MEMO #93–02
June 1993

Higher-Order Statistics in Visual Object Recognition

Thomas M. Breuel
IDIAP, C.P. 609, 1920 Martigny, Switzerland
tmb@idiap.ch

ABSTRACT

In this paper, we develop a higher-order statistical theory of matching models against images. The basic idea is not only to take into account *how much* of an object can be seen in the image, but also *what parts* of it are jointly present. We show that this additional information can improve the specificity (i.e., reduce the probability of false positive matches) of a recognition algorithm.

We demonstrate formally that most commonly used quality of match measures employed by recognition algorithms are based on an independence assumption. Using the Minimum Description Length (MDL) principle and a simple scene-description language as a guide, we show that this independence assumption is not satisfied for common scenes, and propose several important higher-order statistical properties of matches that approximate some aspects of these statistical dependencies. We have implemented a recognition system that takes advantage of this additional statistical information and demonstrate its efficacy in comparisons with a standard recognition system based on bounded error matching.

We also observe that the existing use of grouping and segmentation methods has significant effects on the performance of recognition systems that are similar to those resulting from the use of higher-order statistical information. Our analysis provides a statistical framework in which to understand the effects of grouping and segmentation on recognition and suggests ways to take better advantage of such information.

Keywords: higher-order statistics, object recognition, minimum description length, Bayesian decision theory, grouping, segmentation, error rate.

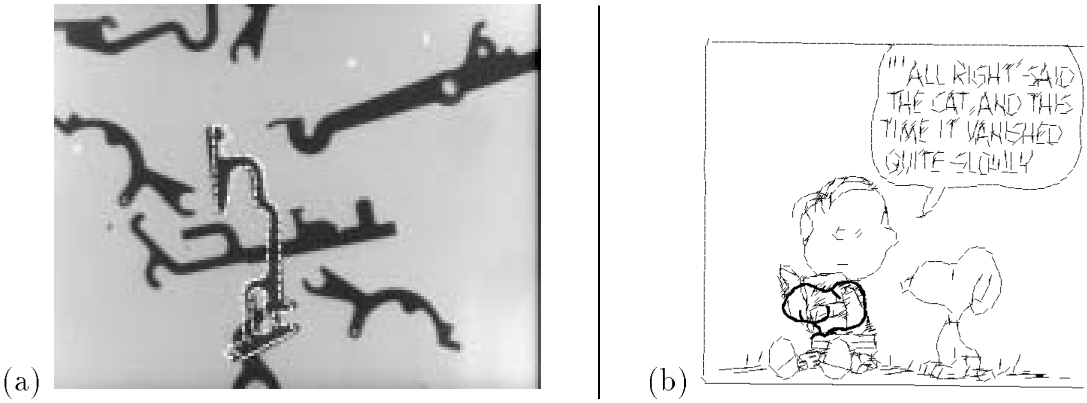
(a)
(b)
"ALL RIGHT" SAID THE CAT, AND THIS TIME IT VANISHED QUITE SLOWLY

Figure 1: Standard recognition algorithms work well for objects with well-defined geometries (a), but fail to recognize even simple natural objects (Snoopy's head, b).

1 Introduction

Most systems for visual object recognition are based on the idea that in order to determine whether an object is present in the scene we need to determine whether a significant fraction of the object is visible in the scene. In addition, such systems make allowances for small variations in shape and appearance of an object due to variations in lighting, sensor error, and model variation.

This basic approach works well for objects with well-defined shapes, like metal widgets, scissors, watch pieces, or other objects encountered in industrial applications (Figure 1a). Unfortunately, experience shows that it fails frequently when objects have less well-defined shapes (Figure 1b), like hand-draw cartoon figures or natural objects like fruits or fish.

The reason is that in the case of objects with well-defined shapes, error bounds only need to account for sensor limitations. In practice, this means that we can use very tight error bounds to achieve high specificity during recognition (high specificity means that the recognition algorithm is unlikely to recognize a random collection of features as some object, i.e., that the recognition algorithm has a low probability of false positive matches). In the case of recognition of natural objects, however, the shape of the object itself can vary greatly, and we need to use other information besides shape to achieve higher specificity during recognition.

In this paper, we argue that we can achieve higher specificity during recognition by not only taking into account *how much* of an object can be seen in the image, but also *what parts* of it can be seen.

That different parts of an object provide different degrees of evidence for the presence of an object in an image is actually not a new idea (see below for

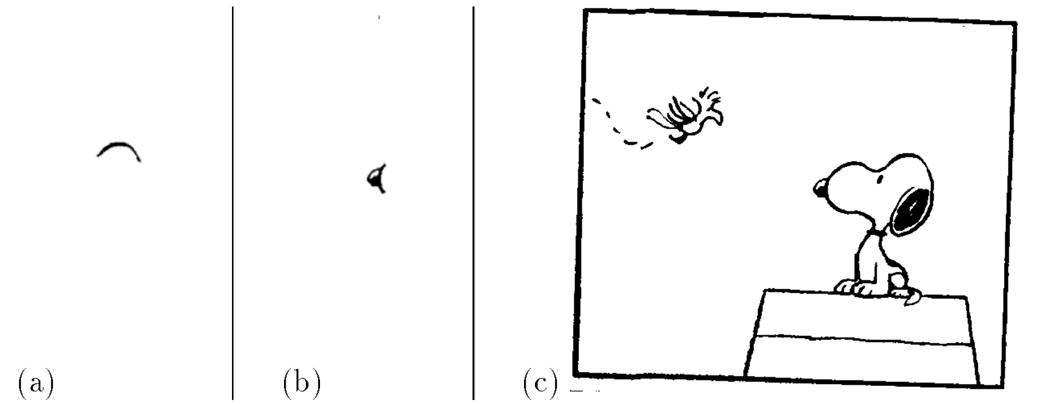

Figure 2: Some object parts are more important than others for recognizing the object. Snoopy's skull provides relatively weak evidence for the presence of a beagle in the image, while the nose is a significantly stronger clue.

references). A simple illustration is given in Figure 2. It is not surprising that a curve segment corresponding to Snoopy's skull provides significantly less evidence for the presence of Snoopy in the image than a picture of Snoopy's nose. This idea has been variously described as the different "saliency", "weight", or "importance" of the parts of an object.

What we will see below is that such notions of saliency are actually only a "first-order" approximation to a statistical theory of object recognition that provides us with rich and very important information about the quality of match between an image and an object.

To illustrate this point, consider the images in Figure 3. Image (a) contains much less total occlusion than image (b). Yet, Charlie Brown is clearly much easier to recognize in image (b) than in image (a). Because many important parts of the image (ears, eyebrows, nose, outline of the head) are almost completely unoccluded in image (a), this difference cannot simply be explained in terms of the occlusion of important parts of the object.

In what follows, we will develop a statistical theory that explains these differences in recognizability and show how we can take advantage of these additional statistical constraints to improve the specificity of recognition algorithms.

Based on this idea, we propose a single, unified statistical framework in which low-level vision (feature extraction), intermediate-level vision (grouping and segmentation), and high-level vision (recognition) can be understood. In particular, we argue that in current recognition systems, the extraction of complex features and the use of grouping and segmentation are not just useful for speeding up recognition, but also are crucial for ensuring specificity of recognition. Furthermore, our statistical theory suggests that vi-

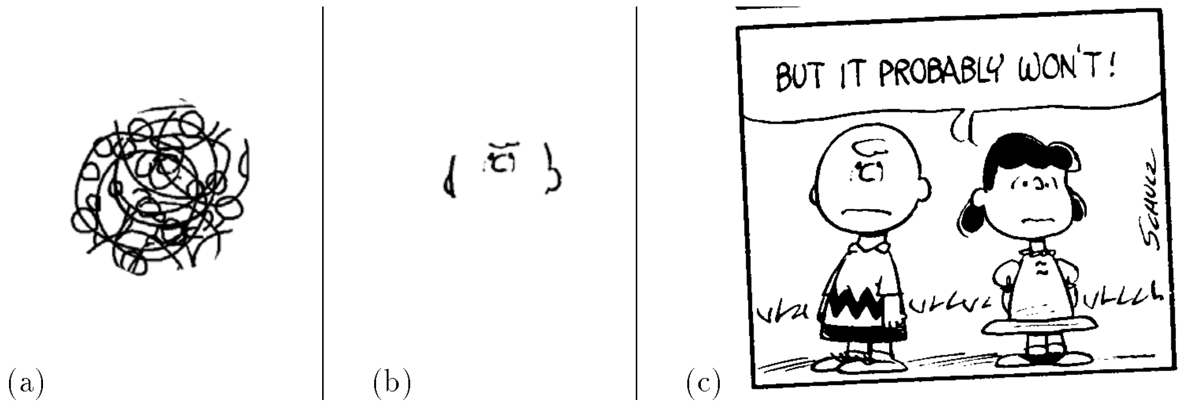

Figure 3: By itself, the presence of important object features in an image is not enough to ensure recognition. There is significantly less total occlusion in (a) than in (b), and yet Charlie Brown is easier to recognize in (b).

sion systems that take advantage of grouping information through higher-order evaluation functions (functions that compute a quality of match using higher-order statistical properties), as proposed in this paper, are likely to be significantly more robust than the current system based on bottom-up, model-independent uses of grouping.

2 Statistical Principles

Why does bounded error recognition fail? Recognition algorithms usually take as a measure of the quality of match between an image and a model the number of features (or fraction of boundary/edge segments) that can be matched between the image and the model subject to given error bounds (see Baird, 1985, and Grimson, 1990, for extensive bibliographies).

As we noted above, bounded error recognition is not very discriminating for real objects: when error bounds become large enough to account for common errors on location, a model will match well in inappropriate places. We find that such inappropriate matches occur when one object model matches small parts of a large number of other objects in the image. This is particularly common in very cluttered images or images that contain textured regions. The figures in this paper provide several examples of such incorrect matches.

To see in more detail how this comes about, consider the match shown in Figure 4. A bounded error recognition algorithm applied to the model of Snoopy's head located the model in the brushes rather than at its correct location. The reason is the following. If we make the error bounds large, in

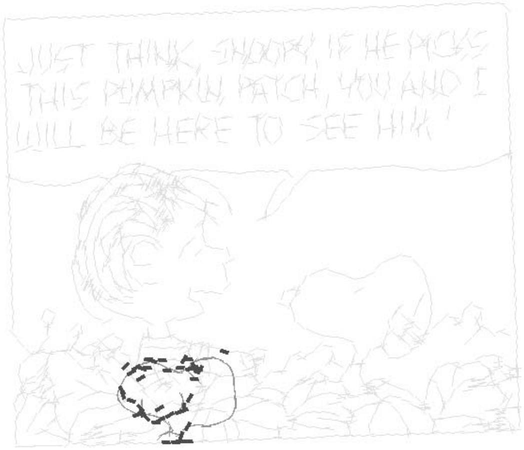
JUST THINK, SNOOPY, IF HE PICKS THIS PUMPKIN PATCH, YOU AND I WILL BE HERE TO SEE HIM!

Figure 4: The matching image segments in a semantically incorrect but geometrically optimal bounded error match.

order to account for the model variation, many of the features making up the brushes might match features from the model well, leading to a spurious match. On the other hand, if we make the error bounds small, then the match between the model and the actual instance of the model in the image is so poor that it will likewise be missed.

In Figure 4, image segments that match some model segments are shown in a dark shade of grey. We note that these segments are scattered somewhat irregularly and form a number of broken up curves. Intuitively, in order to explain such a match, a recognition algorithm would have to postulate a very complicated occluding object that would make a large number of small, disconnected pieces of the matching object visible, and it would have to postulate that the object has undergone considerable local deformations. Clearly, such occlusions and deformations are rather unlikely, and the resulting match should therefore be considered "poor".

Minimum Description Length We can formalize the idea described in the previous paragraph using the *Minimum Description Length* (MDL) principle (Rissanen, 1978). In general, an MDL principle states that the optimal solution to a recognition problem can be found by describing or explaining the input data as concisely as possible in terms of a given (formal) language. The choice of language encodes the prior knowledge about the domain. When applied to the problem of interpreting scenes, this means the following

- We try to describe an unknown image as concisely as possible in terms of known objects, transformations on them, and their mutual occlusions. In different words, we consider that interpretation of an image "optimal" that explains the image in terms of the smallest number of known objects and simplest transformations on them.

We could implement such an MDL approach to object recognition and scene interpretation directly. However, there are several practical reasons for not taking such an approach:

- Finding an optimal or near-optimal description of an image in terms of some language can be a computationally hard problem.
- A complete interpretation of the image may be impossible because it might contain unknown objects.
- Some occluding objects (e.g., trees) may not be describable concisely in terms of a deterministic language, and, in fact, may only need to be *explained* concisely, but not actually described (encoded) concisely.
- An actual recognition system is likely to use adaptive (learning) methods; grammatical inference, however is a difficult problem.

Fortunately, MDL approaches are closely related to statistical methods, and, in particular, Bayesian methods. Essentially, applying an MDL principle and choosing a particular description languages (e.g., for images) corresponds to a choice of a prior probability distribution in a Bayesian framework. The advantage of using an MDL principle to analyze a problem (rather than starting directly with a Bayesian model) is that the MDL principle is a much more intuitive and convenient means for describing and reasoning about higher-order statistical distributions than marginal probability distributions.

Therefore, in what follows, we will be using the MDL principle primarily as a guide to understanding what kinds of statistical constraints might help improve the reliability of a recognition system. But actual implementations of recognition systems will be directly in terms of Bayesian models, and the parameters and, possibly, some structural aspects of such models should ultimately be acquired (learned) directly from visual input.

A Simple Description Language For the purposes of this paper, let us choose a simple description language that corresponds well to a large number of real-world (and cartoon-world) situations.

For the subsequent informal discussion, we will assume that objects are opaque and that they are composed (in 3D or 2D) of simple, approximately convex parts. Furthermore, for transformations, we allow rigid body transformations, small, bounded deviations from the ideal shape, and, possibly, smooth deformations of the object. Note that these assumptions are clearly not satisfied for all objects. For example, some objects are transparent. Natural objects like trees or brushes are better described as textures or fractals. But this does not affect the basic thrust of the argument.

Statistical Implications From our choice of description language above, we can draw several conclusions about statistical properties of matches between a model and an image.

The following two properties follow from the fact that objects (and hence, occluding objects) are composed of only a small number of convex parts:

- If some model feature matches, a nearby model feature is likely to match as well.
- Edges (curved or straight) are unlikely to be broken into a large number of small edges by an occlusion. Hence, the number of edges in a hypothesized match should not be significantly larger than the number of edges in the model.

Intuitively, we can think about this as follows. Consider, for example, a single line in the image. If there are N convex parts occluding this line, it can be broken up into at most $N+1$ parts. By assumption (and this applies to many real-world objects), occluding objects are composed only of few convex parts, and hence N is known to be small. As we will see below, this is a powerful constraint for eliminating spurious matches.

Furthermore, if we consider each edge in the image a sequence of small edge segments ("edgel" features), then the important implication is that it is not just important how many such edge segments are present in the image, or which individual segments are present, but what their joint distribution is. If these differential edge segments are spatially clustered into a few connected components, they are more likely to originate from a single object than if they are clustered into a larger number of connected components.

From the composition of a small number of convex parts and the opaqueness of objects, the following statistical property can be inferred:

- Within the convex closure of matching model features, spurious image features are unlikely.

A final, very important statistical property that has already been used extensively in a bottom-up way in visual object recognition system is that of "non-accidentalness":

- Two model features that define some non-accidental property in the model (e.g. colinearity, parallelism, curvilinear colinearity) are likely to define the same property in the image.

A derivation of this fact and a some actual probability distributions can be found in Lowe, 1986.

Thus, we have seen that some simple knowledge about the world ("objects are opaque and composed of few convex parts") together with the MDL principle imply statistical constraints on the makeup of matches between a model and an image.

We could pursue this line of reasoning by formalizing our world model more carefully and deriving actual probability distributions from it. In fact, some of this work has been carried out already in other fields. For example, probability distributions related to occluding convex objects have been studied extensively in stereology and statistical geometry (Stoyan *et al.*, 1987, Hall, 1988). However, for the purposes of this paper, let it suffice that we make some general observations about the structure and properties of such distributions.

The most important observation is that all the above statistical properties involve not a single feature, but the joint presence and absence of features in the image. We call such properties *higher order*. Evaluation functions (i.e. functions that compute a quality of match between a model and an image for a given transformation) using higher-order statistical properties will be called *higher-order evaluation functions*.

Of significant practical importance is that all the above statistical properties can be evaluated without finding a consistent global interpretation of the image; for example, in order to reject a highly fragmented edge as an unlikely candidate for a match against a single model edge, we do not need to form a complete interpretation of the occlusion that we hypothesize to have given rise to the fragmentation. This is a significant advantage over a strict application of an MDL principle to recognition.

3 Mathematical Model

In the previous section, we presented an informal discussion of how models and transformations together with an MDL principle give rise to higher order statistical constraints. In this section, we will show formally that the most commonly used approaches to recognition correspond to statistical first-order evaluations of the quality of match between a model and an image and are based on a certain independence assumption. We will then motivate and give a specific example of a second-order evaluation function.

Existing Recognition Algorithms A recognition algorithm matches models against images. Usually, a model and an image is a collection of geometrical features associated with their geometric and visual properties (location, orientation, size, configuration, color, texture, etc.).

We can transform this geometric problem into a statistical problem as follows. Assume we are given some transformation T of the model (e.g., a translation). If we apply this transformation to the model, each model feature will be mapped into the image. We say that this model feature *matches* an image feature if it falls within some given error bound of a compatible image feature; compatible means that the model feature and the image feature have similar orientations and similar non-geometric properties (e.g., edge strength, color, etc.). This is the standard definition of bounded error recognition, one of the most commonly used definitions of recognition in computer vision (see, for example, Baird, 1985, Grimson, 1990, for references).

In this way, we can define a *feature vector* $\vec{\phi}(T)$ that contains one entry for each of the model features. $\phi_j(T) = 1$ if model feature j matches some image feature under transformation T. For convenience, in what follows, we will simply not write down the dependence on T explicitly. Associated with each ϕ_j is also some geometric information (location and orientation) in the model and the image; we will notate these as ϕ_j^M and ϕ_j^I, respectively.

Now, let us consider the problem of recognition for a given transformation T as a statistical decision problem. Let there be n different kinds of objects, $\omega_1, \ldots, \omega_n$. Then, a reasonable statistical decision procedure is to return that object ω_i which is most likely given the known input data ϕ (this is a very simplistic version of statistical decision theory and Bayesian analysis and suffices for our purposes here; for more details see Duda and Hart, 1973, Berger, 1980, and Kiefer, 1987):

$$\omega(\vec{\phi}) = \arg\max_{\omega=\omega_1,\ldots,\omega_n} P(\omega|\vec{\phi}) \qquad (1)$$

It is perhaps helpful to relate this formula to the informal discussion in the previous section. There, we considered scenes composed of given objects and discussed the probability of particular constellations of features. That is, as is common in Bayesian analysis, we were setting out with our analysis by considering essentially $P(\vec{\phi}|\omega_j)$ and now have moved to considering $P(\omega_j|\vec{\phi})$. The two points of view are related via Bayes' theorem. The application of Bayes' theorem in practice is not always entirely trivial, but it will turn out that, for the problem of visual object recognition, we can guess useful approximations to the conditional distribution $P(\omega|\vec{\phi})$ directly, using the intuitions developed in the previous section.

Returning to Equation 1, this basic framework provides a nice interpretation of the standard quality of match measures used in computer vision. Commonly used quality of match measures in computer vision are based on the number of model features that have a match in the image, or, similarly, the fraction of the boundaries or edges of a model that are accounted for in an image (see Baird, 1985, and Grimson, 1990, for references).

The following argument shows that this standard evaluation function corresponds to assuming that that the ϕ_j are mutually independent. For, if we assume independence, we can write:

$$P(\omega_i|\vec{\phi}) = P(\omega_i|\phi_1, \ldots, \phi_m) = \prod_j P(\omega_i|\phi_j) \tag{2}$$

If we take logarithms on both sides, we get:

$$\log P(\omega_i|\vec{\phi}) = \sum_j \log P(\omega_i|\phi_j) \tag{3}$$

If we let $P(\omega_i|\phi_j) = \text{const}$, as is often the case in object recognition systems, this means that the quality of match measure is proportional to the number of matching features:

$$\text{quality of match} \sim \#\ \text{matching features} \tag{4}$$

This is exactly the kind of quality measure that is used by most current recognition algorithms. Hence, we see that most existing recognition algorithms assume independence of the probabilities that the individual features of an object match.

In the more general case, if we weigh image features differently, the weights we should choose are simply given by the logarithms of the conditional probabilities $P(\omega_i|\phi_j)$. These are marginal distributions of the true class conditional distribution $P_0(\omega_i|\vec{\phi})$. The class conditional distribution in Equation 3 is then the maximum entropy distribution consistent with these marginals.

Higher-Order Models To improve the performance or recognition systems, we need to extract more information from the matches between the model and the image. This means essentially that we need to decrease the entropy of the distribution $P(\omega_i|\vec{\phi})$. Since the distribution in Equation 3 is the maximum entropy distribution consistent with the 1st order marginals, we can only reduce its entropy further by imposing higher-order marginal constraints.

A plausible representation of statistics with specific first and second order moments is the log-linear representation. This is a good representation to choose because the maximum entropy distribution for given first and second order moments can be represented this way (see Goldman, 1987, for a discussion of these representations).

In the log-linear representation, we express the conditional probabilities as follows:

$$\log P(\omega_i|\vec{\phi}) = \sum_j \alpha^i_j(\phi_j) + \sum_{j,k} \beta^i_{jk}(\phi_j, \phi_k) + \text{const} \tag{5}$$

Recall that the ϕ_j take on values in $\{0, 1\}$. Each α function can therefore be characterized by two numbers, and each β function by four numbers.

Now, it is nice to see that this second-order representation already captures some of the higher-order statistical constraints that we discussed informally in the previous section.

One of the observations was that if some feature matches, then a nearby feature is more likely to match as well. We can express this statistical property by choosing the β^i_{jk} as follows:

$$\beta^i_{jk} = \begin{cases} \text{const} & \text{if distance}(\phi^M_j, \phi^M_k) < \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Recall that ϕ^M_j denotes the geometric information (location, orientation) associated with the model feature j; and distance measures the distance of the two features.

A closely related statistical property was that for most scenes, model edges tend not to be fragmented very much in images. We can express this by considering a model edge as being composed of a small, fixed length edge segment (edgel) and treat each edgel as a separate feature ϕ_j. Then, we can compute the number of breaks in a model edge as:

$$\beta^i_{jk} = \begin{cases} \text{const} & \text{if } \phi_j \neq \phi_k \text{ and } \phi^M_j \text{ is adjacent to } \phi^M_k \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

4 Examples

In what follows, we will give several examples of how the statistical principles derived in the previous sections can be applied in detail and how they can help improve the reliability of a recognition algorithm on some sample images. The images in the examples are cartoon line drawings (Schulz, 1967, Schulz, 1986). The cartoons have been converted into line drawings using morphological operations and thinning. The features that were used consisted of uniformly spaced samples from the boundaries with their associated orientations (bounded length edge segments, edgels); that is, each cartoon image or model consists of a collection of points and associated orientations (usually of the order of 1000-2000).

The feature vector $\vec{\phi}$ contained one component for each of these edge segments. For a given transformation of the model (in this case, translation and scaling), ϕ_j was set to 1 if the transformed model feature fell to within 10 pixels of some image feature, and if the orientation of that corresponding image feature differed from that of the transformed model feature by no more than 18°.

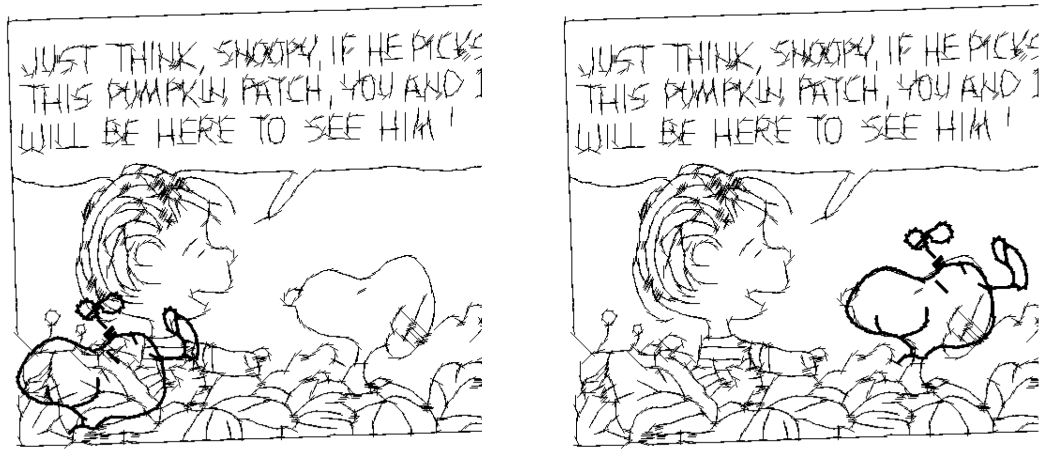
JUST THINK, SNOOPY, IF HE PICKS THIS PUMPKIN PATCH, YOU AND I WILL BE HERE TO SEE HIM!
JUST THINK, SNOOPY, IF HE PICKS THIS PUMPKIN PATCH, YOU AND I WILL BE HERE TO SEE HIM!

Figure 5: Improving the specificity of recognition using second-order statistics (left: bounded error recognition, right: using a second order evaluation function).

In each of the following examples, a higher-order statistical evaluation function was used. The match between a model and an image is defined as the optimal translation and scaling (in the range of $\frac{1}{\sqrt{2}}$ to $\sqrt{2}$ for the given evaluation function. Optimal translations were found by either brute-force search, or using the RAST algorithm (Breuel, 1992).

Second-Order Statistics Figure 5 shows the use of second-order statistics to express the statistical constraint that the presence or absence of nearby features in an image is correlated. That is, for the statistical evaluation function, we used Equation 5 with parameters chosen as in Equation 6. The actual values used in the experiments were $\alpha^i_j = 0$ and $\epsilon = 20$ pixels.

Edge Fragmentation A second method that was implemented is based on edge fragmentation. The evaluation function used is slightly different from the simple second order function shown in Equation 5. Specifically, the evaluation function computes the maximum number of features that can be matched under a given upper bound on the number of occlusions. In the example shown in Figure 6, the number of occlusions that was allowed was set to 0.

Opaqueness As we observed above, for opaque objects, there is another important constraint: portions of the object that are hypothesized to be unoccluded should not contain image features that are not explained by the

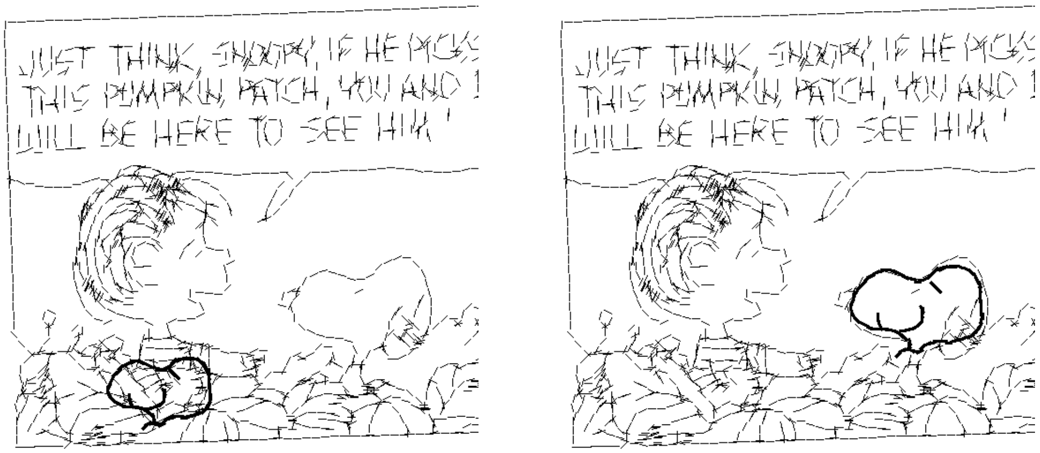
JUST THINK, SNOOPY, IF HE PICKS THIS PUMPKIN PATCH, YOU AND I WILL BE HERE TO SEE HIM!
JUST THINK, SNOOPY, IF HE PICKS THIS PUMPKIN PATCH, YOU AND I WILL BE HERE TO SEE HIM!

Figure 6: The use of bounds on fragmentation to improve the specificity of recognition (left: bounded error recognition, right: using a higher order evaluation function).

model. Furthermore, for the kinds of occlusions that we are assuming, such regions will tend to have relatively simply boundaries (for more details on the shape of unoccluded regions by random collections of convex objects, see Hall, 1988). An example of using this constraint is shown in Figure 7.

Reduction of Positive Mistakes The examples shown above have been qualitative—a small number of models were matched against a few images and some representative optimal matches were shown for different evaluation functions. However, it would be nice to see a more quantitative measure of how much the use of higher order evaluation functions improves the specificity of a recognition algorithm.

In order to do this, we have performed the following experiment. We used a collection of 200 cartoon images (similar to those used in the other experiments shown in this paper) and matched them against the model of an object (an annulus) that is known not to be contained in any of the images. Any match found in one of these images therefore constitutes a positive mistake of the recognition algorithm, i.e., indicating a match when actually none exists.

In order to declare a match, we need to set a threshold for the value returned by the evaluation function. We set this threshold by matching artificially generated, partially occluded pictures of an annulus and recording the value of the evaluation function, for each degree of occlusion.

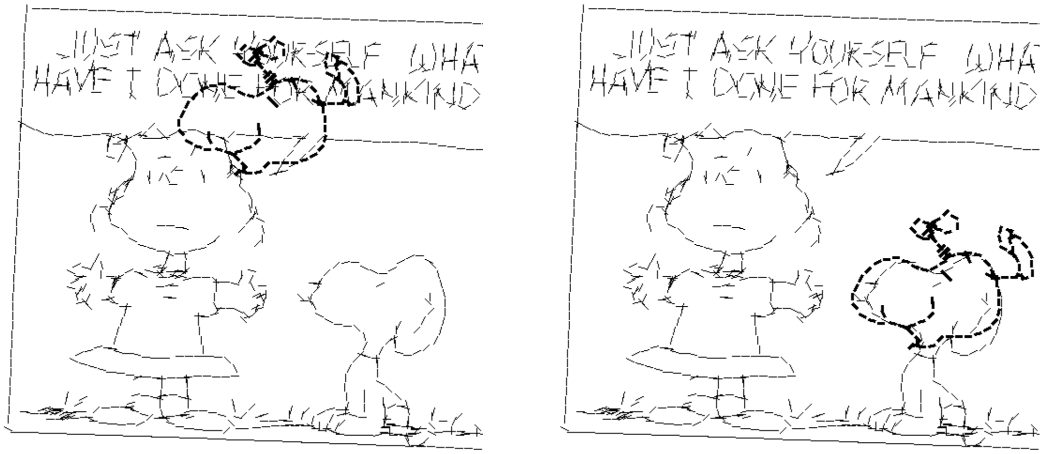

Figure 7: The use of exclusions to improve recognition (left: bounded error recognition, right: using a higher order evaluation function).

The higher-order evaluation function used in this experiment was the same described above under "Edge Fragmentation": the matching algorithm determined the largest match of the model against the image, assuming that it was occluded by at most a given number n of convex objects. The graphs in Figure 8 show the probability of positive mistakes for different n and different degrees of occlusions. The graph labeled "first-order" shows the performance of the standard first-order methods (bounded error recognition, Hough transform, etc.) for this recognition problem.

It is quite apparent from the graph that using higher-order statistical constraints can have a great effect on the specificity of a recognition algorithm. For example, at 40% occlusion, the higher-order method, allowing 1 occlusion, has a rate of positive mistakes of under 10%, while the usual first-order methods have a mistake rate of over 95%.

5 Grouping and Segmentation

Grouping and segmentation has been an active area of research in computer vision. Generally, the goal of a grouping or segmentation algorithm is to identify parts of an image that are likely to belong to the same object. In addition to its practical importance in reducing the complexity of the recognition problem, grouping and segmentation also is a psychophysically important and observable phenomenon (see, for example, Marr, 1982). Closely related to grouping is the extraction of complex features (e.g., vertices, curves).

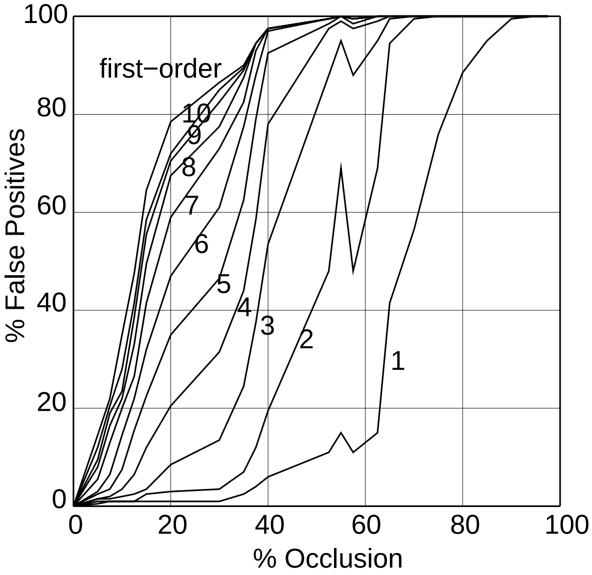

Figure 8: A comparison of the probability of positive mistakes for the usual first-order evaluation function with a family of higher-order evaluation function that take into account edge fragmentation. The small number next to each graph indicates the number of occlusions that is permitted to be present in a match.

The kind of constraints used for grouping are usually based (explicitly or implicitly) on the higher-order statistics of features. For example, Lowe, 1986, considers the probabilities that features are parallel or colinear in scenes of random line segments. Line segments in real images that show parallelism or colinearity that is unlikely to occur at random (i.e., that is non-accidental) are then argued to be like to have originated from the same object. Jacobs, 1988, uses similar constraints on small groups of line segments to argue that they bound a common region. Geman and Geman, 1984, use statistical techniques at a lower level to group pixels into edges.

Traditionally, grouping information has been used to speed up indexing and recognition. In fact, until the recent development of asymptotically recognition algorithms (Cass, 1990, Breuel, 1991) it was widely believed (Grimson, 1990) that such information was essential to achieving polynomial time recognition (usually referred to as "the correspondence problem of recognition").

Grouping and segmentation information allowed exponential time search algorithms to run in polynomial time by excluding most of the combinatorially many correspondences and permutations of model and image features from consideration by the matcher. But this also has a profound effect on recognition accuracy. Intuitively, if we choose incorrect correspondences or permutations of model features, the resulting pose estimate may be completely random, and the recognition algorithm will be performing a match of the model against a "random" image. If many of these "random matches" are carried out, then even an algorithm with a low rate of positive mistakes is bound to return spurious matches for some of the correspondences. By using grouping and segmentation information, we remove a large number of correspondences and permutations from consideration by the recognition algorithm, and the algorithm will carry out much fewer "random matches". Therefore, the overall probability of an accidental positive mistake is reduced. Lowe, 1986, makes the same observation.

We can view the use of grouping and segmentation information as a kind of higher order evaluation function. To see this, assume we have some grouping function $Q(\phi^I_j, \phi^I_k)$ that estimates the probability that two image features ϕ^I_j and ϕ^I_k come from a single object. In a conventional bottom-up grouping algorithm (e.g., Lowe, 1986, Jacobs, 1988), this probability is thresholded to obtain a set of candidates of features that come from the same object, and the resulting group is handed to a recognition algorithm (possibly with a more elaborate subsequent verification stage). More formally, a recognition algorithm that relies on grouping information only considers collections α of features that have a mutual probability higher than some threshold θ of coming from the same object in the image:

$$P(\omega_i|\vec{\phi}) = \max_{\alpha} \sum_j \alpha^i_j(\phi_j) \prod_{k \in \alpha} \left\lceil Q(\phi^I_j, \phi^I_k) > \theta \right\rceil \tag{8}$$

(The notation ⌈predicate⌉ means 1 if predicate is satisfied and 0 otherwise.)

Now, this is a higher-order evaluation function, but when we formulate it as such, it is quite clear that it takes insufficient advantage of the statistical information provided by the grouping function Q. Rather than thresholding Q, it would be better to use the statistical information contained in it directly to approximate the conditional probability $P(\omega_i|\vec{\phi})$. One way of seeing that this thresholding operation is suboptimal and potentially dangerous is that it represents an early commitment; this has been traditionally been viewed as as a problem with grouping algorithms: if the correct group is missed by the grouping algorithm, no subsequent processing recovers the lost information, and the reliability of the overall recognition system is limited by the reliability of the grouping step.

Because Q is not model specific, it can be derived with some confidence from first principles (cf. Lowe, 1986) or determined empirically from comparatively small numbers of images (model-specific parameters, on the other hand, require a significant number of images *per model* to estimate). Therefore, incorporating probabilistic grouping information like Q into higher-order evaluation functions directly is a promising way of obtaining additional robust and useful higher-order constraints. However, deriving a distribution for $P(\omega_i|\vec{\phi})$ from Q is mathematically non-trivial and will be left for a future paper.

Let us close this section with a speculation. The recent development of efficient recognition algorithms means that bottom-up grouping information is not required anymore to make the recognition computationally tractable. Furthermore, an incorporation of grouping information into the match of each model against the image makes the statistical information contained in a bottom-up grouping step preceding recognition redundant. Therefore, might it be that a visual/vision system simply does not require a grouping or segmentation "module" at all? The well-documented human ability for grouping (see, for example, Marr, 1982) would then be an epiphenomenon of recognition: two parts (features) of a scene are grouped together if they are jointly present in one or more matches of models against the image. More specifically, we would *define* the grouping function Q as an average over all known objects ω_i and all transformations T:

$$Q(\phi_a^I, \phi_b^I) = \int \sum_i P(\phi_a = 1 \wedge \phi_b = 1 \mid \omega_i, T)\ P(\omega_i)\ dT \tag{9}$$

In many implementations of a recognition system, this average value would simply be available (or easily computable) as a useful by-product of the recognition process. But it would be available only *after* recognition has taken place, rather than being a prerequisite for recognition.

6 Related Work

Statistical Models A number of other recognition systems have also been formulated in a Bayesian framework. However, the emphasis in most cases has been on error models for the location of individual features (Wells, 1992) or the likelihood that individual object parts or objects are present in particular views (Burns and Kitchen, 1988, Dickinson *et al.*, 1990, Mann and Binford, 1992).

Similarity Measures There has also been some work in trying to develop better measures of similarity among 2D shape. Such work has mostly concentrated on identifying "salient" features (e.g., Shashua and Ullman, 1988, Subirana and Richards, 1991). However, saliency is primarily a first-order constraint: different parts of an image or model simply are more or less important individually for the overall quality of match. Huttenlocher *et al.*, 1991, has approached the question of better 2D similarity measures from a geometrical point of view.

Other uses of MDL in Vision Despite its obvious utility for reasoning about visual object recognition, the Minimum Description Length principle has so far only been applied on low- and intermediate-level vision, primarily segmentation (Pen, 1990, Pentland, 1990, Darrell *et al.*, 1990, Dengler, 1991, Keeler, 1991). Marill, 1992, has also recently used an MDL principle to explain the non-model based disambiguation of line drawings.

Pattern Recognition and Neural Networks There is also a large body of relevant literature in pattern recognition and neural networks. In particular, a significant amount of research has been directed at devising methods that can learn or build "higher-order feature detectors" automatically; examples of this approach to vision are the work of Linsker, 1988, the TRAFFIC system (Zemel *et al.*, 1988), and, from the early days of neural network research, the Pandemonium model (Selfridge, 1959). While, ultimately, adaptive systems like neural networks are almost certainly needed, most neural network models are designed as "black boxes", without a detailed understanding of the combinatorial and statistical structure and constraints of the vision problem. Therefore, the approach described in this paper is complementary to such learning approaches: learning will ultimately be important for estimating the probability distributions, but understanding their structure better will help us choose better algorithms and network structures.

It should be noted that in the area of speech recognition, higher-order statistical models of speech and language are quite common.

7 Conclusions

This paper has demonstrated that higher-order statistical information is an important means for improving the specificity of a recognition algorithm.

We expect that the higher-order evaluation functions used in this paper, based on second-order statistic, edge fragmentation, and opaqueness, will be useful for achieving better performance on real recognition tasks. We have begun to investigate their use in optical character recognition in complex or degraded documents.

The view of recognition by higher-order evaluation functions presented here also provides a new framework for understanding existing techniques in computer vision, such as grouping, the extraction of complex features, and non-accidentalness. We have seen that such methods not only affect the efficiency of a recognition algorithm, but also have a profound (and often desirable) effect on its output. Understanding such effects in a statistical framework will hopefully allow us to make better tradeoffs between speed and correctness.

References

Baird H. S., 1985, *Model-Based Image Matching Using Location*, MIT Press, Cambridge, MA.

Berger J. O., 1980, *Statistical Decision Theory and Bayesian Analysis*, Springer Verlag.

Breuel T. M., 1991, An Efficient Correspondence Based Algorithm for 2D and 3D Model Based Recognition, In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition.*

Breuel T. M., 1992, Fast Recognition using Adaptive Subdivisions of Transformation Space, In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition.*

Burns J., Kitchen L., 1988, Rapid recognition out of a large model base using prediction hierarchies and machine parallelism , In *Intelligent Robots and Computer Vision. Sixth in a Series*, volume vol.848, pages 225–33.

Cass T. A., 1990, Feature Matching for Object Localization in the Presence of Uncertainty, A.I. Memo No. 1133, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Darrell T., Pentland A., Sclaroff S., 1990, Segmentation by minimal-length encoding, In *Proceedings of the International Conference on Computer Vision.*

Dengler J., 1991, Estimation of Discontinuous Displacement Vector Fields with the Minimum Description Length Criterion, In *Proceedings: Computer Vision and Pattern Recognition*, IEEE Computer Society Press.

Dickinson S. J., Pentland A. P., Rosenfeld A., 1990, Qualitative 3d shape reconstruction using distributed aspect graph matching, In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition.*

Duda R. O., Hart P. E., 1973, *Pattern Classification and Scene Analysis*, Wiley, New York.

Geman S., Geman D., 1984, Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6:721–741.

Goldman S. A., 1987, Efficient methods for calculating maximum entropy distributions, Technical Report MIT/LCS/TR-391, MIT Laboratory for Computer Science, Cambridge, MA, USA.

Grimson E., 1990, *Object Recognition by Computer*, MIT Press, Cambridge, MA.

Hall P., 1988, *Introduction to the Theory of Coverage Processes*, John Wiley & Sons, New York, NY.

Huttenlocher D. P., Kedem K., Sharir M., 1991, The Upper Envelope of Voronoi Surfaces and its Applications, In *Proceedings of the Seventh ACM Symposium on Computational Geometry.*

Jacobs D. W., 1988, The Use of Grouping in Visual Object Recognition, A.I. Technical Report No. 1023, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Keeler K., 1991, Map representations and coding-based priors for segmentation, In *Proceedings 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (91CH2983-5)*, pages 420–5.

Kiefer J. C., 1987, *Introduction to Statistical Inference*, Springer Verlag.

Linsker R., 1988, Self-organization in a perceptual network, *IEEE Computer*, 21:105–117.

Lowe D. G., 1986, *Perceptual Organization and Visual Recognition*, Kluwer Academic Publishers, Boston.

Mann W. B., Binford T. O., 1992, An example of 3d interpretation of images using bayesian networks, In *Proceedings Image Understanding Workshop.*

Marill T., 1992, Why Do We See Three-Dimensional Objects?, Technical Report AI–Memo 1366, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Marr D., 1982, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*, W.H. Freeman and Company, San Francisco.

1990, Part segmentation for object recognition, *Neural Computation*, 1:82–91.

Pentland A., 1990, Automatic extraction of deformable part models, *International Journal of Computer Vision*, 4:107–126.

Rissanen J., 1978, Modeling by Shortest Data Description, *Automatica*, 14:465–471.

Schulz C. M., 1967, *You're Something Else, Charlie Brown*, Holt, Rinehart, and Winston, Inc.

Schulz C. M., 1986, *The Way of the Fussbudget is not Easy.*

Selfridge O. G., 1959, Pandemonium: A paradigm for learning., In *The Mechanisation of Thought Processes*, London: H.M. Stationary Office.

Shashua A., Ullman S., 1988, Structural saliency: the detection of globally salient structures using a locally connected network, In *Proceedings of the International Conference on Computer Vision*, pages 321–327, Tarpon Springs, FL, IEEE, Washington, DC.

Stoyan D., Kendall W. S., Mecke J., 1987, *Stochastic Geometry and Its Applications*, John Wiley and Sons.

Subirana B., Richards W., 1991, Perceptual Organization, Figure-Ground, Attention and Saliency, A.I. Memo No. 1218, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Wells W. M., 1992, Posterior marginal pose estimation, In *Proceedings Image Understanding Workshop*, Morgan Kaufmann, San Mateo, CA.

Zemel R. S., Mozer M. C., Hinton G. E., 1988, Traffic: A model of object recognition based on transformation of feature instances, Technical Report CRG–TR–88–7, Department of Computer Science, U. of Toronto.