Measuring the Quality of Hypotheses in Model-Based Recognition*

Daniel P. Huttenlocher[1] *and Todd A. Cass*[2]

[1] Department of Computer Science, Cornell University, Ithaca NY 14853, USA
[2] AI Lab, Massachusetts Institute of Technology, Cambridge MA 02139, USA

Abstract. Model-based recognition methods generally search for *geometrically consistent* pairs of model and image features. The quality of an hypothesis is then measured using some function of the number of model features that are paired with image features. The most common approach is to simply count the number of pairs of consistent model and image features. However, this may yield a large number of feature pairs, due to a single model feature being consistent with several image features and vice versa. A better quality measure is provided by the size of a maximal bipartite matching, which eliminates the multiple counting of a given feature. Computing such a matching is computationally expensive, but under certain conditions it is well approximated by the number of *distinct* features consistent with a given hypothesis.

1 Introduction

A number of different model-based techniques have been used to hypothesize instances of a given model in an image, including searching for possible correspondences of model and image features (e.g., [4]), the generalized Hough transform (e.g., [1]), alignment of a model with an (e.g., [5]), and analysis of the space of model transformation parameters [2]. While these methods differ substantially, they all measure the quality of a given hypothesis as a function of the number of geometric features of the model that are consistent with geometric features extracted from the image. The larger the *consistent set* of model and image features, the better an hypothesis is judged to be.

In this paper we analyze some of the most common methods for assessing the quality of an hypothesis in model-based recognition. We focus on the case in which objects are modeled as a collection of 'atomic' features (such as points). In other words, each individual model feature is either paired with an image feature or not (there is no partial matching of individual features). The three quality measures that we investigate are: (i) the number of pairs of model and image features that are consistent with an hypothesis, (ii) the maximum bipartite matching of such features, and (iii) the number of distinct such features. We find that the first measure, although widely used, often greatly overestimates the quality of a match. The second method is more expensive to compute, but is also much more conservative. In practice we find that the third scoring method is the best. This final method also turns out to be closely related to some recent theoretical work

* This report describes research done in part at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's research is provided in part by an ONR URI grant under contract N00014-86-K-0685, and in part by DARPA under Army contract number DACA76-85-C-0010 and under ONR contract N00014-85-K-0124. DPH is supported at Cornell University in part by NSF grant IRI-9057928 and matching funds from General Electric and Kodak, and in part by AFOSR under contract AFOSR-91-0328.

on using Hausdorff distances for recognition [6]. Finally, we show that if restrictions are placed on the spacing of the model features, then the third measure becomes equivalent to the size of the maximal matching. This provides a more precise meaning to the notion that matching is more difficult when the features of a model are too close together.

Our results have two important implications for model-based recognition systems:

1. For atomic features (such as points), the quality of an hypothesis should be measured based on the number of distinct model and image features that are geometrically consistent with a given hypothesis, not on the number of consistent model and image feature pairs.
2. Object models should be constructed such that no two features are close together, where closeness is a function of the degree of sensory uncertainty.

Existing model-based recognition systems, where appropriate, can be modified with minimal effort to take advantage of these results. Using the number of distinct features as a quality measure simply requires keeping track of which features are accounted for by a given set of feature pairs. Constructing the object models requires that the sensor error estimates used by the recognition method be available when the models are made.

2 Geometrically Consistent Sets of Features

We assume that a model consists of a set of features $M = \{m_i\}_{i=1,\ldots,m}$ measured in some coordinate system $\mathcal{M}$, and the image data consist of a set of features $S = \{s_j\}_{j=1,\ldots,n}$ measured in some coordinate system $\mathcal{I}$. A pair of model and image features, (m_i, s_j) defines a set of possible transformations from $\mathcal{M}$ to $\mathcal{I}$. This set of transformations can be viewed as a volume in a d-dimensional *transformation parameter space*, where each dimension of the space corresponds to one of the parameters of the transformation, $\mathcal{T} : \mathcal{M} \rightarrow \mathcal{I}$. We denote the set, or volume, of transformations consistent with a pair of features (m_i, s_j) as $V(m_i, s_j) \subset \mathcal{T}$. This is the set of all transformations on m_i that map it into $\mathcal{I}$, such that m_i is within some uncertainty region around s_j. We limit the present discussion to the case of point features, where the uncertainty region is modeled using a disk of radius ϵ. These results can be generalized to more complex models of uncertainty, including non-circular positional uncertainty and angular uncertainty (cf. [2, 3]).

As an example of a transformation space volume, consider a model and an image that both consist of a set of points in the plane, and a three-dimensional transformation space, $\mathcal{T}$ with axes u, v and θ (corresponding respectively to the two translations and one rotation in the plane). If there is no sensory uncertainty in the image measurements ($\epsilon = 0$), then the set $V(m_i, s_j)$ is the helical arc ℓ in $\mathcal{T}$ defined as a function of θ by $s_j - \mathbf{R}_\theta m_i$ where $\mathbf{R}_\theta$ is a rotation matrix, and s_j and m_i are the positions of image and model feature points, respectively. Thus the translation is constrained exactly by a given rotation, but any rotation is possible. If the location of the image point is uncertain up to a circle of radius ϵ, then $V(m_i, s_j)$ is a helical tube of circular cross section of radius ϵ about the curve ℓ.

We define a *consistent set* of feature pairs, C, to be a set of model and image feature pairs that specify mutually overlapping sets of transformations. Thus formally $C \subseteq M \times S$ is a consistent set when the intersection of all the volumes specified by the pairs in C is nonempty,

$$\bigcap_{(m_i, s_j) \in C} V(m_i, s_j) \neq \emptyset. \tag{1}$$

The sets of 'geometrically consistent features' computed by most recognition systems are a slight *overestimate* of the volumes specified in equation (1). For example, the interpretation tree approach (e.g., [4]) only ensures pairwise consistency between model and image feature pairs. That is, a path through the interpretation tree specifies a set of feature pairs $J \subseteq M \times S$, such that for each pair $(m_i, s_j) \in J$ and $(m_k, s_l) \in J$, $V(m_i, s_j) \cap V(m_k, s_l) \neq \emptyset$. The tree search does not guarantee that the set J (corresponding to a path in the tree) is a consistent set by the definition in equation (1), because only *pairs* of volumes are required to intersect. Similarly, generalized Hough transform methods (e.g., [1]) overestimate the consistent feature sets, because the transformation space is tessellated into discrete buckets. These buckets are in general larger than the true transformation space volumes (if they were smaller then a match could be missed altogether).

We further note that the size of C, from equation (1) is itself generally an overestimate of the quality of an hypothesis. The key problem is that a given set of consistent feature pairs may contain many pairs that involve a given model feature m_i or a given image feature s_j. Each of these pairs is counted separately, even though it involves the same model or image feature. For atomic features, where there is no partial matching of a single feature, it does not make much sense to count the same feature multiple times. A set of sensor features $T \subseteq S$ will all be paired with the same model feature, $m_i \in M$, in the same consistent set if $\bigcap_{s_j \in T} V(m_i, s_j) \neq \emptyset$. This happens for any set of sensor features T in which all the features are 'close' together, where close means that the corresponding transformation space volumes intersect (i.e., closeness is a function of the sensor uncertainty). Similarly, a set of model features can all be paired with a given image feature.

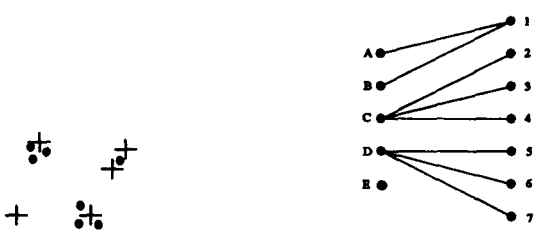

Fig. 1. The number of consistent feature pairs overestimates the quality of a match: a) a superimposed set of model and image features, b) the corresponding bipartite graph.

Thus whenever a group of model features or a group of image features are close together, the size of a consistent set will count the same model or image feature multiple times. For example, Figure 1a shows an alignment of a five point model with an image for which the corresponding consistent set contains eight feature pairs. The model points are shown as crosses and the image points are shown as dots. Any model point m_i that lies within ϵ of an image point s_j defines a pair of features (m_i, s_j) that are consistent with this position of the model in the image. On the other hand, only three of the five model features are paired with distinct image features. In this case the difference in the size of the consistent set (eight) and the number of model features accounted for (three) occurs because a single image feature is paired with more than one model feature and vice versa. Such situations are not merely of theoretical interest, as they occur frequently for reasonable ranges of sensory uncertainty and images containing just tens of features.

3 Bipartite Matching

Maximal bipartite matching can be used to rule out the 'multiple counting' that occurred in the above example, by requiring that each model feature only be paired with a single image feature and vice versa. A consistent set of feature pairs, C, defines a bipartite graph $G = (U, V, E)$ where for each feature pair $(m_i, s_j) \in C$ there is a vertex $u_i \in U$ corresponding to the model feature m_i, a vertex $v_i \in V$ corresponding to the image feature s_j, and an edge $e_{ij} \in E$ connecting u_i to v_j. Each edge is incident on one vertex of U and one vertex of V, so the graph is bipartite. For example, the set of consistent feature pairs corresponding to Figure 1a is $\{(A,1),(B,1),(C,2),(C,3),(C,4),(D,5),(D,6),(D,7)\}$ (where model features are denoted by letters and image features by numbers). This set defines the bipartite graph shown in Figure 1b.

A *matching* in a bipartite graph, G, is a subset of the edges, $F \subseteq E$, such that each vertex of G has at most one edge incident on it. The size of F is the number of edges that it contains, $|F|$. For instance, a trivial matching is a single edge of a bipartite graph. A *maximal matching* is one such that there are no larger matchings in the graph. For our problem, a maximal matching corresponds to the largest set of consistent model and image feature pairs that can be formed without using any model or image feature more than once. For the graph in Figure 1b the set $\{(A,1),(C,2),(D,5)\}$ is a maximal matching. Note that in general there can be more than one maximal matching.

Methods for finding a maximal matching in a bipartite graph require $O(|V|^{1/2} \cdot |E|)$ time, or $O(n^{2.5})$ where $n = |V|$. Methods that are straightforward to implement require time $O(\min(|V|, |U|) \cdot |E|)$ time [7]. In contrast, simply counting the number of pairs in a consistent set (i.e., the number of edges in E) only requires time $O(|E|)$. As model based recognition methods already require substantial amounts of running time, we are concerned with how to estimate the size of the maximal matching, $|F|$, in $O(|E|)$ time.

From the bipartite graph representation of a geometrically consistent set, we can identify several possible measures of the quality of the corresponding hypothesis:

1. The number of feature pairs in the consistent set (i.e., $|C| = |E|$).
2. The number of distinct features accounted for by the consistent set (e.g., $|U|$, $|V|$, or their minimum, maximum, or sum).
3. The size of a maximal matching in the bipartite graph defined by I (i.e., $|F|$ where $F \subseteq E$ is a maximal matching).

Given the way we have constructed a bipartite graph from C, the first of these three quantities is the largest, and the last is the smallest, that is, $|E| \geq \min(|U|, |V|) \geq |F|$. Clearly the first inequality holds because the number of edges in the graph must be at least as large as the number of vertices of each type. The second inequality holds because in a matching each vertex has at most one edge incident on it.

Whereas most recognition systems use the first of these measures, the last is the most conservative measure. Some recognition systems do use bipartite matchings, but these are quite expensive to compute compared with counting the number of consistent feature pairs. Thus we propose using the number of distinct pairs of model and image features, as measured by the quantity $\min(|U|, |V|)$. This is cheap to compute, and measures the minimum number of distinct model or image features accounted for.

The measure $\min(|U|, |V|)$ only overestimates the size of a maximal matching, $|F|$, when there is branching on both sides of the bipartite graph. This corresponds to a situation in which there are several neighboring model features that match a single image feature, and vice versa. If a bipartite graph is guaranteed to only have branching at the vertices on one side of the graph, then situations such as this cannot occur. In that case,

it is trivial to compute the size of a maximal matching – simply count the number of vertices on the side of the graph where the branching occurs. For example if the branching occurs only for vertices in U, then each edge $e \in E$ is incident on a unique vertex $v \in V$ (otherwise there would be branching for some vertex in V). Thus the number of vertices in U determines the size of a maximal matching.

While we cannot in general control the spacing of features in an image, we can do so for the features in a model. More formally, in order for no two model features to match a given image feature, it must be that for each pair of model features, $m_i, m_k \in M$, the volumes produced by intersection with the same sensor feature s_j are disjoint, that is, $\forall_{m_i,m_k \in M} V(m_i, s_j) \cap V(m_k, s_j) = \emptyset$. Another way to view this is that no two model features can be close enough together that when mapped into the image coordinate frame they overlap the uncertainty region around the same image feature. For a rigid-body transformation this can be accomplished by surrounding each model feature with a 'buffer area' based on the positional uncertainty value ϵ. As long as no two buffer areas overlap, no pair of model features can match the same image feature.

As an example, consider the case of points in the plane, where the sensory uncertainty region is a circle of radius ϵ. If each model point is surrounded by a circle of radius ϵ, and the model points are placed such that no two circle intersect, then no two model points can match the same image point. Having done this, the number of distinct features is equal to the size of the maximal bipartite matching, but is much cheaper to compute. In practice, we have found this method to be much better than simply counting the number of feature pairs.

In summary, a good estimate of the size of the size of the maximal bipartite matching is provided by the number of *distinct* model and image features that are consistent with a given hypothesis. Moreover, if models are constructed such that no two model features are close together (as a function of the degree of sensory uncertainty) then the number of distinct features is the same as the size of the maximal bipartite matching. This provides a formal meaning to the intuition that matching is harder when the model features are 'too close together to be resolved by the sensor'.

References

1. D. H. Ballard, 1981. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition* **13**:111.
2. Cass, T.A., 1990, "Feature Matching for Object Localization in the Presence of Uncertainty", MIT Artificial Intelligence Laboratory Memo no. 1133.
3. Grimson, W.E.L. and D.P. Huttenlocher, 1990, "On the Verification of Hypothesized Matches in Model-Based Recognition", *Proceedings of the First European Conference on Computer Vision*, Lecture Notes in Computer Science No. 427, pp. 489-498, Springer-Verlag.
4. Grimson, W.E.L. & T. Lozano-Pérez, 1987, "Localizing overlapping parts by searching the interpretation tree," *IEEE Trans. PAMI* **9**(4), pp. 469–482.
5. Huttenlocher, D.P. and S. Ullman, 1990, "Recognizing Solid Objects by Alignment with an Image",*Intl. Journal of Computer Vision*, vol. 5, no. 2, pp. 195-212.
6. Huttenlocher, D.P. and Kedem, K., 1990, "Efficiently computing the Hausdorff distance for point sets under Translation", proceedings of Sixth ACM Symposium on Computational Geometry, pp. 340-349.
7. Papadimitriou, C.H. and K. Steiglitz, 1982, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall.

This article was processed using the LaTeX macro package with ECCV92 style