MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. Memo No. 1019 February, 1988

The Combinatorics of Object Recognition in Cluttered Environments using Constrained Search

W. Eric L. Grimson

Abstract. The problem of recognizing rigid objects from noisy sensory data has been successfully attacked in previous work by using a constrained search approach. Empirical investigations have shown the method to be very effective when recognizing and localizing isolated objects, but less effective when dealing with occluded objects where much of the sensory data arises from objects other than the one of interest. When clustering techniques such as the Hough transform are used to isolate likely subspaces of the search space, empiricial performance in cluttered scenes improves considerably. In this note, we establish formal bounds on the combinatorics of this approach. Under some simple assumptions, we show that the expected complexity of recognizing isolated objects is quadratic in the number of model and sensory fragments, but that the expected complexity of recognizing objects in cluttered environments is exponential in the size of the correct interpretation. We also provide formal bounds on the efficacy of using the Hough transform to preselect likely subspaces, showing that problem remains exponential, but that in practical terms, the size of the problem is significantly decreased.

Acknowledgements: 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 the System Development Foundation, in part by an Office of Naval Research University Research Initiative grant under contract N00014-86-K-0685, and in part by the Advanced Research Projects Agency of the Department of Defense under Army contract number DACA76-85-C-0010 and under Office of Naval Research contract N00014-85-K-0124.

1. Object Recognition

Recognizing and locating objects from sensory data is a common element of many of the tasks that an intelligent system must perform. Variations of the problem arise in tasks ranging from visual inspection to hand-eye coordination to autonomous vehicle localization. In all of these domains, the recognition problem can be generally characterized as follows: Given a set of object models, and given sensory data about some environment, find all the instances of the models in the environment, both identifying the existence of an instance, and identifying the location of that instance. Each solution to the recognition problem usually consists of a specification of which subset of the sensory data accounts for the object instance and the transformation needed to map the object model from its own inherent coordinate frame into the sensor's coordinate frame, in order to account for the sensory data.

Clearly, the information contained in the sensory data can significantly influence possible approaches to the problem. In general, the data may come from any of a number of modalities, including visual, range and tactile data, and that data is generally noisy, partially occluded and partially spurious. Although other approaches are possible, we shall restrict our attention to the case in which the sensory data, from any of these modalities, can be processed to derive measurements about the geometry of local portions of the object's boundary. In order to be robust, a recognition system must be able to deal with measurements of the position and orientation of a patch of surface that are noisy. As well, the data may come from environments in which much of the data is spurious, arising from objects other than the one of interest, and in which much of the object of interest is occluded, so that sensory data is available only for some portions of the object.

The problem of recognizing rigid objects from noisy sensory data has been successfully attacked in previous work by using a constrained search approach [Grimson and Lozano-Pérez 84, 87]. Empirical investigations have shown the method to be very effective when recognizing and localizing isolated objects, but less effective when dealing with occluded objects where much of the sensory data arises from objects other than the one of interest. When clustering techniques such as the Hough transform are used to isolate likely subspaces of the search space, empiricial performance in cluttered scenes improves considerably. In this note, we establish formal bounds on the combinatorics of this approach. Under some simple assumptions, we show that the expected complexity of recognizing isolated objects is quadratic in the number of model and sensory fragments, but that the expected complexity of recognizing objects in cluttered environments is exponential in the size of the correct interpretation. We also provide formal bounds on the efficacy of using the Hough transform to preselect likely subspaces, showing that problem remains exponential, but that in practical terms, the size of the problem is significantly decreased.

In the remainder of this section, we briefly describe the constrained search method used to solve the recognition problem. In section 2, we consider the com-

binatorics of unoccluded objects, obtaining general expressions for the expected search. These results are extended to occluded objects in section 3. Specific bounds relating the combinatorics to the object recognition problem are derived in sections 4 and 5. The impact of Hough transforms on the problem are considered in section 6.

1.1 Definition of a solution

In more formal terms, a solution to the recognition problem consists of a triplet,

$$\langle \texttt{object}_i, \{(d_{i_1}, m_{j_1}), (d_{i_2}, m_{j_2}), \ldots (d_{i_k}, m_{j_k})\}, \mathcal{T} \rangle$$

where $\texttt{object}_i$ identifies which object from a library of known objects, the (d,m) pairings are associations of a subset of the sensory data, d, with model elements, m, from $\texttt{object}_i$ and $\mathcal{T}$ is a transformation from model coordinates to sensor coordinates such that each data fragment agrees with its transformed model element, to within noise bounds.

Stated in such general terms, there are a variety of possibilities for specifying the recognition problem, with variations in the types of models, the specifics of the sensor data, and the method used to find the transformation. In this article, we will restrict attention to the following specific case.

- We will assume that the objects are modeled as polygons or polyhedra, so that each m_j is a linear segment. The models need not be complete, so that gaps are allowed.
- We will also assume that the sensory data, d_i, can be processed to produce estimates of the geometry of linear fragments of the object's boundary, either line segments in the case of two-dimensional data, or planar patches in the case of three-dimensional data.
- We will assume that the objects are rigid, so that the transformation $\mathcal{T}$ maps points $\mathbf{v}_m$ in model coordinates into points $\mathbf{v}_s$ in sensor coordinates by

  $$\mathbf{v}_d = R\mathbf{v}_m + \mathbf{v}_0$$

  where R is a rotation matrix, and $\mathbf{v}_0$ is a translation vector.

Even in this case,there are a variety of techniques for finding the solution, most of which can be considered as different forms of search. Successful approaches have included maximal clique techniques [e.g. Bolles and Cain 82, Bolles et al. 84], hypothesize and test methods [e.g. Ayache and Faugeras 86] and constrained search [e.g. Grimson and Lozano-Pérez 84, 87]. In this article, we are interested in the constrained search approach.

1.2 Constrained search as applied to recognition

The basic idea is to find legitimate pairings of data and model fragments by a depth first search of an *interpretation tree* (IT). We begin by associating the first

data fragment with the first model face, and represent this by a node at the first level of the tree. If this association is feasible, we consider associating the second data fragment with the first model face, represented as a node at the second level of the tree, which is a son of the first node. If this pair of associations is still feasible, we continue downward in the tree, associating model faces with the next data fragment. If this pair of associations is not feasible, then we backtrack, and consider associating the data fragment with the second model face, and so on. Once we have considered the association of a data fragment with all of the m model faces, we also consider excluding the data fragment from the interpretation, by associating it with the *wild card* $(*)$ or *null* branch. Thus, each node of the tree describes a partial interpretation of the data, and implicitly contains a set of pairings of data fragments and model faces. Nodes at the i^{th} level of the tree define assignments for the first i data fragments. Each node branches at the next level in up to $m+1$ ways, where m is the number of model faces in the object. The last branch is a *wild card* or *null* branch and has the effect of excluding the data fragment corresponding to the current level of the tree from the interpretation defined at that node. An example is shown in Figure 1. With the inclusion of the wild card branch, any node at level i defines a mapping from a subset of the first i data points to actual faces of the object model.

Given s data fragments, any leaf of the tree specifies an interpretation

$$\{(d_1, m_{j_1}), (d_2, m_{j_2}), \ldots (d_s, m_{j_s})\},$$

where some of the m_{j_h} may be the wild card character. By excluding such matches, the leaf yields a partial interpretation

$$\left\{(d_{i_1}, m_{j_{i_1}}), (d_{i_2}, m_{j_{i_2}}), \ldots (d_{i_k}, m_{j_{i_k}})\right\}$$

where $1 \leq i_1 < i_2 < \ldots < i_k$, but these indices may not include the entire set from 1 to s. This interpretation may then be used to solve for a rigid, scaled transformation that maps model faces into corresponding data fragments, if such a transformation exists. This transformation must map the faces so that both the position and the orientation of the face are consistent with the associated data point, modulo noise in the measurements. Thus, by searching for leaves of the tree and testing that the interpretation there yields a legal transformation, we can find possible instances of object models in the data, and solve the recognition problem.

Because this search process is inherently an exponential problem, the key to an efficient solution is to use constraints to remove large subtrees from consideration without explicitly having to explore them, thereby providing a specific definition for the notion of *feasible* in the above discussion. In [Grimson and Lozano-Pérez 84, 87] we describe a constrained search method called RAF (for Recognition and Attitude Finder), that uses a set of constraints based on the relative shape of parts of objects, either in two dimensions or in three. In this work, the object models and the sensory data consist of linear edge or face fragments. The constraints include the following:

- The length (area) of a data fragment must be smaller than the length (area) of a corresponding model fragment, up to some bounded measurement error;

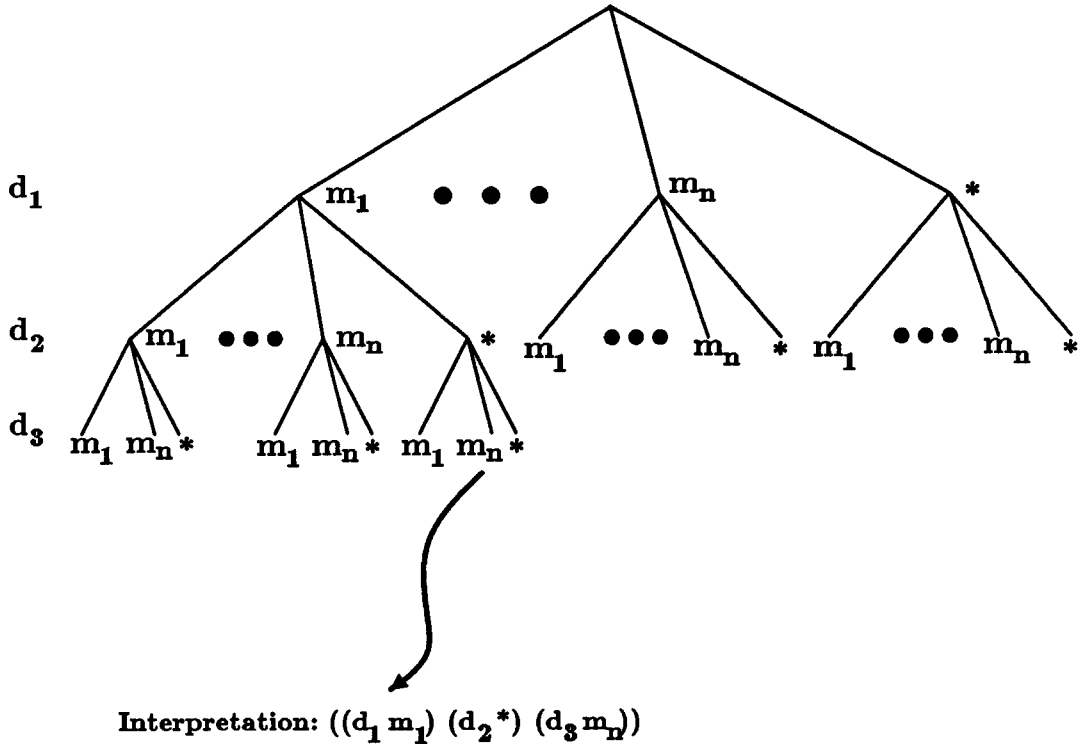

Figure 1. An Interpretation Tree. Each node of the tree defines a partial interpretation, where the level of each ancestor defines a sensory data point, and the branch leading to each such node defines the corresponding model fragment. An example of a partial interpretation is shown, where d_i denotes the i^{th} data point and m_k denotes the k^{th} model fragment. The * indicates the wild card branch, corresponding to the exclusion of the associated data point from the interpretation.

- The angle between the normals to a pair of data fragments must differ from the angle between the normals of the corresponding model fragments by no more than a bounded measurement error;
- The range of distances between two data fragments must lie within the range of distances of the corresponding model fragments, where the model range has been expanded to account for measurement errors;
- The range of components of a vector spanning the two data fragments in the direction of each of the fragments' normal must lie within the corresponding range of components for vectors spanning the model fragments, modulo measurement error.
- A data fragment assigned to the wild card is always consistent.

It is possible to extend these constraints to handle the recognition of curved objects in two dimensions [Grimson 87], but here we stay with linear elements.

1.3 The constraints reduce the search

Given these unary and binary constraints, the constrained search process consists of a depth first search, with downward termination based on constraint consistency. Suppose the search process is currently at some node at level k in the interpretation tree and with a **consistent** partial interpretation given by

$$I_k = \{(d_1, m_{j_1}), (d_2, m_{j_2}), \ldots (d_k, m_{j_k})\}.$$

We now consider the next data fragment d_{k+1}, and its possible assignment to model face $m_{j_{k+1}}$, where j_{k+1} varies from 1 to $m+1$. This leads to a potential new interpretation

$$I_{k+1} = \{(d_1, m_{j_1}), (d_2, m_{j_2}), \ldots (d_{k+1}, m_{j_{k+1}})\}$$

The following rules hold.

- If $m_{j_{k+1}}$ is the wild card match, then the new interpretation I_{k+1} is consistent, and we continue downward in our search.
- If $m_{j_{k+1}}$ is a real model edge fragment, we must verify that the length constraint holds for matching d_{k+1} to $m_{j_{k+1}}$, and that the angle, distance and component constraints hold for the pairings $[(d_{k+1}, m_{j_{k+1}}), (d_i, m_{j_i})]$, for $1 \leq i \leq k$.
- If all of these constraints are true, then the new interpretation I_{k+1} is a consistent partial interpretation, and we continue our depth first search. If one of them is false, then the partial interpretation is inconsistent. In this case, we increment the model face index j_{k+1} by 1 and try again with a new I_{k+1}, until $j_{k+1} = m+1$.

If the search process is currently at some node at level k in the interpretation tree, and has an **inconsistent** partial interpretation given by

$$I_k = \{(d_1, m_{j_1}), (d_2, m_{j_2}), \ldots (d_k, m_{j_k})\}$$

then it is in the process of backtracking. If $j_k = m+1$ (the wild card) we backtrack up another level, otherwise we increment j_k and continue.

1.4 Model tests

Once the search process reaches a leaf of the interpretation tree, we have accounted for all of the data points. We are now ready to determine if the interpretation is in fact globally valid. To do this, we solve for a rigid transformation mapping points $\mathbf{v}_m$ in model coordinates into points $\mathbf{v}_d$ in sensor coordinates,

$$\mathbf{v}_d = R\mathbf{v}_m + \mathbf{v}_0$$

where R is a rotation matrix, and $\mathbf{v}_0$ is a translation vector. We can solve for this transformation in a number of ways [e.g. Grimson and Lozano-Pérez 84, 87, Ayache and Faugeras 86].

Given such a transformation, which is usually found using some type of least squares fit, we must then ensure that the interpretation actually satisfies it. We do

this by considering each of the data fragments associated with a real model fragment in the interpretation, and transforming the associated model fragment by the computed transform. For each such fragment, we then verify that the transformed fragment differs in position and orientation from its associated data fragment by amounts that are less than some acceptable error bounds. These bounds on transform error can be obtained from the predefined bounds on the sensor error [Grimson 86b]. Any interpretation that passes such a model test is a consistent interpretation of the data.

1.5 Additional search reductions

While the constrained search technique described above will succeed in finding all consistent interpretations of the sensory data, for a given object model, it is not particularly efficient. This is mostly due to the problem of segmenting the data to determine subsets that belong to a single object. Indeed, if all of the sensory data do belong to one object, the described method is known to be quite efficient, as has been verified both empirically [Grimson and Lozano-Pérez 84, 87] and theoretically [Grimson 86a]. In order to improve the efficiency of the method, we can add two additional methods to our search process, both previously discussed for the case of linear fragments in [Grimson and Lozano-Pérez 87], and extended to circular fragments in [Grimson 87].

The first is to use a parameter hashing scheme, such as a Hough transform, to hypothesize small subspaces of the entire search space that are likely to contain an interpretation (a more detailed treatment of the Hough transform appears in a later section). The second is to use a measure of goodness of match, such as the portion of the object perimeter (in 2D) or the object surface area (in 3D) correctly accounted for by the matched sensory data, to prematurely terminate the search process. That is, as soon as an interpretation is found whose value under that measure exceeds a predefined threshold, the search process is terminated, with that interpretation taken as the correct solution. Both of these methods are known empirically to considerably reduce the search needed.

1.6 Empirical Performance

The recognition method described in the previous sections has been tested on a variety of data, including two dimensional recognition from grey-level images [Grimson and Lozano-Pérez 84, 87], and three-dimensional recognition from laser range data [Grimson and Lozano-Pérez 87], silhouettes [Van Hove 87], stereo data [Porrill, et al. 87], motion data [Murray 87] and tactile data [Grimson and Lozano-Pérez 84]. In all of these cases, the method typically finds a unique interpretation quite rapidly, in the presence of varying amounts of sensor noise.

For example, in [Grimson and Lozano-Pérez 87], we report on experiments in which an object containing 50 model edges was correctly identified in scenes contain-

ing 100 data edges, when as little as .25 percent of the object was visible in the scene. Over 100 different trials, the median search effort involved the exploration of 59000 nodes of the interpretation tree when using a Hough transform. In elapsed time, such exploration typically took only a few seconds on a Symbolics Lisp Machine.

2. The Combinatorics of Isolated Objects

Given that the RAF recognition technique has good empirical performance, our goal is to prove that such empirical observations are generally valid. We begin by considering the combinatorics of recognizing isolated objects, that is, situations in which all of the sensory data is known to lie on a single object. An earlier study of this problem is presented in [Grimson 86a]. In that study, we used a simple model of the recognition system to develop estimates of the performance of the system. In this section, we use a more complete model to develop better bounds on the performance of the system.

Because we formulate it as a search process, our approach to object recognition can be considered as a problem of constraint satisfaction, or consistent labeling. There are several general results available concerning the characteristics of consistent labeling techniques [e.g. Freuder 78, 82, Gaschnig 79, Haralick and Elliot 80, Haralick and Shapiro 79, Mackworth 77, Mackworth and Freuder 85, Montanari 74, Nudel 83, Waltz 75]. 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. We will use a specific instance of the framework provided by these results to derive explicit bounds on our version of the recognition problem.

2.1 Model of consistency – unoccluded case

We are particularly interested in bounds on the number of interpretations delivered by the system, and in bounds on the amount of work performed by the system, in this case measured as the number of nodes of the search tree actually explored by the system. Since our method uses both unary and binary constraints, we need to model the probability that a data-model assignment is consistent and the probability that a pair of data-model assignments are consistent.

We let $q_{i,I}$ denote the probability that assigning the i^{th} data element to the I^{th} model element is consistent, and we let $q_{i,j;I,J}$ denote the probability that the pair of assignments $i \mapsto I, j \mapsto J$ is consistent. Our model of the recognition problem is defined as follows.

For a single data-model pairing, if the pairing is part of the correct interpretation, the probability of consistency is simply 1. If it is not correct, we let the

probability of consistency be p_1. Thus, we have

$$q_{i,I} = \begin{cases} 1 & \text{if } i \mapsto I \text{ is correct} \\ p_1 & \text{otherwise.} \end{cases}$$

For a pair of assignments, suppose we are considering a match in which data fragments i, j are paired with model fragments I, J respectively. We will model the situation by saying that the consistency of this pair of pairs has probability 1 if these pairings are part of the correct interpretation, and has probability p_2 otherwise. Note that this is essentially assuming a random distribution of edges. It is also assuming that pairs of model edges are distinctive, so that objects with partial symmetries are excluded. Thus, we have

$$q_{i,j;I,J} = \begin{cases} 1 & \text{if } i \mapsto I, j \mapsto J \text{ is correct} \\ p_2 & \text{otherwise.} \end{cases}$$

Because the data is known to lie on a single object, we do not need to use the wild card branch of the search tree, so that each node of the search tree has only m branches in this case. Thus, the search tree has m^k nodes at level k. However, not all of these are actually reached by the algorithm.

In general, a node at the k^{th} level of the tree, with assignment $1 \mapsto I_1, \ldots, k \mapsto I_k$ has a probability of consistency:

$$\prod_{i=1}^{k} q_{i,I_i} \prod_{i=1}^{k} \prod_{j=i+1}^{k} q_{i,j;I_i,I_j}.$$

2.2 Simple bounds on the problem

We let n_k denote the number of consistent nodes at the k^{th} level of the interpretation tree, under this model of consistency. The expected number of consistent nodes is simply the sum of the probability above take over all mappings. We are interested in bounds on n_s, the number of interpretations of the s sensory data fragments. Simple bounds on the number of interpretations are given by the following result. In the interests of clarity of presentation, the proof is deferred to the appendix.

Proposition 1: If all of the k sensory measurements are known to lie on a single object with m faces, then the number of interpretations n_k is bounded by

$$n_k \leq \left[1 + (m-1)p_1 p_2^{\frac{k-1}{2}}\right]^k.$$

and by

$$n_k \geq 1 + \left[p_2^{\frac{1}{2}} + p_1(m-1)\right]^k p_2^{\frac{k(k-1)}{2}} - p_2^{\frac{k^2}{2}}$$

where p_1 is the probability of a random data-model assignment satisfying unary consistency, and p_2 is the probability of a pair of random data-model assignments satisfying binary consistency.∎

This provides us with formal bounds on the number of k-interpretations. Bounds on the number of nodes explored in the tree can be obtained by

$$N_s = \sum_{k=1}^{s-1} m n_k,$$

because the algorithm must look at each of the nodes below a consistent node, even if not all of these subsidiary nodes are themselves consistent.

In principle, the bounds on the number of k-interpretations are exponentials in k. But because $p_1, p_2 < 1$, we can see that as k increases, the base of the exponent decreases. This suggests that n_k may decrease as k gets large enough, but to establish this formally, we need to relate the probabilities p_1, p_2 to properties of the object, in particular to m. Before we do that, we consider formal bounds on the case of occluded recognition.

3. The Combinatorics of Occluded Objects

We want to extend our analysis to the case in which the scene is cluttered, so that much of the object of interest may be occluded, and so that much of the data obtained may come from objects other than the one of interest. To model this, we will again assume the object has m faces, that there are s sensory fragments, of which c actually lie on the object to be recognized. We need to determine bounds on n_s^*, the number of interpretations, and N_s^*, the number of nodes of the interpretation tree actually examined.

3.1 Model of consistency – occluded case

A node at the k^{th} level of the tree defines an k-interpretation, assigning model faces to the first k data fragments. Each such interpretation can be specified by choosing j (out of c) of the data points lying on the object to be correctly matched to a model face, and choosing $r - j$ of the remaining data points (either lying on the object or not) to be incorrectly matched, with the remaining data points assigned to the wild card. Such an interpretation would have r actual matches, and $k - r$ wild card matches. We denote by $n_{k,r}$ the number of such k, r-interpretations.

We need to determine which of these interpretations are consistent. For the unary constraints, any wild card match is consistent with probability 1, as is any correct match. The remaining $r - j$ incorrect matches each have probability of consistency p_1. Thus, we have

$$q_{i,I} = \begin{cases} 1 & \text{if } i \mapsto I \text{ is correct} \\ 1 & \text{if } I \text{ is the wild card character,} \\ p_1 & \text{otherwise.} \end{cases}$$

Any pair of assignments, both of which are correct, is consistent with probability 1. Any pair of assignments, at least one of which is assigned to the wild card also is consistent with probability 1. Thus, we have

$$q_{i,j;I,J} = \begin{cases} 1 & \text{if } i \mapsto I, j \mapsto J \text{ is correct} \\ 1 & \text{if either } I \text{ or } J \text{ are the wild card character,} \\ p_2 & \text{otherwise.} \end{cases}$$

Using this model of consistency, we can establish the following bounds. The proof is deferred to the appendix.

Proposition 2: Given an object with m faces and given k sensory data points, of which c actually lie on the object, the number of interpretations n_k^* is bounded by

$$n_k^* \leq 2^c - [1+p_2]^c + [1+mp_1p_2^{\frac{1}{2}}]^{k-c}[p_2+1+mp_1p_2^{\frac{1}{2}}]^c + mp_1[1-p_2^{\frac{1}{2}}][1+p_2]^{c-1}[k+p_2(k-c)]$$

and by

$$n_k^* \geq 2^c - [1+p_2^{\frac{k-c}{2}}]^c + [1+(m-1)p_1p_2^{\frac{k-1}{2}}]^{k-c}[1+(m-1)p_1p_2^{\frac{k-1}{2}}+p_2^{\frac{k-1}{2}}]^c + p_1(m-1)[1+p_2]^{c-1}[k+p_2(k-c)] - p_1(m-1)p_2^{\frac{k-1}{2}}[1+p_2^{\frac{k-c}{2}}]^{c-1}[k+p_2^{\frac{k-c}{2}}(k-c)]$$

where p_1 is the probability of a random data-model assignment satisfying unary consistency, and p_2 is the probability of a pair of random data-model assignments satisfying binary consistency.∎

As in the non-occluded case, bounds on the number of nodes explored in the tree can be obtained from

$$N_s^* = \sum_{k=1}^{s-1} mn_k^*.$$

In order to make sense out of these rather messy equations, we again need to relate the probabilities of consistency p_1, p_2 to properties of the objects.

4. Bounding the probability of consistency

In the previous sections, we have derived bounds on the problem, as a function of the probability of consistency. It is desirable, however, to reduce these expressions to ones involving parameters of the problem, in particular, to characteristics of the object models and the sensory data. In the following sections, we derive such expressions, under some simplifying assumptions.

4.1 Consistency in the two dimensional case

We begin with the probability of unary consistency, p_1. If ℓ is the length of the data fragment, and L is the length of the model segment, then the probability that this pairing, made at random, is consistent is given by the probability that

$$\ell \leq L + \epsilon$$

where ϵ is a bound on the error in measuring the length of the data edge. If we let $f(\ell)$ denote the distribution of data lengths, and $F(L)$ denote the distribution of model lengths, then the probability of consistency is simply given by

$$\int_{L=0}^{D} \int_{\ell=0}^{\min(L+\epsilon, D)} f(\ell) F(L)\, d\ell dL$$

where D is the dimension of the image. In the worst case, this is just 1, which holds, for example, when the model segments all have the same length and all of the data fragments are smaller than this length. If other models of length distribution are chosen, a different probability can be derived.

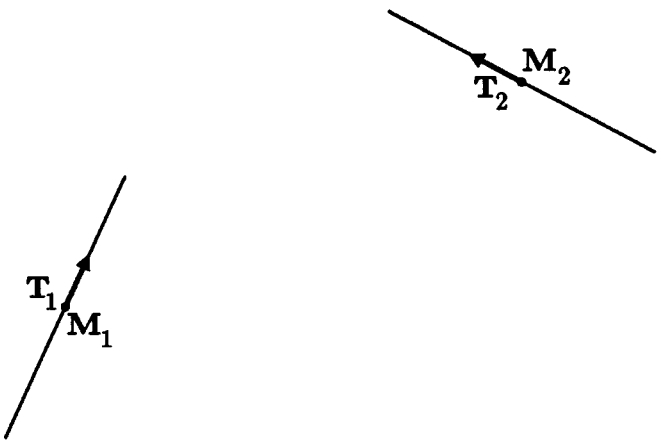

Figure 2. A pair of model edges.

Now we turn to the probability of binary consistency. In the RAF system, a pair of fragments is characterized by the relationship between the fragment normals, and by the components of the family of separation vectors between the fragments. We first transform this representation into a more convenient one.

Claim 1: A pair of edges whose relationship is defined by the ranges of the constraints used in the RAF system are equivalently described by the relative transformation need to align one with the other.

Proof: Consider two model edges, each given by a midpoint, $\mathbf{M}_i$, a unit tangent, $\hat{\mathbf{T}}_i$, and a length L_i, as shown in Figure 2. We can characterize the two edges by the relative transformation needed to transform edge i into edge j. This is given

by the angle θ_{ij} needed to align the tangent vector $\hat{\mathbf{T}}_i$ with the tangent vector $\hat{\mathbf{T}}_j$, and the translation $\mathbf{t}_{ij}$ needed to shift $\mathbf{M}_i$ to $\mathbf{M}_j$. We must first show that such a representation is equivalent to the one used in the constrained search process.

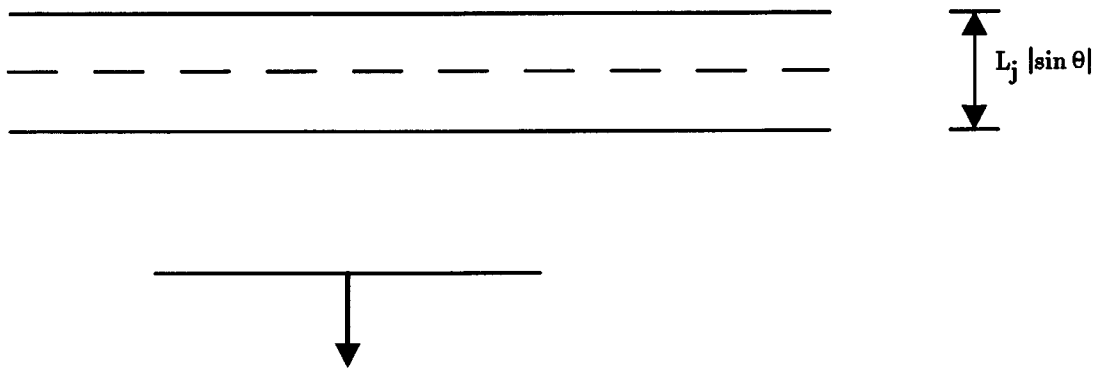

Figure 3. The range of positions for edge j, given a component constraint.

Consider edge i. We are given a range $[c_{i\ell}, c_{ih}]$ of values defining the range of possible components of a separation vector in the direction of the normal to edge i. In general, a separation vector between the two edges is given by

$$\mathbf{S}(\alpha,\beta) = \mathbf{M}_i + \alpha\hat{\mathbf{T}}_i - \mathbf{M}_j - \beta\hat{\mathbf{T}}_j$$

where $\alpha \in [-L_i/2, L_i/2]$ and $\beta \in [-L_j/2, L_j/2]$. Now the actual range of components is given by

$$\langle \mathbf{S}(\alpha,\beta), \hat{\mathbf{T}}_i^{\perp} \rangle = \langle \mathbf{M}_i - \mathbf{M}_j, \hat{\mathbf{T}}_i^{\perp} \rangle - \beta \langle \hat{\mathbf{T}}_j, \hat{\mathbf{T}}_i^{\perp} \rangle$$

where $<,>$ denotes the standard Euclidean inner product. Because β ranges from $-L_j/2$ to $L_j/2$,

$$c_{ih} - c_{i\ell} = L_j |\sin\theta|.$$

Thus any edge that lies entirely within the region shown in Figure 3 is consistent with this constraint.

Now edge j must lie at an angle θ with respect to edge i, and must lie entirely within the range of positions shown in Figure 3. Given the length of edge j and its orientation relative to edge i, this implies that the center of edge j must lie somewhere along the line midway between the two bounds shown. But the same analysis holds relative to edge j, i.e. there is a range of distances perpendicular to it, within which edge i must lie. This is shown in Figure 4. This implies that edge j must have its midpoint along line X such that edge i lies inside the region shown.

As a consequence, there is only one position along the line X such that the midpoint of edge i lies along the line Y. Note that while we have demonstrated this geometrically, it can also be established algebraically.∎

This claim implies that the angle and component constraints used by our recognition system are equivalent to the specification of two edges in terms of their relative

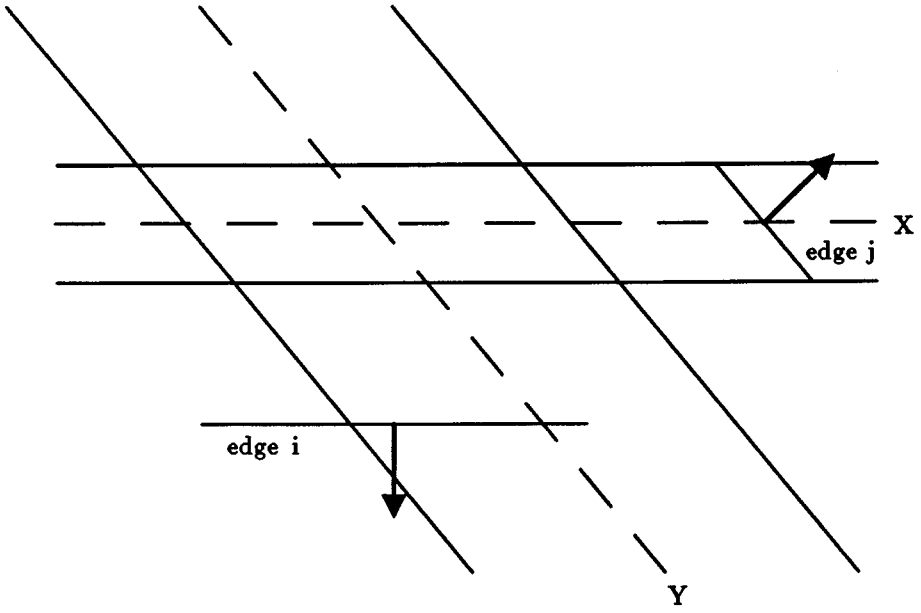

Figure 4. The range of positions for edge i, given a component constraint.

transformation. Hence, a pair of model edges can be equivalently specified in our system in terms of a relative transformation $(\theta_{ij}, \mathbf{t}_{ij})$.

The idea is to use the characterization of a pair by their relative transformation to determine the consistency. Since binary consistency uses pairs of segments, we must relate a pair of data edges to a corresponding pair of model edges. Suppose we are considering the consistency of matching a pair of data edges to a pair of model edges. We know that a pair of model edges are specified by their relative transformation. We need to determine the set of relative transformations that could correspond to a pair of data edges. Note that this is not just the relative transformation between the two data edges. Rather, we want the set of relative transformations of the associated model edges assigned to these data edges. This is important because the problem is compounded by the fact that the data edges may be occluded, so that only part of the corresponding model edge is accounted for, and by the fact that the data edges will be noisy. We assume that position measurements in the data are accurate to within $\pm\epsilon_p$ and that angular measurements are accurate to within $\pm\epsilon_a$.

Because we are only interested in relative transformations, without loss of generality, we position the midpoint of data edge i at the origin of a coordinate frame, with its normal pointing along the negative y axis. The position of the second data edge j relative to this coordinate frame is shown in Figure 5.

Initially, we ignore the effects of noise. Because data edge j may be occluded, the position of the midpoint of the corresponding model edge, if it were transformed into this coordinate frame, would lie along the line defined by the tangent of edge j and the midpoint of the edge, within a distance $\frac{L_j - \ell_j}{2}$ of the midpoint of the

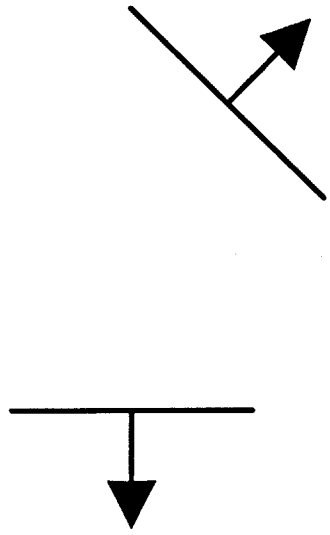

Figure 5. The relative position of data edge j with respect to data edge i.

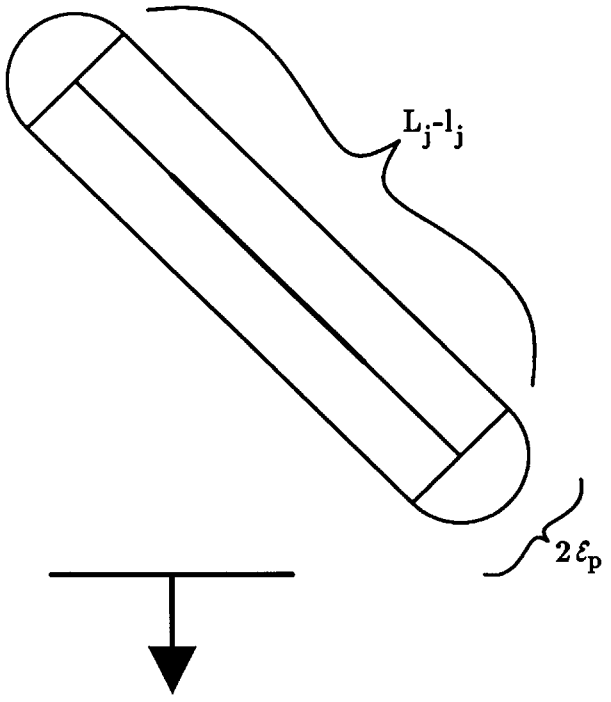

Figure 6. Set of positions for model edge center, given fixed edge.

data edge. When we allow for noise, we must consider any point that lies within a distance ϵ_p of this line. This region of possible positions for the midpoint of the model edge corresponding to data edge j is obtained by sweeping a ball of radius ϵ_p along the line, through a distance of length $L_j - \ell_j$ centered about the midpoint of the edge. This region is shown in Figure 6.

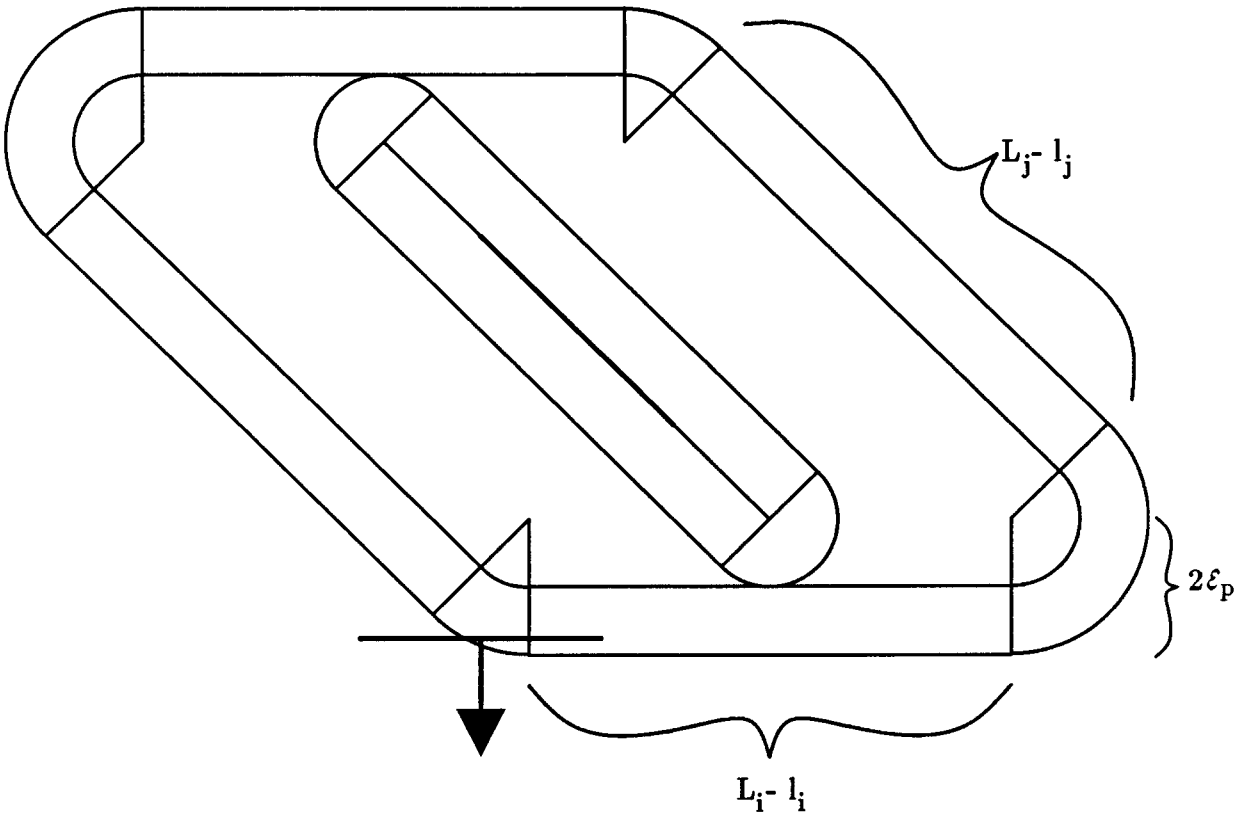

Figure 7. Set of possible positions for relative transformation between two model edges associated with a pair of data edges.

This region shows the range of possible positions for the midpoint of the model edge corresponding to data edge j, given that data edge i is fixed. Because the midpoint lies at the origin, this also gives a set of relative translations. But edge i has the same problem, namely that the centerpoint of the corresponding model edge may actually vary in position. Because we are interested in relative transformations, we can obtain the full set of possible transforms by sweeping the entire region shown in Figure 6 over a distance of $\pm\frac{L_i-\ell_i}{2}$ along the x axis, and then take the set of points lying within a distance ϵ_p of this region. This new region is shown in Figure 7.

This analysis implies that any model edge pair whose relative translation component lies within this area can be considered for consistency. The analysis was performed assuming that the orientation was correctly known. But the relative orientation could also vary within $\pm\epsilon_a$ of the measured angle θ between the data edge normals. For each value, there is a corresponding region of consistent relative translation, which actually changes shape and position, with the center tracing a helical path in this space. Hence, the volume of relative transformation space consistent with a pair of data-model pairings is a skewed extension of the region shown in Figure 7. In the analysis that follows, however, this skewing is not critical.

To estimate the probability of consistency p, we need to know the probability

that a pair of model faces have a relative transformation that falls within the volume described above. To obtain useful results, we will assume that the data edges are uniformly distributed in transform space, so that the probability of consistency is a function only of the relative size of the volume, and not on its actual position in transform space.

To obtain an expression for the volume, we begin with the area shown in Figure 7. By breaking the region into subareas, we find that the total area is given by

$$A(\theta) = 4\pi\epsilon_p^2 + 4\epsilon_p[L_i - \ell_i + L_j - \ell_j] + (L_i - \ell_i)(L_j - \ell_j)|\cos\theta|$$

where θ is the angle between the two edges.

As we have noted, this region will change, as θ varies over the range of values consistent, to within the error bounds, with the measured value, $[\theta_0 - \epsilon_a, \theta_0 + \epsilon_a]$. Thus, the volume of transform space consistent with a pair of data-model assignments is

$$\begin{aligned} V &= \int_{\theta=\theta_0-\epsilon_a}^{\theta=\theta_0+\epsilon_a} A(\theta)\, d\theta \\ &= 2\epsilon_a\left[4\pi\epsilon_p^2 + 4\epsilon_p[L_i - \ell - i + L_j - \ell_j]\right] + 2|\cos\theta_0|\sin\epsilon_a(L_i - \ell_i)(L_j - \ell_j). \end{aligned}$$

To get an estimate of the expected probability of consistency, we will make some simple assumptions. First, we will assume that all the model edges have the same length $L_i = L, \forall i$. We will also assume that the measured edge fragments have at least some minimum length h.

Clearly, the worst case volume occurs for $|\cos\theta_0| = 1$, and $\ell_i = \ell_j = h$. In this case, we have

$$V_w = 8\epsilon_a[\pi\epsilon_p^2 + 2\epsilon_p(L - h)] + 2\sin\epsilon_a(L - h)^2.$$

A more likely case is one in which the data edge lengths are uniformly and independently distributed over the range $[h, L]$. In this case, by evaluating the appropriate integrals, we find that the expected volume is given by

$$V_u = 8\epsilon_a[\pi\epsilon_p^2 + \epsilon_p(L - h)] + |\cos\theta_0|\sin\epsilon_a\frac{(L - h)^2}{2}.$$

If we also assume that θ_0 is uniformly distributed, then

$$V_u = 8\epsilon_a[\pi\epsilon_p^2 + \epsilon_p(L - h)] + \sin\epsilon_a\frac{(L - h)^2}{\pi}.$$

Other models are possible, but these will suffice for our purposes.

Now, we need to relate two factors, the relative transformation associated with a pair of model edges, and the set of relative transformations consistent with a pair of model edges that have been assigned to a pair of data edges. Suppose we consider some point $\mathbf{t}$ in relative translation space. We need to have an expression that denotes the probability that a pair of model edges is consistent at $\mathbf{t}$, which we call $f(t, \theta)$. We also need the probability that a pair of data edges would be consistent at $\mathbf{t}$ (or rather that a pair of model edges matched to this pair of data edges would be consistent at $\mathbf{t}$). We will assume that the data edges are uniformly distributed over

relative transformation space, which has a range of $[0, 2\pi]$ in the rotational dimension and which has a range of $[-D/2, D/2]$ in each of the translation dimensions, where D is the dimension of the image. In this case, the expected probability of a data pair being consistent at any point is simply given by the relative volumes. If we let

$$\epsilon_p^* = \frac{\epsilon_p}{L} \qquad h^* = \frac{h}{L}$$

then the relative volumes are

$$V_w^* = \left[\frac{4\epsilon_a}{\pi}\left[\pi(\epsilon_p^*)^2 + 2\epsilon_p^*(1-h^*)\right] + \frac{\sin\epsilon_a}{\pi}(1-h^*)^2\right]\left[\frac{L}{D}\right]^2$$

$$V_u^* = \left[\frac{4\epsilon_a}{\pi}\left[\pi(\epsilon_p^*)^2 + \epsilon_p^*(1-h^*)\right] + \frac{\sin\epsilon_a}{2\pi^2}(1-h^*)^2\right]\left[\frac{L}{D}\right]^2$$

for the worst case and uniform distribution case respectively. Thus, the expected consistency is given by

$$\int_{\omega=0}^{2\pi}\int_{t=0}^{D} t \text{ Prob(model consistent) Prob(data consistent)}\, dt\, d\omega$$
$$= V^* \int_{\omega=0}^{2\pi}\int_{t=0}^{d} t f(t,\omega)\, dt\, d\omega$$
$$= V^*.$$

Because we assumed that the model edges were of equal length, then

$$L = \frac{P}{m}$$

where P is the perimeter of the object. This finally reduces the probability of consistency to

$$p_2 = \left[\frac{\kappa}{m}\right]^2$$

where

$$\kappa = \kappa_w = \sqrt{\frac{4\epsilon_a}{\pi}\left[\pi(\epsilon_p^*)^2 + 2\epsilon_p^*(1-h^*)\right] + \frac{\sin\epsilon_a}{\pi}(1-h^*)^2}\left[\frac{P}{D}\right]$$

in the worst case, and

$$\kappa = \kappa_u = \sqrt{\frac{4\epsilon_a}{\pi}\left[\pi(\epsilon_p^*)^2 + \epsilon_p^*(1-h^*)\right] + \frac{\sin\epsilon_a}{2\pi^2}(1-h^*)^2}\left[\frac{P}{D}\right]$$

in the uniform distribution case.

Note that κ is a constant that depends only on the error bounds on the sensor, the perimeter of the object, and the size of the image. Unless the perimeter P is very large compared to the size of the image D, we have $\kappa \leq 1$. If the sensor error in measuring position and the minimum edge length are small relative to the length of the model edges, then the constant reduces to

$$\kappa_u = \frac{P}{D}\sqrt{\frac{\sin\epsilon_a}{2\pi^2}}.$$

This leads to:

Proposition 3: Given a two dimensional object with m equal sized edges of length L, and given sensory data that is distributed uniformly in transform space with a uniform distribution of lengths, the expected probability of two random data-model pairings being consistent, p_2, is given by

$$p_2 = \left[\frac{\kappa}{m}\right]^2$$

where

$$\kappa = \kappa_w = \sqrt{\frac{4\epsilon_a}{\pi}\left[\pi(\epsilon_p^*)^2 + 2\epsilon_p^*(1-h^*)\right] + \frac{\sin\epsilon_a}{\pi}(1-h^*)^2}\left[\frac{P}{D}\right]$$

in the worst case, and

$$\kappa = \kappa_u = \sqrt{\frac{4\epsilon_a}{\pi}\left[\pi(\epsilon_p^*)^2 + \epsilon_p^*(1-h^*)\right] + \frac{\sin\epsilon_a}{2\pi^2}(1-h^*)^2}\left[\frac{P}{D}\right]$$

in the uniform distribution case, and where ϵ_a is a bound on the error in measuring orientation, ϵ_p is a bound on the error in measuring position, h is the minimum length data edge, $\epsilon_p^* = \frac{\epsilon_p}{L}$, $h^* = \frac{h}{L}$, P is the perimeter of the object, and D is the dimension of the image.∎

4.2 Consistency in the three dimensional case

A similar analysis can be performed for the case of three dimensional recognition. As in the two dimensional case, we use the angle between two face normals, and the range of components between two faces, in the direction of each of the normals and in the direction of the cross product of the normals, to prune the search. Here we assume for simplicity that each object is modeled by m faces, each a square of size L. Using methods similar to those employed in the previous section, we can show that the probability of consistency is proportional to L^3. In this case, the surface area of the object S is related to the number of faces in the object by $S = mL^2$. Hence we can obtain:

Proposition 4: Given a three dimensional object with m equal sized square faces of side L, and given sensory data that is distributed uniformly in transform space, the probability of two random data-model pairings being consistent p is bounded by

$$p_2 = \frac{\kappa_3^2}{m^{\frac{3}{2}}}$$

where the constant κ_3 is a dimensionless unit depending on bounds on the error in the sensory data and on the ratio of the surface area S to the size of the image.∎

5. Specific Bounds on Recognition

The point of this analysis is that we can relate the probability of consistency p_2 to properties of the recognition problem, specifically to the amount of sensory error relative to the object parameters, $(\epsilon_a, \epsilon_p^*, h^*)$, and the actual parameters of the object itself, (number of faces m and perimeter P). We can now use this to establish particular bounds on the recognition method.

5.1 Bounds on the non-occluded case

We begin with explicit bounds on the number or interpretations obtained in the case of data obtained from a single object. In the appendix, we provide a proof of the following assertion.

Proposition 5: If all of the k sensory measurements are known to lie on a single two-dimensional object with m equal sized edges of length L, and the sensory data is distributed uniformly in transform space, with a uniform length distribution, then the number of k-interpretations is rapidly assymptotic to 1.∎

This is not surprising, because it says that if we exclude objects with symmetries from consideration, and if we have enough data fragments from a single object, there will only be one interpretation. On the other hand it is reassuring to see that the analysis correctly predicts this effect. For most objects and most sensory error ranges, the upper bound rather rapidly approaches 1, so that even with $k = 3$, the expected number of interpretations is basically 1. This is consistent with empirical data.

For the amount of search needed to find the interpretations, we show in the appendix that under some simple assumptions on the amount of noise, the search is at most quadratic in the size of the problem.

Proposition 6: If all of the k sensory measurements are known to lie on a single two-dimensional object with m equal sized edges of length L, $m \geq 2$, the sensory data is distributed uniformly in transform space, with a uniform length distribution, and if the noise is small enough, then the expected amount of search needed to find the interpretation is bounded by

$$m^2 \leq N_s \leq m^2 + ams$$

where a is a constant that depends on the object characteristics and the amount of noise in the sensory measurements.∎

In the appendix, we provide a proof of this, giving a specific definition of "small enough", and a specific definition of the constant a. In particular, we note that the conditions for the definition of "small enough" are satisified for most sensing situtations. For example, if the relative sensing error and the minimum edge length are .1, that is, the error in determining position is no more than one tenth the length of the model edges, then so long as the perimeter of the object is less than 5 times the dimension of the image, the proposition is satisfied. Even when the error rises to .5, the perimeter can be roughly as large as the image dimensions.

Note that the two bounds are reasonably close. Also note that under the assumptions of the analysis, in general, we need only explore $m(s+m)$ nodes. Because there are ms possible initial hypotheses for pairing data edges with model edges, this implies that the constrained search method will rapidly converge to the correct interpretation.

This analysis has been performed using a model in which the consistency of a pair of model-data assignments was taken as 1 if the assignment were correct, and as p if not. This excludes objects with partial symmetries from consideration. Note that we could amplify our analysis by generalizing the notion of consistency to:

$$p = \begin{cases} 1, & \text{if both assignments are correct} \\ q, & \text{if only one assignment is correct} \\ p, & \text{if neither assignment is correct.} \end{cases}$$

For the case of three dimensional recognition, a similar result holds:

Proposition 7: If all of the k sensory measurements are known to lie on a single three-dimensional object with m equal sized edges of dimension L, $m \geq 2$, the sensory data is distributed uniformly in transform space, with a uniform area distribution, and if the noise is small enough, then the number of interpretations is assymptotic to 1, and the expected amount of search needed to find the interpretation is bounded by

$$m^2 \leq N_s \leq m\left[m + \kappa_3^2 m^{\frac{1}{2}} + 2\kappa_3 m^{\frac{1}{4}} + s\right]. \blacksquare$$

Both of these results indicate that while the total number of possible interpretations is exponential, namely m^s, the constrained search method is quite effective at finding the correct interpretation, requiring only a quadratic amount of search. This result is reflected in empirical studies. It suggests that the constraints, even in the presence of sensor noise, are quite powerful. The analysis has excluded objects with symmetries, so that in practical situations the amount of search may be larger, but it is expected to remain polynomial in the problem size.

5.2 Bounds on the occluded case

We can use similar methods to reduce the rather messy expressions we derived

earlier for the expected number of interpretations in the case of occluded data. The appendix contains a proof of the following.

Proposition 8: If c_0 of the k sensory measurements lie on a two-dimensional object with m equal sized edges of length L, the sensory data is distributed uniformly in transform space, with a uniform length distribution, and if the noise is small enough, then the expected number of interpretations, for m large, is bounded by

$$2^{c_0} \leq n_k^* \leq 2^{c_0} + \left[1 + p_1\kappa\right]^k + p_1 mk\left(1 - \frac{\kappa}{m}\right)\left[1 + \frac{\kappa^2}{m^2}\right]^{c_0}$$

where κ is a constant the depends on the object characteristics and the amount of sensor noise, and p_1 is the probability of a random data-model assignment satisfying unary consistency.∎

The lower bound is not as tight as we could make it, but we will use this simple bound for convenience.

Note that the bounds in Proposition 8 make intuitive sense. Consider the correct interpretation, which involves the correct assignment of c_0 of the data points. Not only will this assignment lead to an intepretation, but so will any subset of this assignment. Hence, there must be at least the power set of c_0 possible interpretations, which accounts for the 2^{c_0} term. Any interpretation of length 1 will also be included, because only pairwise constraints are used to reduce the search. This accounts for the mk term. The remaining terms essentially imply that if the sensory error bounds are large enough, some additional interpretations will also be included. If, however, the sensory error bounds are small enough that $\kappa \ll 1$, then basically only the interpretations described above will be found.

Note, of course, that these interpretations involve different amounts of real matches. As discussed in [Grimson and Lozano-Pérez 87], we can adjust our recognition method to accept the longest (in terms of number of data points accounted for) interpretation. This adjustment will in fact reduce the overall amount of search required, because the depth first search may be terminated at any node such that even if all the nodes below that point were to be correctly matched, the length of the resulting interpretation will be less than the best interpretation found so far.

For the amount of search expected in the occluded case, we can use the above result to obtain the following (a proof is found in the appendix).

Proposition 9: If c_0 of the k sensory measurements lie on a two-dimensional object with m equal sized edges of length L, the sensory data is distributed uniformly in transform space, with a uniform length distribution, and if the noise is small enough, then the expected amount of search needed to find the interpretations, for

m large, is bounded by

$$N_s^* \leq m\left[\frac{[1+p_1\kappa]^s}{p_1\kappa} + 2^{c_0}[s-c_0+1] + p_1 m\left[\frac{1}{\alpha^2} + [1+\alpha]^{c_0}\left[\binom{s}{2} - \binom{c_0}{2} + \frac{c_0}{\alpha(1+\alpha)}\right]\right]\right]$$

$$N_s^* \geq m\left[2^{c_0+1} + s - c_0 - 3\right]$$

where

$$\alpha = \frac{\kappa^2}{m^2}$$

and where κ is a constant the depends on the object characteristics and the amount of sensor noise, and p_1 is the probability of a random data-model assignment satisfying unary consistency.∎

We can see from this result that the introduction of the wild card match puts our search method back into the exponential domain, although the amount of work is still considerably less than the normal British Museum algorithm search. The bounds are not tight, since we used a number of approximations in deriving them. Note that the lower bound consists of two terms

$$m2^{c_0+1} \qquad \text{and} \qquad m(s-c_0).$$

Depending on the actual values for the parameters, one of these two terms will dominate, but for most situations, the exponential term is likely to be the larger.

For the upper bound, there are essentially four different major terms

$$m\left[1+p_1\kappa\right]^s \qquad m(s-c_0)2^{c_0} \qquad m^2(s^2-c_0^2)\left[1+\frac{\kappa^2}{m^2}\right]^{c_0} \qquad \frac{m^6}{\kappa^4}.$$

Again, depending on the actual values of the parameters, one of these terms will dominate. For example, if the noise in the sensory data is large, and there are a large number of spurious measurements, the first term will dominate. On the other hand, if the noise is small, the second term is likely to dominate.

Nonetheless, the analysis implies that in general the introduction of spurious data and the use of the wild card branch in a constrained search method forces the expected complexity of the method into the exponential domain.

A similar analysis may be done for the three-dimensional case.

5.3 Branch and bound search

One way to decrease the work involved in finding an interpretation is to use a type of branch and bound search. In particular, suppose that at each stage during the constrained search, we keep track of the longest (measured in terms of the number of data points assigned to non-wild card model faces) interpretation we have found so far. Suppose we reach a non-leaf node of the interpretation tree, such that the sum of the non-wild card matches assigned so far, plus the number of remaining data points to consider (i.e. the remaining levels of the tree between the current point and

the leaves of the tree) is less than the length of the best interpretation so far found. In this case, we cannot find a better interpretation below this point in the tree, so we can terminate our downward search and backtrack. In principle, such a branch and bound technique should reduce the amount of search performed in finding the best interpretation. We can place a bound on the amount of search in this case by noting that in the best possible case, we would discover an interpretation of length c_0 along the first branch of the tree. As a consequence, the remainder of the search would only have to consider a tree of depth $s - c_0$. Unfortunately, this does not change the lower bound, only the upper bound. Hence, to reduce the search further, we need some additional techniques for restricting the size of the search space. We next consider the use of Hough transforms.

6. Hough transforms

The analysis in the previous sections argues that while the constrained search technique is quite effective when it is known that all of the sensory data comes from a single object, the expected search effort is exponential in the size of the correct interpretation when spurious data is allowed. This increase in required search has also been observed in empirical tests. The increased cost arises in part because the use of the wild card branch as a means of separating real from spurious data is not particularly efficient. One way to improve the performance of our recognition engine is to provide a method for selecting candidate subspaces of the search space, that are much smaller than the full search space and that have a high likelihood of containing little or no spurious data. In our experimental work, we have done this using a Hough transform [e.g. Hough 62, Merlin and Farber 75, Sklansky 78, Ballard 81].

In brief, we use the Hough transform as follows. Each possible pose of an object can be described by specifying the parameters of the rigid transformation needed to take the object from its inherent coordinate system into the sensory coordinate system. In the case of two dimensional data, for example, a transformation can be described by an angle of rotation and a two dimensional vector of translation. Each transformation can be represented as a point in a space of transformations, having one dimension for the rotation angle, and one dimension for each of the translation components. We tesselate this space into buckets, using some predefined spacing, h_θ, h_x, h_y.

One way to extract candidate subspaces of the search space is to find pairings of data and model segments that are consistent with the same pose of the object. Thus, for each sensory data fragment d_i, we compute the transformation needed to align that fragment with each of the model fragments, m_j, in turn. Then, that pairing (d_i, m_j) is placed into the tesselation bucket in the transform space in which

the corresponding transform lies. We do this for all pairings of data and model fragments. When completed, each Hough bucket contains a limited set of data fragments, each of which is associated with a limited set of model patches. The expectation is that random data-model pairings will be dispersed in the tesselated space, while the correct data-model pairings will all fall within the same bucket, because they correspond tothe same pose of the object. Hence, by sorting the buckets on the number of votes (or pairings) they contain, we can isolate likely candidate subspaces.

In the ideal case, the bucket with the largest vote will actually identify the correct interpretation, and since the bucket also defines the associated transformation, in principle, we are done. In practice, however, the Hough transform is not sufficient on its own for solving the recognition problem posed here. There are several reasons for this. The first is that in practice one cannot use infinitesimal sized tesselation buckets. Since the Hough bucket has a finite size, any data-model pairing that falls within that bucket will contribute to the vote in that bucket. As the size of the bucket grows, the difference in transform between data-model pairs that will be associated together also grows. This means that spurious data-model pairings may be accidentally grouped together, potentially scoring a larger vote than the correct interpretation. As well, spurious data-model pairings may be accidentally included with the correct pairings, meaning that additional effort is needed to isolate the correct pairs in a bucket, in order to find the actual size of the interpretation. Secondly, a data-model pairing will in general cast a vote in several Hough buckets, not just a single one. Error in the sensory data will give rise to a set of consistent transformations, rather than a single one. Also, occlusion may cause a data edge to correspond to only a part of a model edge. As a consequence, there is a set of corresponding transformations, one for each possible position of the smaller data edge on the model edge. This implies that each data-model pairing contributes to several Hough buckets, say r, so that the noise level in the transform space is amplified considerably. Finally, while the spurious data-model pairings may well be distributed in the Hough space, the sheer number of such pairings may potentially drown out the size of the vote in the correct Hough bucket. For example, if the replication factor is r as above, and there are m model fragments and s data fragments, then there are ms different pairings of which c_0 are correct. This means that there are $msr - c_0$ noisy pairings distributed throughout the Hough space. If there are b buckets, then the average noise contribution to a Hough bucket is $\frac{msr-c_0}{b}$ which can clearly be of significant size relative to c_0, the size of the correct interpretation.

The effect of all this is that while the Hough transform can be used to order candidate subspaces, it is likely in practical circumstances both that the Hough buckets with the largest number of entries may not contain a correct interpretation, and that a Hough bucket containing a correct interpretation is also likely to have some spurious data fragments included and to have some additional model patches associated with correct data fragments. We see this effect in running the **RAF** system. Hence, in our empirical studies, we have use the Hough transform to select candidate

subspaces, ranked in order. We then apply the RAF technique to the subtree defined by the Hough bucket, that is, we use constrained search on a tree whose levels correspond only to those data fragments that are contained within the bucket, and for each such fragment, we only consider those model fragments associated with it as possible matches. We take the Hough buckets in order, applying the RAF technique to each in order, terminating the search when a correct interpretation of sufficient size is found within a bucket.

6.1 Bounds on occluded recognition, using Hough

This argument implies that one cannot assume that the data-model pairings defined by the contents of a Hough bucket correspond to a correct segmentation of the data into elements that are guaranteed to lie on the object. This is unfortunate, since it means that the expected complexity is still in the exponential domain. Fortunately, for practical purposes, the actual size of the search complexity is considerably reduced, since the parameters of the search problem are also reduced.

We can demonstrate this as follows. Suppose that the contents of a Hough bucket define a new interpretation tree, in which the number of model fragments associated with a data fragment is m', where $m' \ll m$ (as we have observed in practice). Also, suppose that the probability of a random data-model pairing falling within a bucket is given by P_r, so that the expected number of data points contributing to a Hough bucket containing the correct interpretation is $s' = c_0 + P_r(s - c_0)$. The bounds on the amount of search required to isolate the correct interpretation are given by the results of Proposition 9, with s replaced by s' and m replaced by m'. While the expressions are still exponential in form, the key is to observe that the parameters have been reduced from their previous values. In the limit, as $m' \to 1$ and $s' \to c_0$, the bounds tend to

$$N^*_{s'} \leq \frac{[1 + p_1\kappa]^{c_0}}{p_1\kappa} + 2^{c_0} + p_1\left[\frac{1}{\alpha^2} + [1+\alpha]^{c_0}\left[\frac{c_0}{\alpha(1+\alpha)}\right]\right]$$

$$N^*_{s'} \geq 2^{c_0+1} - 3$$

where

$$\alpha = \kappa^2.$$

Hence, the expressions remain exponential, but are tighter than the previous ones. In fact, much tighter upper bounds can be established in this case, but the key point is that the bounds remain exponential. In practical terms, this suggests that for many problems, the constrained search approach may still be applicable, if the characteristics of the problem are small enough. In our empirical testing of the RAF system, for example, elapsed times on the order of a few seconds are commonly observed. As the problem size grows, however, and especially when the scenes become complex, the combinatorics suggests that an exponential search is required and this suggests that other techniques are needed to reduce the cost of recognition.

7. Implications of the combinatorics

The goal of this paper was to establish a theoretical basis in support of empirical observations of the utility of a constrained search approach to object recognition. Our experience with **RAF** suggested that when the sensory data could be assumed to all lie on a single object, the system was very efficient at finding correct interpretations. When spurious data was introduced, however, the use of a wild card branch as the last resort to remove data fragments from consideration lead to a strong increase in the amount of work required to find correct interpretations. The analysis in this paper supports this observation, showing that, under some simple assumptions, the expected search in the case of isolated data is quadratic in the number of data fragments and the number of sensory fragments, while the expected search in the case of spurious data is bounded by an expression that is general dominated by the product of the number of data fragments, the number of model fragments and an exponential denoting the magnitude of the power set of the correct interpretation. While the size of this bound is considerably smaller than that associated with British Museum search, it is still exponential.

To some extent, these results are not surprising. Search methods are well known to be computationally expensive. Indeed, some very successful approaches to recognition use maximal clique techniques to find the correct interpretations [Bolles and Cain 82, Bolles et al. 84], and the maximal clique problem is known to be NP-complete. This simply implies that as the characteristics of the problem domain grow, such approachs may lead to poor solutions, but that for many instances of the problem, the performance is acceptable.

At the same time, however, the analysis implies that a general solution to the recognition problem will require additional methods to reduce the combinatorics. One class of methods involves the use of measures of fit to terminate the search. For example, one can terminate the search once an interpretation is found that accounts for some predefined percentage of the object model. We have used such a technique in applying **RAF** [Grimson and Lozano-Pérez 87], and have found that it can signficantly reduce the search cost. The drawback, of course, is in deciding what constitutes an appropriate measure, and what constitutes an appropriate threshold for termination. Depending on the threshold chosen, such termination procedures may run the danger of accepting false positives.

A second approach is to use grouping to reduce the search, and the analysis in this note suggests strong support for the importance of grouping in recognition. If one can identify groups of sensory fragments that are likely to have come from a single object, without exponential cost in identifying such groups, then it is likely that the expected cost of the search process associated with recognizing an object can be reduced to practical levels. While the Hough transform provides a simple method for doing this, more robust techniques are also emerging, for example, [Jacobs 88]. As such grouping techniques continue to develop, the efficiency and robustness of

associated recognition methods should also improve.

References

Ayache, N. J. and Faugeras, O. D. 1986. HYPER: A new approach for the recognition and positioning of two-dimensional objects. *IEEE Trans. PAMI* **8**(1):44–54.

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

Bolles, R. C., and Cain, R. A. 1982. Recognizing and locating paritally 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. In *Robotics Research: The First International Symposium*, M. Brady and R. Paul, Eds. Cambridge, Mass: MIT Press, pp. 413–424.

Freuder, E. C. 1978. Synthesizing constraint expressions. *Comm. of the ACM*, **21**(11), pp. 958–966.

Freuder, E. C. 1982. A sufficient condition for backtrack-free search. *J. ACM*, **29**(1), pp. 24–32.

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

Grimson, W. E. L., 1986a. The combinatorics of local constraints in model-based recognition and localization from sparse data. *J. ACM* **33**(4):658–686.

Grimson, W. E. L., 1986b. Sensing strategies for disambiguating among multiple objects in known poses, *IEEE J. Rob. & Aut.*, **2**, 196–213.

Grimson, W. E. L. 1987a. On the recognition of curved objects. MIT AI Lab Memo 983.

Grimson, W. E. L., 1987b. Recognition of object families using parameterized models, First Intl. Conf. on Computer Vision, London, England, June 1987, pp. 93–101.

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-Pŕez, T. 1987. Localizing overlapping parts by searching the interpretation tree. *IEEE Trans. PAMI* **9**(4):469–482.

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

Haralick, R. M. and Shapiro, L. G. 1979. *IEEE Trans. Pattern Anal. Machine Intell.* PAMI-1(4):173–184.

Hough, P. V. C. 1962. Methods and means for recognizing complex patterns. *U.S. Patent 3069654.*

Jacobs, D. 1988. The use of grouping in visual object recognition. M.S. Thesis, Dept. of EE and CS, Massachusetts Institute of Technology.

Mackworth, A. K. 1977. Consistency in networks of constraints. *Artificial Intelligence*, Vol. 8, pp. 99–118.

Mackworth, A. K. and Freuder, E. C. 1985. The complexity of some polynomial network consistency algorithms for constraint satisfaction problems. *Artificial Intelligence*, Vol. 25, pp. 65–74.

Merlin, P. M., and Farber, D. J. 1975. A parallel mechanism for detecting curves in pictures. *IEEE Trans. Computers*, **13**:96–98.

Montanari, U. 1974. Networks of constraints: Fundamental properties and applications to picture processing. *Inform. Sci.*, Vol. 7, pp 95–132.

Murray, D. W. 1987. Model-based recognition using 3d shape alone. *Computer Vision, Graphics, and Image Processing* **40**(2):250–266.

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

Porrill, J., Pollard, S. B., Pridmore, T. P., Bowen, J. B., Mayhew, J. E. W., and Frisby, J. P. 1987. TINA: The Sheffield AIVRU vision system. University of Sheffield AIVRU Memo.

Sklansky, J. 1978. On the Hough technique for curve detection. *IEEE Trans. Computers* **27**:923–926.

Van Hove, P. 1987. Model-based silhouette recognition. *IEEE Workshop on Computer Vision* Miami Beach, pp. 88–93.

Waltz, D. 1975. Understanding line drawings of scenes with shadows. in *The Psychology of Computer Vision*, P. Winston, Ed. New York:McGraw Hill, pp 19 – 91.

Appendix

In the appendix, we establish formal proofs for the results cited in the text. We begin with bounds on the number of interpretations, for data from a single object.

Proposition 1: If all of the k sensory measurements are known to lie on a single object with m faces, then the number of interpretations n_k is bounded by

$$n_k \leq \left[1 + (m-1)p_1 p_2^{\frac{k-1}{2}}\right]^k.$$

and by

$$n_k \geq 1 + \left[p_2^{\frac{1}{2}} + p_1(m-1)\right]^k p_2^{\frac{k(k-1)}{2}} - p_2^{\frac{k^2}{2}}$$

where p_1 is the probability of a random data-model assignment satisfying unary consistency, and p_2 is the probability of a pair of random data-model assignments satisfying binary consistency.∎

Proof: To determine the number of nodes of the tree at the k^{th} level of the tree, we note that each such node defines a k-interpretation, that is, an assignment of model faces to the first k data fragments. For such an interpretation, there can be i correct assignments, where $i = 0, 1, \ldots, k$. The i data points that are correctly assigned to model faces may be chosen in

$$\binom{k}{i}$$

ways. For the remaining $k-i$ incorrect assignments, there are $m-1$ possible choices for the assignment of each such incorrect label. By considering all possible values for i, we see that there are

$$\sum_{i=0}^{k} \binom{k}{i}(m-1)^{k-i}$$

nodes at this level. We need to determine which of these are actually consistent. For each node, there are $k-i$ incorrect assignments, and the probability that these all pass the unary constraint is

$$p_1^{k-i}.$$

There are also

$$\binom{k}{2}$$

different pairwise constraints, of which

$$\binom{i}{2}$$

involve correct pairs, that have probability of consistency of 1. The rest of the pairs have a probability of consistency p_2. Thus, the probability of a node being consistent with the binary constraints is given by

$$p_2^{\binom{k}{2}-\binom{i}{2}}.$$

Putting this all together, we obtain

$$n_k = \sum_{i=0}^{k} \binom{k}{i} (m-1)^{k-i} p_1^{k-i} p_2^{\binom{k}{2}-\binom{i}{2}}. \tag{1}$$

Note, by the way, that if $p_1 = p_2 = 1$, this reduces to

$$n_k = \sum_{i=0}^{k} \binom{k}{i} (m-1)^{k-i} = (m-1+1)^k = m^k$$

which is the correct expression for the total number of nodes possible at level k of the tree.

Now, we want to obtain bounds on the expression in equation (1). To obtain an upper bound on the expression, we can substitute a smaller exponent for the power of p_2, because $p_2 \leq 1$ implies that a lower exponent will result in a larger expression. In particular, we have

$$n_k \leq \sum_{i=0}^{k} \binom{k}{i} (m-1)^{k-i} p_1^{k-i} p_2^{\frac{(k-1)(k-i)}{2}}.$$

But this simplifies to

$$n_k \leq \left[1 + (m-1) p_1 p_2^{\frac{k-1}{2}}\right]^k. \tag{2}$$

For a lower bound, we can first expand out the $i = k$ term, and then replace the exponent for p by a larger expression:

$$n_k = 1 + \sum_{i=0}^{k-1} \binom{k}{i} (m-1)^{k-i} p_1^{k-i} p_2^{\binom{k}{2}-\binom{i}{2}}$$

$$n_k \geq 1 + \sum_{i=0}^{k} \binom{k}{i} (m-1)^{k-i} p_1^{k-i} p_2^{\frac{k(k-1)+i}{2}} - p_2^{\frac{k^2}{2}}$$

$$n_k \geq 1 + \left[p_2^{\frac{1}{2}} + p_1(m-1)\right]^k p_2^{\frac{k(k-1)}{2}} - p_2^{\frac{k^2}{2}}. \tag{3}$$

∎

For occluded objects, bounds on the expected number of interpretations is given by the following result.

Proposition 2: Given an object with m faces and given k sensory data points, of which c actually lie on the object, the number of interpretations n_k^* is bounded by

$$n_k^* \leq 2^c - \left[1 + p_2\right]^c + \left[1 + m p_1 p_2^{\frac{1}{2}}\right]^{k-c} \left[p_2 + 1 + m p_1 p_2^{\frac{1}{2}}\right]^c$$
$$+ m p_1 \left[1 - p_2^{\frac{1}{2}}\right] \left[1 + p_2\right]^{c-1} \left[k + p_2(k-c)\right]$$

and by

$$n_k^* \geq 2^c - \left[1 + p_2^{\frac{k-c}{2}}\right]^c + \left[1 + (m-1) p_1 p_2^{\frac{k-1}{2}}\right]^{k-c} \left[1 + (m-1) p_1 p_2^{\frac{k-1}{2}} + p_2^{\frac{k-1}{2}}\right]^c$$
$$+ p_1(m-1)\left[1 + p_2\right]^{c-1} \left[k + p_2(k-c)\right]$$
$$- p_1(m-1) p_2^{\frac{k-1}{2}} \left[1 + p_2^{\frac{k-c}{2}}\right]^{c-1} \left[k + p_2^{\frac{k-c}{2}}(k-c)\right]$$

where p_1 is the probability of a random data-model assignment satisfying unary consistency, and p_2 is the probability of a pair of random data-model assignments satisfying binary consistency.

Proof:

A node at the k^{th} level of the tree defines an k-interpretation, assigning model faces to the first k data fragments. Each such interpretation can be specified by choosing j (out of c) of the data points lying on the object to be correctly matched to a model face, and choosing $r-j$ of the remaining data points (either lying on the object or not) to be incorrectly matched, with the remaining data points assigned to the wild card. Such an interpretation would have r actual matches, and $k-r$ wild card matches. We denote by $n_{k,r}$ the number of such k,r-interpretations. Note that for each of the $r-j$ selections, there is an upper bound of m possible assignments, and a lower bound of $m-1$ assignments.

We need to determine which of these interpretations are consistent. For the unary constraints, any wild card match is consistent with probability 1, as is any correct match. The remaining $r-j$ incorrect matches each have probability of consistency p_1. Thus, we have

$$p_{i,I} = \begin{cases} 1 & \text{if } i \mapsto I \text{ is correct} \\ 1 & \text{if } I \text{ is the wild card character,} \\ p_1 & \text{otherwise.} \end{cases}$$

Any pair of assignments, both of which are correct, is consistent with probability 1. Any pair of assignments, at least one of which is assigned to the wild card also is consistent with probability 1. Thus, we have

$$q_{i,j;I,J} = \begin{cases} 1 & \text{if } i \mapsto I, j \mapsto J \text{ is correct} \\ 1 & \text{if either } I \text{ or } J \text{ are the wild card character,} \\ p_2 & \text{otherwise.} \end{cases}$$

Hence, to derive bounds on the number of consistent nodes, we need only consider pairs of assignments chosen from the r actual matches. There are $\binom{r}{2}$ such pairs. Of these, however, $\binom{j}{2}$ have a consistency of 1, because they correspond to correct matches. Thus, the number of interpretations of length r from k sensory points is bounded by

$$n^*_{k,r} \geq \sum_{j=0}^{c} \binom{c}{j}\binom{k-j}{r-j}(m-1)^{r-j}p_1^{r-j}p_2^{\binom{r}{2}-\binom{j}{2}}$$

$$n^*_{k,r} \leq \sum_{j=0}^{c} \binom{c}{j}\binom{k-j}{r-j}m^{r-j}p_1^{r-j}p_2^{\binom{r}{2}-\binom{j}{2}}. \tag{4}$$

Finding tight, closed form expressions for the bounds in equation (4) is somewhat difficult. Instead, we consider the total number of intepretations,

$$n^*_k = \sum_{r=0}^{k} n_{k,r}.$$

We first consider an upper bound on this expression:

$$n_k^* \leq \sum_{r=0}^{k} \sum_{j=0}^{c} \binom{c}{j} \binom{k-j}{r-j} m^{r-j} p_1^{r-j} p_2^{\binom{r}{2}-\binom{j}{2}}.$$

We begin by considering the sum over r:

$$\sum_{r=0}^{k} \binom{k-j}{r-j} m^{r-j} p_1^{r-j} p_2^{\binom{r}{2}} = \sum_{t=0}^{k-j} \binom{k-j}{t} m^t p_1^t p_2^{\binom{t+j}{2}}.$$

The dominant terms in this sum will be for small t, because $p_2 < 1$, hence we expand out the first few terms, yielding

$$p_2^{\binom{j}{2}} + m p_1 (k-j) p_2^{\binom{j+1}{2}} + \sum_{t=2}^{k-j} \binom{k-j}{t} m^t p_1^t p_2^{\binom{t+j}{2}}.$$

To get an upper bound on this expression, we need to replace the exponent of p_2 with a smaller linear expression in t, so that the above sum is bounded above by

$$p_2^{\binom{j}{2}} + m p_1 (k-j) p_2^{\binom{j+1}{2}} + \sum_{t=2}^{k-j} \binom{k-j}{t} m^t p_1^t p_2^{\frac{(j+1)(t+j)}{2}}$$

or

$$p_2^{\binom{j}{2}} - p_2^{\binom{j+1}{2}} + m p_1 (k-j) \left(p_2^{\binom{j+1}{2}} - p_2^{\frac{(j+1)^2}{2}} \right) + \left[1 + m p_1 p_2^{\frac{j+1}{2}} \right]^{k-j} p_2^{\binom{j+1}{2}}. \tag{5}$$

We can now consider the summation over j, treating each of the terms above in turn. Taking the first two terms of (5) yields

$$\sum_{j=0}^{c} \binom{c}{j} p_2^{-\binom{j}{2}} \left[p_2^{\binom{j}{2}} - p_2^{\binom{j+1}{2}} \right] = \sum_{j=0}^{c} \binom{c}{j} [1 - p_2^j] = 2^c - [1+p_2]^c. \tag{6}$$

We can bound the third term of (5) as follows

$$\sum_{j=0}^{c} \binom{c}{j} [1 + m p_1 p_2^{\frac{j+1}{2}}]^{k-j} p_2^{\binom{j+1}{2}-\binom{j}{2}} \leq \sum_{j=0}^{c} \binom{c}{j} [1 + m p_1 p_2^{\frac{1}{2}}]^{k-j} p_2^j$$

$$= [1 + m p_1 p_2^{\frac{1}{2}}]^{k-c} [p_2 + 1 + m p_1 p_2^{\frac{1}{2}}]^c. \tag{7}$$

The final two terms of equation (5) become

$$m p_1 \sum_{j=0}^{c} \binom{c}{j} (k-j) p_2^j \left(1 - p_1^{\frac{j+1}{2}} \right)$$

and this can be bounded above by replacing the exponent with a smaller expression,

$$m p_1 \left(1 - p_2^{\frac{1}{2}} \right) \sum_{j=0}^{c} \binom{c}{j} (k-j) p_2^j.$$

To reduce this, we note that if we let

$$f(x) = (1+x)^n = \sum_{i=0}^{n} \binom{n}{i} x^i$$

then

$$\frac{df(x)}{dx} = n(1+x)^{n-1} = \sum_{i=0}^{n} \binom{n}{i} i x^{i-1}$$

so that

$$\sum_{i=0}^{n} \binom{n}{i} i x^i = nx(1+x)^{n-1}.$$

Hence,

$$\sum_{i=0}^{n} \binom{n}{i} (a-i)x^i = a(1+x)^n - nx(1+x)^{n-1} = (1+x)^{n-1}[a + x(a-n)].$$

Thus, the final two terms of (5) reduce to

$$mp_1\left(1 - p_2^{\frac{1}{2}}\right)(1+p_2)^{c-1}[k + p_2(k-c)]. \tag{8}$$

By combining equations (6)–(8), we get

$$\begin{aligned} n_k^* \leq 2^c - [1+p_2]^c + [1 + mp_1 p_2^{\frac{1}{2}}]^{k-c}[p_2 + 1 + mp_1 p_2^{\frac{1}{2}}]^c \\ + mp_1[1 - p_2^{\frac{1}{2}}][1+p_2]^{c-1}[k + p_2(k-c)]. \end{aligned} \tag{9}$$

We can use a similar approach to obtain a lower bound on the number of interpretations. We have

$$n_k^* \geq \sum_{r=0}^{k} \sum_{j=0}^{c} \binom{c}{j} \binom{k-j}{r-j} (m-1)^{r-j} p_1^{r-j} p_2^{\binom{r}{2} - \binom{j}{2}}.$$

As before, we begin with the summation over r, which reduces to

$$\sum_{t=0}^{k-j} \binom{k-j}{t} (m-1)^t p_1^t p_2^{\binom{t+j}{2}}.$$

Expanding out terms yields

$$p_2^{\binom{j}{2}} + (m-1)p_1(k-j)p_2^{\binom{j+1}{2}} + \sum_{t=2}^{k-j} \binom{k-j}{t} (m-1)^t p_1^t p_2^{\binom{t+j}{2}}.$$

In this case, we need to replace the exponent for p_2 with a linear expression in t which is greater than the current one, because this will lead to smaller expressions in p_2. We obtain

$$p_2^{\binom{j}{2}} + (m-1)p_1(k-j)p_2^{\binom{j+1}{2}} + \sum_{t=2}^{k-j} \binom{k-j}{t} (m-1)^t p_1^t p_2^{\frac{(k-1)(t+j)}{2}}$$

or

$$\begin{aligned} p_2^{\binom{j}{2}} + (m-1)p_1(k-j)p_2^{\binom{j+1}{2}} + [1 + (m-1)p_1 p_2^{\frac{k-1}{2}}]^{k-j} p_2^{\frac{j(k-1)}{2}} \\ - (m-1)p_1(k-j)p_2^{\frac{(j+1)(k-1)}{2}} - p_2^{\frac{j(k-1)}{2}}. \end{aligned}$$

Using the same methods as before, this reduces to

$$\begin{aligned} n_k^* \geq 2^c - [1 + p_2^{\frac{k-c}{2}}]^c + [1 + (m-1)p_1 p_2^{\frac{k-1}{2}}]^{k-c}[1 + (m-1)p_1 p_2^{\frac{k-1}{2}} + p_2^{\frac{k-1}{2}}]^c \\ + p_1(m-1)[1+p_2]^{c-1}[k + p_2(k-c)] \\ - p_1(m-1)p_2^{\frac{k-1}{2}}[1 + p_2^{\frac{k-c}{2}}]^{c-1}[k + p_2^{\frac{k-c}{2}}(k-c)]. \end{aligned} \tag{10}$$

∎

Once we have a relationship between the probability of consistency and the parameters of the problem, we can derive specific bounds on the number of interpretations. In section 4 of the paper, we derive such relationships.

Proposition 5: If all of the k sensory measurements are known to lie on a single two-dimensional object with m equal sized edges of length L, and the sensory data is distributed uniformly in transform space, with a uniform length distribution, then the number of k-interpretations is assymptotic to 1.

Proof: From equations (2) and (3) we have

$$n_k \leq \left[1 + (m-1)p_1 p_2^{\frac{k-1}{2}}\right]^k \tag{2}$$

$$n_k \geq 1 + \left[p_2^{\frac{1}{2}} + p_1(m-1)\right]^k p_2^{\frac{k(k-1)}{2}} - p_2^{\frac{k^2}{2}}. \tag{3}$$

In the case of two dimensional recognition, we substitute from Proposition 3 to get:

$$\begin{aligned} n_k &\leq \left[1 + \frac{\kappa^{k-1}}{m^{k-1}} p_1(m-1)\right]^k \\ n_k &\geq 1 + \left[\left(\frac{\kappa}{m}\right)^{k-1}\left(\frac{\kappa}{m} + p_1(m-1)\right)\right]^k - \left[\frac{\kappa}{m}\right]^{k^2}. \end{aligned} \tag{11}$$

To establish the result, we need to show that

$$(1 + ax^k)^k \geq (1 + ax^{k+1})^{k+1}$$

for some k, where $x < 1$. This is equivalent to showing that

$$\left(1 + \frac{ax^k(1-x)}{1 + ax^{k+1}}\right)^k \geq 1 + ax^{k+1}.$$

If we can show that

$$1 + \frac{kax^k(1-x)}{1 + ax^{k+1}} \geq 1 + ax^{k+1}$$

then we are done, because the left hand side is just the first two terms of the expanded product. To establish this, we simply need to show that

$$k(1-x) \geq (1 + ax^{k+1})x$$

for some k, but this is clearly true if $x < 1$.

Thus, for k large enough, both the upper and the lower bounds tend to 1.∎

To establish bounds on the amount of search needed, we use the following:

Proposition 6: If all of the k sensory measurements are known to lie on a single two-dimensional object with m equal sized edges of length L, $m \geq 2$, the sensory data is distributed uniformly in transform space, with a uniform length distribution, and if the noise is small enough, then the expected amount of search needed to find the interpretation is bounded by

$$m^2 \leq N_s \leq m^2 + ams$$

where a is a constant that depends on the object characteristics and the amount of noise in the sensory measurements.∎

Proof: To get bounds on the amount of search in the two dimensional case, recall that this amount is given by:

$$N_s = \sum_{k=1}^{s-1} mn_k.$$

To bound this, we could simply find the largest term in the summation, and use ms times that term as an upper bound, since there are s terms in the sum. To do this explicitly, we first consider the constant κ, given by Proposition 3

$$\kappa = \kappa_u = \sqrt{\frac{4\epsilon_a}{\pi}[\pi(\epsilon_p^*)^2 + \epsilon_p^*(1-h^*)] + \frac{\sin\epsilon_a}{2\pi^2}(1-h^*)^2}\left[\frac{P}{D}\right]$$

To ease the analysis, we will restrict our attention to cases in which $\kappa < 1$, although a similar analysis will hold for other cases. To do this, we note that the error in determining angles can be obtained as a function of the error in determining position, by considering the worst case deviation, which yields $\epsilon_a = \tan^{-1} 2\epsilon_p^*$. Thus, we have:

Claim: If the perimeter of an object P, the dimension of the image D and the error in measuring positions relative to the length of a model edge $\epsilon_p^* = \frac{\epsilon_p}{L}$ satisfy the relationship:

$$P\sqrt{4\tan^{-1}(2\epsilon_p^*)\left((\epsilon^*)^2 + \frac{\epsilon_p^*}{\pi}(1-\epsilon_p^*)\right) + \frac{\epsilon_p^*}{\pi^2}\frac{(1-\epsilon_p^*)^2}{\sqrt{1+4(\epsilon_p^*)^2}}} < D$$

then

$$\kappa < 1.∎$$

This follows naturally from Proposition 3. It is worth noting that the conditions for this proposition are satisfied for most situations. For example, if the relative sensing error and the minimum edge length are .1, that is, the error in determining position is no more than one tenth the length of the model edges, then so long as the perimeter of the object is less than 5 times the dimension of the image, the proposition is satisfied. Even when the error rises to .5, the perimeter can be roughly as large as the image dimensions.

If the proposition holds, it is straightforward to show that the upper bound for n_k given in equation (11) is a maximum for $k = 1$, in this case being equal to $p_1 m + 1 - p_1$. Because there are roughly s terms in the summation, this leads to the bound

$$N_s \leq m^2 s.$$

(Note that since p_1 is generally a constant, independent of m, using the upper bound of $p_1 \leq 1$ does not radically change the derived bound.) We can improve on this,

however, by noting that if $\kappa \leq 1$, then, from equation (11),

$$n_1 \leq p_1 m + 1 - p_1$$
$$n_2 \leq [1 + p_1 \kappa]^2$$
$$n_3 \leq \left[1 + p_1 \frac{\kappa^2}{m}\right]^3$$

We want to show that under the conditions of Proposition 6, the upper bound on n_k is monotically decreasing. By taking the derivative of this expression with respect to k and considering the worst case, in which $\kappa = 1, p_1 = 1$, we need to establish that

$$(1 + m^{2-k}) \log(1 + m^{2-k}) + m^{2-k} k \log \frac{1}{m} \leq 0.$$

For $k \geq 3$, we can approximate the first log by its second argument, so that we need to establish that

$$1 + m^{2-k} \leq k \log m.$$

Since the left hand side decreases with increasing k and the right hand side increases with increasing k, we need only establish this for $k = 3$. This expression holds for $k = 3$ if $m \geq 1.698$, which is trivial to assume. Hence, the expression is monotically decreasing for $k \geq 3$, and case analysis shows this also holds for $k = 1, 2$. Hence, for large m, s we have

$$N_s \leq p_1 m^2 + (1 + p_1 \kappa)^2 ms + m(1 - p_1). \tag{12}$$

Note that if a tighter constant is desired, we can expand out several more terms in the summation, before bounding the remainder.

Similarly, the lower bound on n_k given in equation (11) is a maximum for $k = 1$, having the value m, provided $\kappa < 1$. Thus

$$N_s \geq p_1 m^2 + m(1 - p_1). \tag{13}$$

If we simply let $p_1 = 1$, we establish the proposition.∎

For the three-dimensional case, we have a similar argument.

Proposition 7: If all of the k sensory measurements are known to lie on a single three-dimensional object with m equal sized edges of dimension L, $m \geq 2$, the sensory data is distributed uniformly in transform space, with a uniform area distribution, and if the noise is small enough, then the number of interpretations is assymptotic to 1, and the expected amount of search needed to find the interpretation is bounded by

$$m^2 \leq N_s \leq m\left[m + \kappa_3^2 m^{\frac{1}{2}} + 2\kappa_3 m^{\frac{1}{4}} + s\right].$$

Proof:

For the case of three dimensional recognition, we substitute from Proposition 4 into equations (2) and (3), to get:

$$1 + \left[\frac{\kappa_3}{m^{\frac{3}{4}}} + m - 1\right]^k \frac{\kappa_3^{k(k-1)}}{m^{\frac{3k(k-1)}{4}}} - \frac{\kappa_3^{k^2}}{m^{\frac{3k^2}{4}}} \leq n_k \leq \left[1 + \frac{(m-1)\kappa_3^{k-1}}{m^{\frac{3(k-1)}{4}}}\right]^k. \tag{14}$$

For $k = 1$, we have

$$m \leq n_1 \leq m$$

for $k = 2$, we have

$$1 + \kappa_3^2\left[m^{\frac{1}{4}} - \frac{1}{m^{\frac{3}{4}}} + \frac{\kappa_3}{m^{\frac{3}{2}}}\right]^2 - \frac{\kappa_3^4}{m^3} \leq n_2 \leq \left[1 + \kappa_3 m^{\frac{1}{4}} - \frac{\kappa_3}{m^{\frac{3}{4}}}\right]^2$$

and for $k = 3$, we have

$$1 + \kappa_3^6\left[\frac{1}{m^{\frac{1}{2}}} - \frac{1}{m^{\frac{3}{2}}} + \frac{\kappa_3}{m^{\frac{9}{4}}}\right]^3 - \frac{\kappa_3^9}{m^{\frac{27}{4}}} \leq n_3 \leq \left[1 + \kappa_3^2\left(\frac{1}{m^{\frac{1}{2}}} - \frac{1}{m^{\frac{3}{2}}}\right)\right]^3.$$

Again, as k continues to increase, we have

$$n_k \Rightarrow 1.$$

As before, we can substitute to obtain the desired expressions.∎

For the case of occluded objects, we can use equations (9) and (10):

$$\begin{aligned} n_k^* \leq\ & 2^c - \left[1 + p_2\right]^c + \left[1 + mp_1 p_2^{\frac{1}{2}}\right]^{k-c}\left[p_2 + 1 + mp_1 p_2^{\frac{1}{2}}\right]^c \\ & + mp_1\left[1 - p_2^{\frac{1}{2}}\right]\left[1 + p_2\right]^{c-1}\left[k + p_2(k-c)\right] \qquad (9) \\ n_k^* \geq\ & 2^c - \left[1 + p_2^{\frac{k-c}{2}}\right]^c + \left[1 + (m-1)p_1 p_2^{\frac{k-1}{2}}\right]^{k-c}\left[1 + (m-1)p_1 p_2^{\frac{k-1}{2}} + p_2^{\frac{k-1}{2}}\right]^c \\ & + p_1(m-1)\left[1 + p_2\right]^{c-1}\left(k + p_2(k-c)\right) \\ & - p_1(m-1)p_2^{\frac{k-1}{2}}\left[1 + p_2^{\frac{k-c}{2}}\right]^{c-1}\left[k + p_2^{\frac{k-c}{2}}(k-c)\right]. \qquad (10) \end{aligned}$$

Two dimensional case

To relate these bounds on the number of interpretations to characteristics of the objects, we substitute from (11). This gives

$$\begin{aligned} n_k^* \leq\ & 2^c - \left[1 + \frac{\kappa^2}{m^2}\right]^c + \left[1 + \kappa p_1\right]^{k-c}\left[1 + \kappa p_1 + \frac{\kappa^2}{m^2}\right]^c \\ & + mp_1\left[1 - \frac{\kappa}{m}\right]\left[1 + \frac{\kappa^2}{m^2}\right]^{c-1}\left[k + \frac{\kappa^2}{m^2}(k-c)\right] \qquad (15a) \\ n_k^* \geq\ & 2^c - \left[1 + \left(\frac{\kappa}{m}\right)^{k-c}\right]^c \\ & + \left[1 + (m-1)p_1\left(\frac{\kappa}{m}\right)^{k-1}\right]^{k-c}\left[1 + (m-1)p_1\left(\frac{\kappa}{m}\right)^{k-1} + \left(\frac{\kappa}{m}\right)^{k-1}\right]^c \\ & + p_1(m-1)\left(1 + \frac{\kappa^2}{m^2}\right)^{c-1}\left(k + \frac{\kappa}{m}(k-c)\right) \\ & - p_1(m-1)\left(\frac{\kappa}{m}\right)^{k-1}\left[1 + \left(\frac{\kappa}{m}\right)^{k-c}\right]^{c-1}\left[k + \left(\frac{\kappa}{m}\right)^{k-c}(k-c)\right]. \qquad (15b) \end{aligned}$$

To reduce this to a more manageable form, we will assume that the conditions of Proposition 6 hold. If m is also large, then this rather messy bound reduces to the following.

Proposition 8: If c_0 of the k sensory measurements lie on a two-dimensional object with m equal sized edges of length L, the sensory data is distributed uniformly in transform space, with a uniform length distribution, and if the noise is small enough, then the expected number of interpretations, for m large, is bounded by

$$2^{c_0} \leq n_k^* \leq 2^{c_0} + [1 + p_1\kappa]^k + p_1 mk\left(1 - \frac{\kappa}{m}\right)\left[1 + \frac{\kappa^2}{m^2}\right]^{c_0}. \blacksquare$$

Now we turn to the problem of bounding the amount of search required in this case. We establish the following claim.

Proposition 9: If c_0 of the k sensory measurements lie on a two-dimensional object with m equal sized edges of length L, the sensory data is distributed uniformly in transform space, with a uniform length distribution, and if the noise is small enough, then the expected amount of search needed to find the interpretations, for m large, is bounded by

$$N_s^* \leq m\left[\frac{[1 + p_1\kappa]^s - [1 + p_1\kappa]}{p_1\kappa} + 2^{c_0}[s - c_0 + 1] - 2 \right.$$
$$\left. + p_1 m\left(1 - \frac{\kappa}{m}\right)\left[\frac{1}{\alpha^2} + [1 + \alpha]^{c_0}\left[\binom{s}{2} - \binom{c_0}{2} + \frac{\alpha(c_0 - 1) - 1}{\alpha^2(1 + \alpha)}\right]\right]\right]$$
$$N_s^* \geq m\left[2^{c_0+1} + s - c_0 - 3\right]$$

where

$$\alpha = \frac{\kappa^2}{m^2}.$$

Proof:

The upper bound on the search is given by:

$$N_s^* \leq m\left[\sum_{k=1}^{s-1} 2^{c(k)} + [1 + p_1\kappa]^k + p_1 mk\left(1 - \frac{\kappa}{m}\right)\left[1 + \frac{\kappa^2}{m^2}\right]^{c(k)}\right].$$

The second term is simply a geometric series, and is easily reduced to closed form. To obtain explicit bounds on the other terms of the summation, however, we need to know something about the subset of the data fragments that are part of the correct interpretation, that is, we need to know how c changes with k.

In general,

$$c(k) = \begin{cases} 0, & k < i_1 \\ 1, & i_1 \le k < k_2 \\ \vdots & \\ c_0, & i_{c_0} \le k \le s-1 \end{cases}$$

in which case

$$\sum_{k=1}^{s-1} 2^{c(k)} = (i_1 - 1)2^0 + (i_2 - i_1)2^1 + \ldots + (i_{c_0} - i_{c_0-1})2^{c_0-1} + (s - 1 - i_{c_0})2^{c_0}.$$

The worst case for this sum is when $i_j = j$, in which case, the sum reduces to

$$(s - c_0 - 1)2^{c_0} + \sum_{i=0}^{c_0-1} 2^i$$

$$(s - c_0 - 1)2^{c_0} + 2^{c_0} - 2$$

$$2^{c_0}[s - c_0 + 1] - 2.$$

Now consider the term

$$\left(1 - \frac{\kappa}{m}\right) p_1 m \sum_{k=1}^{s-1} k[1 + \alpha]^{c(k)} \qquad \alpha = \frac{\kappa^2}{m^2}.$$

By the above assumption about $c(k)$ the summation part of this becomes

$$\sum_{k=1}^{i_1-1} k[1+\alpha]^0 + \sum_{k=i_1}^{i_2-1} k[1+\alpha]^1 + \ldots \sum_{k=i_{c_0-1}}^{i_{c_0}-1} k[1+\alpha]^{c_0-1} + [1+\alpha]^{c_0} \sum_{k=i_{c_0}}^{s-1} k$$

and again the worst case is when $i_j = j$, in which case, the sum reduces to

$$\sum_{k=1}^{c_0-1} k[1+\alpha]^{k-1} + [1+\alpha]^{c_0} \sum_{k=i_{c_0}}^{s-1} k$$

$$= [1+\alpha]^{c_0} \left[\binom{s}{2} - \binom{c_0}{2} \right] + \sum_{k=1}^{c_0-1} k[1+\alpha]^{k-1}.$$

To bound the remaining summation, we can use the arithmetico-geometric progression:

$$\sum_{k=0}^{n-1} (a + kr)q^k = \frac{a - [a + (n-1)r]q^n}{1-q} + \frac{rq(1 - q^{n-1})}{(1-a)^2}.$$

In our case we have $a = 0, r = 1$ and $q = 1 + \alpha$, so that

$$\sum_{k=1}^{c_0-1} k[1+\alpha]^{k-1} = \frac{1 - [1+\alpha]^{c_0-1}}{\alpha^2} + \frac{(c_0 - 1)[1+\alpha]^{c_0-1}}{\alpha}$$

$$= \frac{1 + (\alpha(c_0 - 1) - 1)[1+\alpha]^{c_0-1}}{\alpha^2}.$$

This yields:

$$N_s^* \le m\left(\frac{[1+p_1\kappa]^s - [1+p_1\kappa]}{p_1\kappa} + 2^{c_0}[s-c_0+1] - 2 \right.$$
$$\left. + p_1 m\left(1-\frac{\kappa}{m}\right)\left[\frac{1}{\alpha^3} + [1+\alpha]^{c_0}\left[\binom{s}{2} - \binom{c_0}{2} + \frac{c_0(c_0-1)-1}{\alpha^3(1+\alpha)}\right]\right]\right).$$

For the lower bound, we have

$$N_s^* \ge m\sum_{h=1}^{s-1} 2^{c(h)}.$$

Here, the worst case occurs when c is 0 for the first $s - c_0 - 1$ terms, and then increases linearly, yielding

$$N_s^* \ge m\left[2^{c_0+1} + s - c_0 - 3\right].\blacksquare$$

This blank page was inserted to preserve pagination.

CS-TR Scanning Project
Document Control Form

Date : 5/5/95

Report # AIM-1019

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
☐ Laboratory for Computer Science (LCS)

Document Type:

☐ Technical Report (TR) ☒ Technical Memo (TM)
☐ Other:_____

Document Information

Number of pages: 41 (47-IMAGES)
Not to include DOD forms, printer intstructions, etc... original pages only.

Originals are:
☒ Single-sided or
☐ Double-sided

Intended to be printed as :
☐ Single-sided or
☒ Double-sided

Print type:
☐ Typewriter ☐ Offset Press ☒ Laser Print
☐ InkJet Printer ☐ Unknown ☐ Other:_____

Check each if included with document:

☒ DOD Form(2) ☐ Funding Agent Form ☐ Cover Page
☐ Spine ☐ Printers Notes ☐ Photo negatives
☐ Other: _____

Page Data:

Blank Pages(by page number):_____

Photographs/Tonal Material (by page number):_____

Other (note description/page number):

Description :	Page Number:
IMAGE MAP (1) UNNUM'ED TITLE PAGE	
(2-41) PAGES #'ED 1-40	
(42-44) SCANCONTROL, DOD (2)	
(45-47) TRGTS (3)	

Scanning Agent Signoff:
Date Received: 5/5/95 Date Scanned: 5/9/95 Date Returned: 5/11/95

Scanning Agent Signature: Michael W. Cook

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AI Memo 1019	2. GOVT ACCESSION NO.: AD-A196224
3. RECIPIENT'S CATALOG NUMBER:	
4. TITLE (and Subtitle): The Combinatorics of Object Recognition in Cluttered Environments Using Constrained Search	5. TYPE OF REPORT & PERIOD COVERED: memorandum
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): W. Eric L. Grimson	8. CONTRACT OR GRANT NUMBER(s): DACA76-85-C-0010 N00014-85-K-0124
9. PERFORMING ORGANIZATION NAME AND ADDRESS: MIT Artificial Intelligence Laboratory 545 Technology Square Cambridge, MA 02139	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:
11. CONTROLLING OFFICE NAME AND ADDRESS: Advanced Research Projects Agency 1400 Wilson Blvd. Arlington, VA 22209	12. REPORT DATE: February 1988 13. NUMBER OF PAGES: 41
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): Office of Naval Research Information Systems Arlington, VA 22217	15. SECURITY CLASS. (of this report): Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

object recognition
Hough transform
combinatoric complexity

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Abstract. The problem of recognizing rigid objects from noisy sensory data has been successfully attacked in previous work by using a constrained search approach. Empirical investigations have shown the method to be very effective when recognizing and localizing isolated objects, but less effective when dealing with occluded objects where much of the sensory data arises from objects other than the one of interest. When clustering techniques such as the Hough transform are used to isolate likely subspaces of the search space, empiricial performance in cluttered scenes improves considerably. In this note, we establish

Block 20 cont.

formal bounds on the combinatorics of this approach. Under some simple assumptions, we show that the expected complexity of recognizing isolated objects is quadratic in the number of model and sensory fragments, but that the expected complexity of recognizing objects in cluttered environments is exponential in the size of the correct interpretation. We also provide formal bounds on the efficacy of using the Hough transform to preselect likely subspaces, showing that problem remains exponential, but that in practical terms, the size of the problem is significantly decreased.

Scanning Agent Identification Target

Scanning of this document was supported in part by the **Corporation for National Research Initiatives,** using funds from the **Advanced Research Projects Agency** of the **United states Government** under Grant: **MDA972-92-J1029.**

The scanning agent for this project was the **Document Services** department of the **M.I.T Libraries.** Technical support for this project was also provided by the **M.I.T. Laboratory for Computer Sciences.**