\right)^2 \cos^2\gamma} \right] \\ s_1 &= \frac{\ell'}{\sqrt{\pi}} + s_2 \\ V_2 &= \frac{\gamma}{\pi} s_1 s_2 \\ V_3 &= \frac{\gamma}{\pi} s_1^2 s_2 \end{aligned}$$

and the upper bounds on the probability of consistency, as in equation 1, are given by V_2 in the case of three degrees of freedom, and by V_3 in the case of six degrees of freedom.

6.4 Degradation with Noise

Having derived specific bounds on the effective pruning of the local constraints, it is easy to see how the presence of sensor error affects the expected number of consistent hypotheses. To demonstrate the graceful degradation of the constraints, we have plotted the expected number of interpretations as a function of the number of sensory points, for several different error conditions.

In Figure 8, we graph an upper bound on the expected number of interpretations as a function of the sensor error, by using equation 5. Each graph represents a different value of ϵ' ranging from 0.05 to 0.5. The ordinate of each graph is the expected number of interpretations. The abscissa is the number of sensory data points. For all of the graphs shown here, the number of model faces was fixed at $m = 1000$ and the size of the model faces was fixed at $\ell' = 0.01$. Parts a, b and c show the cases of angle error $\gamma = \pi/5, \pi/10$ and $\pi/15$ respectively. In Figure 9, we plot similar graphs for the case of $\ell' = 0.1$.

It can be seen from these graphs, as might be expected from the earlier theoretical analysis, that the number of expected interpretations reaches an extremum for a small number of sensed points, and then quickly reduces as the number of sensed points is increased. This suggests that for a wide range of errors in the sensory measurements the constraint based recognition and localization algorithm will be remarkably effective.

6.5 Predictions for Sensing

Given estimates for the probability of consistency between faces of an object, we can use equation 3 to predict the number of sensory data points needed to reduce the expected number of interpretations to 1. Some samples are listed in Tables 1 and 2.

It can be seen from Tables 1 and 2 that the number of sensory data points required is quite small, even for large amounts of error in the sensory measurements.

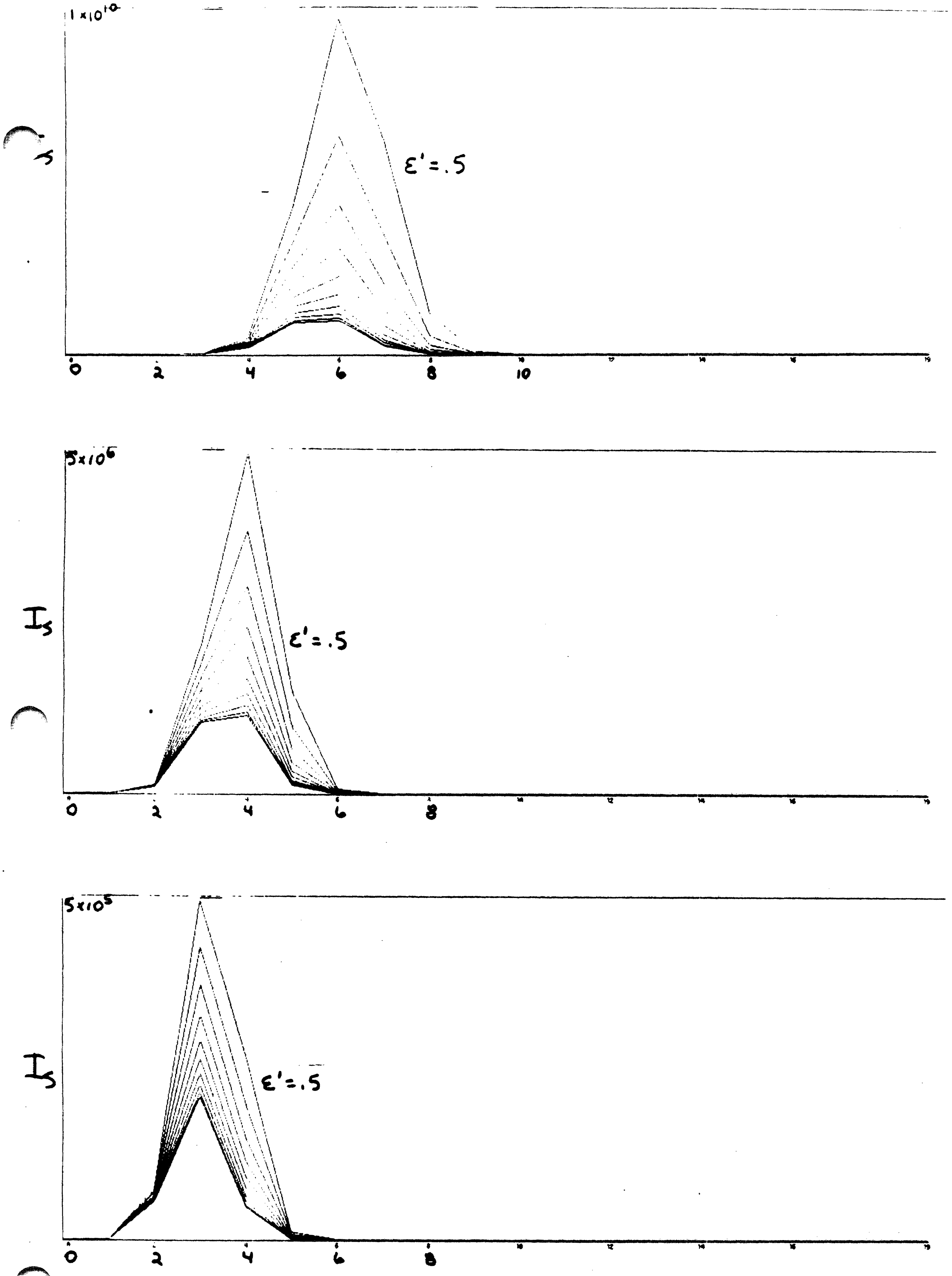

Figure 8. The expected number of interpretations as a function of the sensor error. Each graph represents a different value of ϵ' ranging from 0.05 to 0.5. The ordinate of each graph is an upper bound on the expected number of interpretations. The abscissa is the number of sensory data points. For all of the graphs shown here, the number of model faces was fixed at $m = 1000$ and the size of the model faces was fixed at $\ell' = 0.01$. Parts a, b and c show the cases of angle error $\gamma = \pi/5, \pi/10$ and $\pi/15$ respectively.

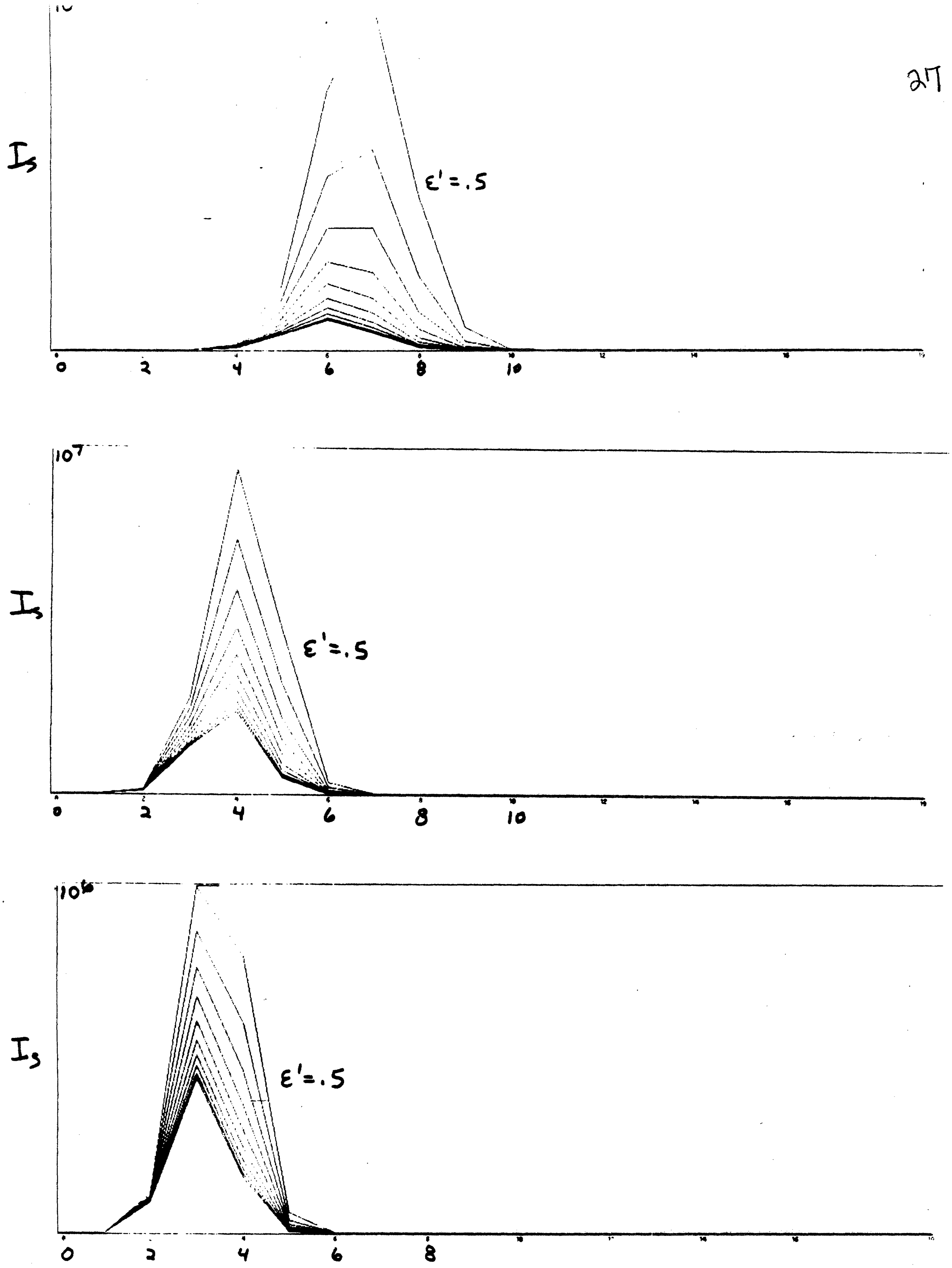

Figure 9. The expected number of interpretations as a function of the sensor error. Each graph represents a different value of ϵ' ranging from 0.05 to 0.5. The ordinate of each graph is an upper bound on the expected number of interpretations. The abscissa is the number of sensory data points. For all of the graphs shown here, the number of model faces was fixed at $m = 1000$ and the size of the model faces was fixed at $\ell' = 0.1$. Parts a, b and c show the cases of angle error $\gamma = \pi/5, \pi/10$ and $\pi/15$ respectively.

	.05	.1	.15	.2	.25	.3	.35	.4	.45	.5
$\pi/50$	4	4	4	4	4	4	4	4	4	4
$2\pi/50$	5	5	5	5	5	5	5	5	5	5
$3\pi/50$	5	5	6	6	6	6	6	6	6	6
$4\pi/50$	6	6	6	6	6	6	7	7	7	7
$5\pi/50$	7	7	7	7	7	7	7	7	8	8
$6\pi/50$	8	8	8	8	8	8	8	8	9	9
$7\pi/50$	9	9	9	9	9	9	9	9	10	10
$8\pi/50$	10	10	10	10	10	10	10	10	11	11
$9\pi/50$	11	11	11	11	11	11	11	11	12	12
$10\pi/50$	12	12	12	12	12	12	12	12	13	13

Table 1 – Number of sensory points needed for a unique interpretation. The number of model faces is fixed at 1000, and the ratio of model face size to object diameter is fixed at $\ell' = 0.1$. The rows show increasing positional error, measured in ϵ' and the columns show increasing angular error, measured in γ.

	.05	.1	.15	.2	.25	.3	.35	.4	.45	.5
$\pi/50$	5	5	5	5	5	5	5	5	5	5
$2\pi/50$	6	6	6	6	6	6	6	6	6	6
$3\pi/50$	7	7	7	7	7	7	7	7	7	7
$4\pi/50$	8	8	8	8	8	8	8	8	9	9
$5\pi/50$	9	9	9	9	9	10	10	10	10	10
$6\pi/50$	11	11	11	11	11	11	11	11	12	12
$7\pi/50$	12	12	12	12	13	13	13	13	14	14
$8\pi/50$	14	14	14	14	15	15	15	15	16	16
$9\pi/50$	16	16	16	17	17	17	17	18	18	19
$10\pi/50$	19	19	19	19	19	20	20	21	21	22

Table 2 – Number of sensory points needed for a unique interpretation. The number of model faces is fixed at 1000, and the ratio of model face size to object diameter is fixed at $\ell' = 0.5$. The rows show increasing positional error, measured in ϵ' and the columns show increasing angular error, measured in γ.

7. Application of the Theory

We began our investigation of the problem of model-based recognition and localization by establishing a set of criteria that should be satisfied by constraints between sensory data and model elements. In particular, we argued that the constraints should be coordinate-frame-independent, simple, sensor-independent, combinatorially powerful and degrade gracefully with error. We have spent the bulk of this paper deriving such a set of constraints and establishing on theoretical grounds that these criteria are satisfied. This theoretical basis was an outgrowth of earlier work [Grimson and Lozano-Pérez 1984] in which an isomorphic set of constraints was proposed and tested.

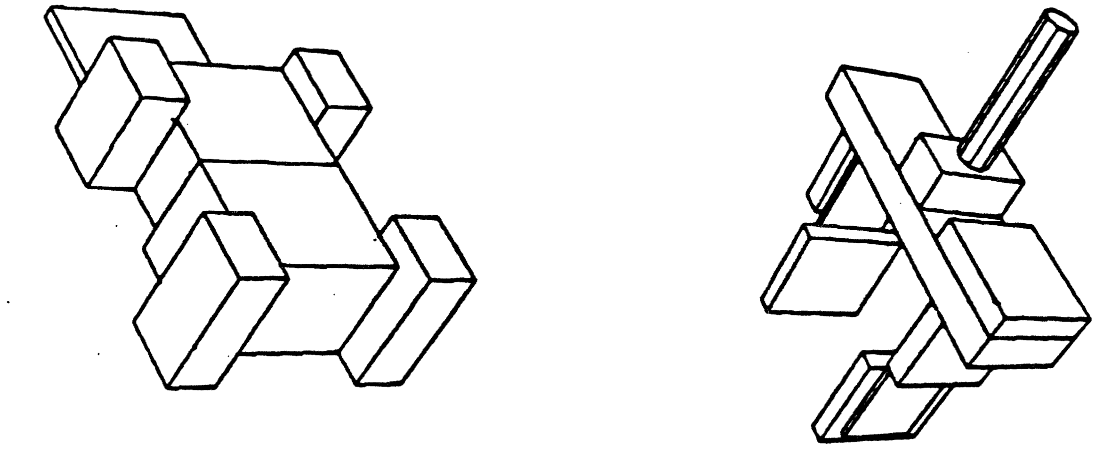

Figure 10. Sample objects used in empirical studies as reported in [Grimson and Lozano-Pérez 84]. Table 3 indicates the agreement between the theoretical results and the empirically observed behavior.

The original motivation for this theoretical investigation was the surprising (to the authors) success of a recognition algorithm based on the simple geometric constraints developed here, and reported in [Grimson and Lozano-Pérez 84]. In the original development and presentation of the recognition algorithm, a large set of simulations were run on a series of test objects, for varying types of error conditions. Even though the assumptions used in development above are not completely met by the objects used in the empirical studies, we can use the empirical results to demonstrate a general agreement between the theoretical bounds derived in this paper, and the empirically observed results. Two sample objects are shown in Figure 10. A comparison of the theoretical and empirical results is shown in Table 3. Here I_{exp} is computed from equations 1 and 3, and is the theoretically prediction of the number of interpretations. I_{obs} is the empirically observed number of interpretations. s_{min} is computed using equation 3 and denotes the predicted number of data points needed to force a unique interpretation. Since the object does not consist of equal sized faces, we used the average length of the edge of a face as the value for ℓ in our computations. The table indicates that the values in the I_{exp} and the I_{obs} are almost identical, and the slight deviations can probably be accounted for by deviations of the objects from the simplifying assumptions made in the theoretical analysis. Similar agreement has been observed for other objects.

We can relate the theoretical results to empirical ones in another way. Equation 2 predicts the number of sensory points for which the expected number of interpretations reaches a maximum, provided that an estimate for p is available. In [Grimson and

γ	ϵ'	$V_3 = p$	s_{min}	I_{exp}	I_{obs}
$\pi/15$	.003	.0970	5	1	1
$\pi/15$	.017	.0971	5	1	1
$\pi/15$	.033	.0972	5	1	1
$\pi/10$	.003	.1695	6	1	1
$\pi/10$	.017	.1696	6	1	1
$\pi/10$	.033	.1697	6	1	2
$\pi/8$	.003	.2322	7	1	1
$\pi/8$	.017	.2323	7	1	1
$\pi/8$	.033	.2326	7	1	2

γ	ϵ'	$V_3 = p$	s_{min}	I_{exp}	I_{obs}
$\pi/12$	.0013	.1023	5	1	1
$\pi/12$	.0065	.1024	5	1	1
$\pi/10$	.0013	.1328	6	1	1
$\pi/10$	.0065	.1328	6	1	1
$\pi/8$	.0013	.1833	6	1	2
$\pi/8$	.0065	.1834	6	1	2

Table 3 – Comparison of theoretical and empirical results. The top table corresponds to the Motor Housing, and the bottom table to the Hand. γ and ϵ' indicate the normalized errors in the sensing. s_{min} is the minimum number of sensory points needed to reduce the expected number of interpretations I_{exp} to 1. I_{obs} is the median number of observed interpretations. using 12 points of sensed data, and based on 100 trials.

Lozano-Pérez 84] the median number of interpretations, based on a series of tests, was presented as a function of the number of sensory points. For the objects in Figure 10, it was shown that the maximum number of interpretations occurred for 3 sensory points, after which point the number of interpretations decreased. If we evaluate equation 2, given the estimate of p obtained above, we find that the theoretical bounds also predict a peak in the number of interpretations for 3 sensory points (when taken to the nearest integer).

8. Summary

Previously, [Grimson and Lozano-Pérez 84, 85a, 85b] presented a technique for the recognition and localization of objects from sparse, noisy sensory data. The technique's performance on extensive simulations and real data was used as support for its effectiveness. In this paper, we have further supported that recognition technique, by showing that local, coordinate-frame-independent constraints between sparse, noisy sensory data are very effective, in general, in handling the recognition and localization problem.

We provided a method for estimating the probability of consistency between data points and corresponding object patches, and used this method to actually compute the expected number of interpretations, under a variety of sensing conditions. It was seen that this number rapidly decreases with increasing sensory data, and that shows graceful degradation with the presence of increasing noise. The predictions of the theory were also shown to be in agreement with the results of experiments.

Acknowledgments

Tomás Lozano-Pérez was critical to the development of this work, both in terms of the underlying recognition technique and in terms of this particular presentation. Matt Mason scrutinized versions of the manuscript with extreme care, and provided many valuable suggestions and criticisms. John Hollerbach and Berthold Horn provided useful comments at various stages. The referees greatly improved the presentation of the original manuscript by providing valuable criticisms and suggestions.

9. References

Ballard, D. H. 1981. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition* 13(2):111–122.

Bausch and Lomb. 1976. Bausch and Lomb Omnicon Pattern Analysis System. Analytic Systems Division brochure. Rochester, New York: Bausch and Lomb.

Bolles, R. C., and Cain, R. A. 1982. Recognizing and locating partially visible objects: The Local-Feature-Focus method. *Int. J. Robotics Res.* 1(3):57–82.

Bolles, R. C., Horaud, P., and Hannah, M. J. 1984. 3DPO: A three-dimensional part orientation system. *Robotics Research*, ed. Michael Brady and Richard Paul. Cambridge, Mass.: MIT Press, pp. 413–424.

Brady, M. 1982. Smoothed local symmetries and frame propagation. *Proc. IEEE Pattern Recog. and Im. Proc.*.

Brooks, R. 1981. Symbolic reasoning among 3-dimensional models and 2-dimensional images. *Artificial Intell.* 17:285–349.

Brou, P. 1983. Finding objects in depth maps. Ph.D. thesis, Massachusetts Institute of Technology Department of Electrical Engineering and Computer Science.

Faugeras, O. D., and Hebert, M. 1983 (Aug. Karlsruhe, W. Germany). A 3-D recognition and positioning algorithm using geometrical matching between primitive surfaces. *Proc. Eighth Int. Joint Conf. Artificial Intell.* Los Altos: William Kaufmann, pp. 996–1002.

Gaschnig, J. 1979. Performance measurement and analysis of certain search algorithms, Ph. D. Thesis, Dept. Computer Science, Carnegie-Mellon University.

Gaston, P. C., and Lozano-Pérez, T. 1984. Tactile recognition and localization using object models: The case of polyhedra on a plane. *IEEE Trans. Pattern Anal. Machine Intell.* PAMI-6(3):257–265.

Gleason, G., and Agin, G. J. 1979 (Mar). A modular vision system for sensor-controlled manipulation and inspection. *Proc. Ninth Int. Symp. Industrial Robots*. Dearborn, Mich.: Society of Manufacturing Engineers, pp. 57–70.

Grimson, W. E. L. 1984. The combinatorics of local constraints in model-based recognition and localization from sparse data. AIM-763. Cambridge, Mass.:Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Grimson, W. E. L., and Lozano-Pérez, T. 1984. Model-based recognition and localization from sparse range or tactile data. *Int. J. Robotics Res.* 3(3):3–35.

Grimson, W. E. L., and Lozano-Pérez, T. 1985a. Recognition and localization of overlapping parts from sparse data. AIM-841. Cambridge, Mass.:Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Grimson, W. E. L., and Lozano-Pérez, T. 1985b (Mar., St. Louis, MO). Recognition and localization of overlapping parts from sparse data in two and three dimensions. *Proc. IEEE Intern. Conf. on Robotics and Automation.* Silver Spring: IEEE Computer Society Press, pp. 61–66.

Haralick, R. M. and Elliot, G. L. 1980. Increasing tree search efficiency for constraint satisfaction problems. *Artificial Intelligence* 14:263–313.

Holland, S. W. 1976 (Feb.) A programmable computer vision system based on spatial relationships. General Motors Publ. GMR-2078. Detroit: General Motors.

Horn, B. K. P., and Ikeuchi, K. 1983. Picking parts out of a bin. AIM-746. Cambridge, Mass.:Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Horn, B. K. P. 1983. Extended Gaussian images. AIM-740. Cambridge, Mass.:Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Ikeuchi, K. 1983. Determining attitude of object from needle map using extended gaussian image. AIM-714. Cambridge, Mass.:Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Marr, D. 1982. *Vision*. San Francisco:W. H. Freeman and Company.

Marr, D., and Nishihara, H. K. 1978. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. R. Soc. Lond. B* 200:269–294.

Machine Intelligence Corporation. 1980. Model VX-100 machine vision system. Product description. Mountain View, Calif.: Machine Intelligence Corporation.

Nevatia, R. 1974. Structured descriptors of complex curved objects for recognition and visual memory. Ph.D. thesis, Stanford University. AIM 250. Stanford, Calif.: Stanford University Artificial Intelligence Laboratory.

Nevatia, R., and Binford, T. O. 1977. Description and recognition of curved objects. *Artificial Intell.* 8:77–98.

Nudel, B. 1983. Consistent-Labeling problems and their algorithms: Expected-complexities and theory-based heuristics. *Artificial Intelligence* 21:135–178.

Perkins, W. A. 1978. A model-based vision system for industrial parts. *IEEE Trans. Comput.* C-27:126–143.

Reinhold, A. G., and VanderBrug, G. J. 1980. Robot vision for industry; the autovision system. *Robotics Age*, Fall, pp. 22–28.

Stockman, G., and Esteva, J. C. 1984. Use of geometrical constraints and clustering to determine 3D object pose. TR84-002. East Lansing, Mich.:Michigan State University Department of Computer Science.

Sugihara, K. 1979. Range-data analysis guided by a junction dictionary. *Artificial Intell.* 12:41–69.

Tsuji, S., and Nakamura, A. 1975 (Aug., Cambridge, Mass.). Recognition of an object in a stack of industrial parts. *Proc. Fourth Int. Joint Conf. Artificial Intell.* Los Altos: William Kaufmann, pp. 811–818.