TECHNICAL REPORT

A New Bayesian Framework for Object Recognition

Yuri Boykov, Daniel Huttenlocher
Computer Science Department
Cornell University
Ithaca, NY 14853, USA
{yura,dph}@cs.cornell.edu

October 28, 1998

Abstract

We describe a new approach to feature-based object recognition, using maximum a posteriori (MAP) estimation under a Markov random field (MRF) model. The main advantage of this approach is that it allows explicit modeling of dependencies between individual features of an object. For instance, we use the approach to model the fact that mismatched features due to partial occlusions tend to form spatially coherent groups rather than being independent. Efficient computation of the MAP estimate in our framework can be accomplished by finding a minimum cut on an appropriately defined graph. An even more efficient approximation, that does not use graph cuts, is also presented. This approximation technique, which we call spatially coherent matching (SCM), is closely related to generalized Hausdorff matching. We report some Monte Carlo experiments showing that the SCM technique improves substantially on the tradeoff between correct detection and false alarms compared with previous feature matching methods such as the Hausdorff distance.

1 Introduction

In this paper we present a new Bayesian approach to object recognition using Markov random fields (MRF's). As with many approaches to recognition we assume that an object is modeled as a set of features. The recognition task is then to determine whether there is a match between some subset of these object features and features extracted from an observed image. The central idea underlying our approach is to explicitly capture dependencies between individual features of an object. Markov random fields provide a good theoretical framework for representing such dependencies between features. Recent algorithmic developments make it quite practical to compute the maximum a posteriori (MAP) estimate for the MRF model that we employ (e.g., [2], [5]). Moreover, in several cases of practical interest, we present even faster approximation methods that do not require graph algorithms.

Our approach contrasts with most feature-based object recognition techniques, as they do not explicitly account for dependencies between features of the object. It is desirable to be able to account for such dependencies, because they occur in real imaging situations. For example, a common case occurs with partial occlusion of objects, where features that are near one another in the image are likely to be occluded together. In our model, we assume that the process of matching individual object features is described *a priori* by a Gibbs distribution[1] associated with a certain Markov random field. This model captures pairwise dependencies between features of the object. We then use *maximum a posteriori* (MAP) estimation to find a match between the object and the scene or to show that there is no such match. While a number of probabilistic approaches to recognition have been reported in the literature (e.g., [10], [9],[12]) these methods do not provide an explicit model of dependencies between features.

We show that finding the best match using the Hausdorff fraction [6], [11] can be viewed as a special case of our technique, where the dependencies between all pairs of features in the object are equally strong. Therefore, our Bayesian framework can be seen as providing a probabilistic understanding of generalized Hausdorff matching. With this view of Hausdorff matching, it becomes apparent that one of the main limitations of the Hausdorff approach is its failure to take into account the spatial coherence of matches between neighboring features. That is, the Hausdorff approach does not account for the fact that in a local neighborhood there tends to be a higher correlation between features. We thus suggest a closely related method, which we call *spatially coherent matching* (SCM). This method requires that matching features be more than some critical distance from features that do not match, thus ensuring spatially contiguous sets of matching features. We present some Monte Carlo experiments demonstrating that the SCM approach is a substantial improvement over Hausdorff and other previous matching techniques, in cases where the image is cluttered with many irrelevant features and there is substantial occlusion of the object to be recognized.

In the following section we present the MRF framework for recognition and the resulting MAP estimation problem. Then in Section 3 we briefly discuss how recent results on graph algorithms for solving MRF's provide an efficient solution to this estimation problem. In Section 4 we show that the generalized Hausdorff matching problem can be seen to be a special case of this MAP-MRF framework, and then introduce the spatially coherent matching approach to improve upon Hausdorff matching. Finally in Section 5 we present some Monte Carlo experiments comparing the new SCM approach with Hausdorff matching.

2 The MAP-MRF Recognition Framework

In this section we describe our object matching framework in more detail. We represent an object by a set of features, indexed by integers in the set $M = \{1, 2, \ldots, m\}$. Each feature corresponds to some vector M_i in a feature space of the object. Commonly the vectors M_i will simply specify a feature location (x, y) in a fixed coordinate system of the object,

[1]In Section 2.1 we briefly discuss the Gibbs distribution. See [8] for more details.

although more complex feature spaces fit within the framework.

A given image I is a set of observed features from some underlying true scene. Each feature $i \in I$ corresponds to a vector I_i in a feature space of the image. The true scene can be thought of as some unknown set of features I^{tr} in the same feature space. Similarly, I_i^{tr} is a vector describing the feature $i \in I^{tr}$ in the feature space of the image. We are interested in finding a match between the object features M and the true scene features I^{tr}, using the observed image features I.

A match of the object M to the true scene I^{tr} is described by a pair $\{S, L\}$ where $S = \{S_1, S_2, \ldots, S_m\}$ is a collection of boolean variables and L is a location parameter. If $S_i = 1$ then the ith feature of the object has a matching feature in I^{tr} and if $S_i = 0$ then it does not. In the latter case we say the feature is mismatched. For example, the event

$$\{S_1 = \ldots = S_k = 1,\ S_{k+1} = \ldots = S_m = 0,\ L = l\}$$

implies that for $1 \leq i \leq k$, feature i of M has a matching feature $j \in I^{tr}$, such that $I_j^{tr} = M_i \oplus L$. Moreover, the last $(m-k)$ features are mismatched, meaning they have no such matching features. The operation $\oplus$ depends on the type of mapping from the object to the image feature space, which varies for the particular recognition task. In this paper we will use translation (vector summation), but other transformations are possible.

To determine the values of $\{S, L\}$ we use the maximum a posteriori (MAP) estimate

$$\{S^*, L^*\} \quad = \quad \arg\max_{S,L} \ \Pr(S, L|I).$$

Bayes rule then implies

$$\{S^*, L^*\} \quad = \quad \arg\max_{S,L} \ \Pr(I|S, L) \cdot \Pr(S) \cdot \Pr(L), \tag{1}$$

assuming that the distributions of S and L are a priori independent. The prior distributions $\Pr(S)$ and $\Pr(L)$ are discussed in Section 2.1. We assume that the prior distribution of S is described by a certain Markov random field, thus allowing for spatial dependencies among the S_i. The likelihood function $\Pr(I|S, L)$ is then discussed in Section 2.2.

Let $\mathcal{L}$ denote a set of possible locations of the object in the true scene. Then the range of the location parameter L is $\mathcal{L} \cup \emptyset$ where the extra value $\emptyset$ implies that the object is not in the scene. The basic idea of our recognition framework is to report a match between the object and the observed scene if and only if

$$S^* \neq \bar{0} \quad \text{and} \quad L^* \neq \emptyset. \tag{2}$$

In Section 2.3 we develop the test in (2) for the model specified in Sections 2.1 and 2.2.

2.1 Prior Knowledge

We assume that the prior distribution of the location parameter L can be described as

$$\Pr(L) \quad = \quad (1-\rho) \cdot f(L) \quad + \quad \rho \cdot \delta(L = \emptyset) \tag{3}$$

where $f(L) = \Pr(L|L \in \mathcal{L})$, the parameter ρ is the prior probability that the object is not present in the scene, and $\delta(\cdot)$ equals 1 or 0 depending on whether condition "$\cdot$" is true or false. Generally the distribution function $f(L)$ is uniform over $\mathcal{L}$. However in some applications $f(L)$ can reflect additional information about the object's location. For example, such information might be available in object tracking since the current location of the object can be estimated from its previous location. The value of the constant ρ may be anywhere in the range $[0, 1)$. In Section 2.3 we will see that ρ appears in our recognition technique only as a threshold for deciding whether or not the object is present given the image.

We assume that the collection of boolean variables, S, indicating the presence or absence of each feature, forms a Markov random field independent of L. More specifically, the prior distribution of S is described by the Gibbs distribution

$$\Pr\{S\} \quad \propto \quad \exp\left\{-\sum_{i\in M} \alpha\cdot(1-S_i) \quad - \quad \sum_{\{i,j\}} \beta_{\{i,j\}}\cdot\delta(S_i \neq S_j)\right\} \tag{4}$$

where the second summation is over all distinct unordered pairs of features of the object.

This model captures the probability that features will not be matched even though they are present in the true scene, given some fixed location, L. Such non-matches could be due to occlusion, feature extraction error, or other causes. The parameter $\alpha \geq 0$ is a penalty for such non-matching features. The coefficient $\beta_{\{i,j\}} \geq 0$ specifies a strength of interaction between features i and j of the object. For tractability, this model captures only pairwise interaction between features. Nevertheless, the pairwise interaction model provided by this form of Gibbs distribution is rich enough to capture one important intuitive property: a priori it is less likely that a feature will be un-matched if other features of the object have a match. Note that if all $\beta_{\{i,j\}} = 0$ then there is no interaction between the features and the S_i's become independent Bernoulli variables with probability of success $\Pr(S_i = 1) = e^{\alpha}/(1+e^{\alpha}) \;\geq\; 0.5$.

2.2 Likelihood Function

The features of the observed image I may appear differently from the features of the unknown true scene I^{tr} due to a number of factors. This includes sensor noise, errors of feature extraction algorithms (e.g. edge detection), and others. It is the purpose of the likelihood function to describe these differences in probabilistic terms.

We use the likelihood function

$$\Pr(I|S, L) \quad \propto \quad \prod_{i\in M} g_i(I|S_i, L) \tag{5}$$

where $g_i(\cdot)$ is a likelihood function corresponding to the ith feature of the object. If $S_i = 0$ or $L = \emptyset$ then $g_i(I|S_i, L)$ is the likelihood of I given that the true scene does not contain the ith feature of the object. We assume that all cases of a mismatching feature have the same likelihood

$$g_i(I|1,\emptyset) = g_i(I|0,\emptyset) = g_i(I|0,L) = C_0 \qquad\qquad \forall i \in M, \quad \forall L \in \mathcal{L} \tag{6}$$

where C_0 is a positive constant.

If $L \in \mathcal{L}$ then $g_i(I|1,L)$ is the likelihood of observing image I given that the i-th feature of the object is at location $(L \oplus M_i)$ in the feature space of the true scene I^{tr}. The choice of $g_i(I|1,L)$ for $L \in \mathcal{L}$ will depend on the particular application.

Example 1 (Recognition based on edges) Consider an edge-based object matching problem, where all features of the object are edge pixels. We observe a set of image features I obtained by an intensity edge detection algorithm. One reasonable choice of $g_i(I|1,L)$ for $L \in \mathcal{L}$ is

$$g_i(I|1,L) \;=\; C_1 \cdot g(d_I(L \oplus M_i)) \tag{7}$$

where $d_I(\cdot)$ is a distance transform of the image features I. That is, the value of $d_I(p)$ is the distance from p to the nearest feature in I. The function $g(\cdot)$ is some probability distribution that is a function of the distance to the nearest feature. Normally, g is a distribution concentrated around zero. The underlying intuition is that if the true scene I^{tr} has an edge feature located at $(L \oplus M_i)$ then the observed image I should contain an edge nearby. Thus the distance transform $d_I(L \oplus M_i)$ will be small with large probability. A number of existing recognition schemes use functions of this form, including Hausdorff matching [6] and Chamfer matching [1]. □

2.3 MAP Estimation

By substituting (3), (4), (5) into (1) and then taking the negative logarithm of the obtained equation we can show that MAP estimates $\{S^*, L^*\}$ minimize the value of the posterior energy function

$$E(S,L) = \begin{cases} H_L(S) - \ln f(L) & - \;\; \ln(1-\rho) & \text{if } L \in \mathcal{L} \\ H_L(S) & - \;\; \ln \rho & \text{if } L = \emptyset \end{cases}$$

where

$$H_L(S) \;=\; \sum_{\{i,j\}} \beta_{\{i,j\}} \cdot \delta(S_i \neq S_j) \;+\; \sum_{i \in M} \left(\alpha \cdot (1 - S_i) - \ln g_i(I|S_i, L)\right). \tag{8}$$

Our goal is to find $\{S^*, L^*\}$. The main technical difficulty is to determine $\{\hat{S}, \hat{L}\}$ that minimize $H_L(S) - \ln f(L)$ for $L \in \mathcal{L}$. In Section 3 we show how this can be done in the most general case. For the moment simply assume that $\{\hat{S}, \hat{L}\}$ are given.

Consider $H_L(S)$ for $L = \emptyset$. Equation (6) implies that $H_\emptyset(S)$ is minimized by the configuration $S = \bar{1}$ where all $S_i = 1$. If $E(\hat{S}, \hat{L}) > E(\bar{1}, \emptyset)$ then $\{S^*, L^*\} = \{\bar{1}, \emptyset\}$. According to (2), in this case we report that the object is not recognized in the scene. If $E(\hat{S}, \hat{L}) \leq E(\bar{1}, \emptyset)$ then $\{S^*, L^*\} = \{\hat{S}, \hat{L}\}$. In this case $L^* \in \mathcal{L}$. Nevertheless, if $\hat{S} = \bar{0}$ we would still report the absence of the object in the scene.

Finally, our recognition framework can be summarized as follows. The match between the object and the observed scene is reported if and only if $\hat{S} \neq \bar{0}$ and

$$H_{\hat{L}}(\hat{S}) - \ln f(\hat{L}) \;\leq\; m \cdot \ln \frac{1}{C_0} + \ln \frac{1-\rho}{\rho} \tag{9}$$

where (9) is derived from $E(\hat{S}, \hat{L}) \leq E(\bar{1}, \emptyset)$. The right hand side of (9) is a constant that represents a certain decision threshold. Note that this decision threshold depends on two things: first, the prior probability of occlusion, ρ; and second, the product of the number of features of the object, m, with the log-likelihood of a mismatch, C_0.

3 Energy Minimization

In our framework, the recognition problem is formulated as finding a pair $\{\hat{S}, \hat{L}\}$ that minimizes $H_L(S) - \ln f(L)$ for $L \in \mathcal{L}$. In this section we briefly explain how to perform this minimization in the most general case. In Section 4 we consider some special cases where no sophisticated algorithmic scheme is needed to obtain $\{\hat{S}, \hat{L}\}$.

In the simplest formulation, this minimization problem can be solved in two steps. The first step is to find an S_L that minimizes $H_L(S)$ for each fixed value of $L \in \mathcal{L}$, and the second step is to find an L in $\mathcal{L}$ that gives the smallest value for $H_L(S_L) - \ln f(L)$. More sophisticated schemes can be envisioned that use properties of $H_L(S)$ to avoid computing its minimum for each possible value of L. For example, the techniques for pruning the search space in Hausdorff matching [6, 11] can be applied in the special cases considered in Sections 4.1 and 4.3. For the moment we simply consider checking all possible values of L.

The first step, of minimizing $H_L(S)$, appears computationally quite difficult as the number of possible configurations of S is 2^m. This number is astronomical in practice since we consider objects where the number of features $m > 100$. Fortunately, the graph cut methods developed in [5] and [2] can be applied to minimize $H_L(S)$ exactly and efficiently. We now briefly explain their technique in the context of our problem.

To minimize $H_L(S)$ we need to find an optimal assignment of labels 1 and 0 to the variables S_i for $i \in M$. Consider a graph $G = \langle V, E\rangle$ where V is a set of vertices and E is a set of edges. The set V consists of m vertices indexed by the names of the variables S_i and two terminal vertices indexed by the integers 1 and 0. The set of edges E consist of n-links connecting pairs of S_i vertices with each other and t-links connecting S_i vertices with one of the terminals. The structure of the graph is shown in Figure 1.

A cut C of the graph G is a subset of edges E such that the terminals 1 and 0 are completely separated on the induced graph $G(C) = \langle V, E - C\rangle$. Note that a cut C of G corresponds to a certain assignment of labels. If the vertex S_i is connected to the terminal 1 on the induced graph $G(C)$ then $S_i = 1$ and if S_i is connected to the terminal 0 then $S_i = 0$.

We need to specify a cutting cost for all edges in E. The n-link connecting the vertices S_i and S_j costs $\beta_{\{i,j\}}$. Therefore, the stronger the interaction between features i and j according to the prior distribution in (4) the costlier it is to assign them different labels, that is, to cut the edge between them.

The t-link between S_i and the terminal 1 costs $(\alpha - \ln C_0)$ and the t-link between S_i and the terminal 0 costs $-\ln g_i(I|1, L)$. According to (6), these edge cutting costs represent the penalties in the second summation in (8) when S_i is assigned 0 or 1, correspondingly. Note that if the likelihood $g_i(I|1, L)$ is large then the cutting cost of the t-link between S_i and the terminal 0 becomes small. Intuitively speaking, this encourages assigning label 1 to S_i. Note also that C_0 and g_i have values between zero and one and, therefore, the weights

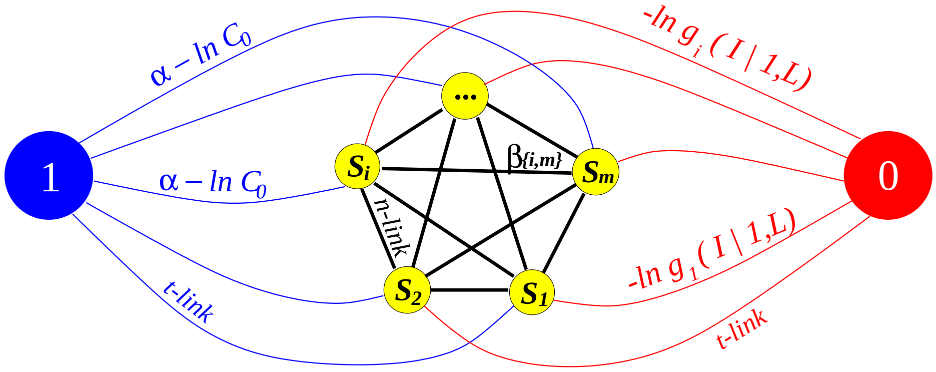


Figure 1: The graph $G = \langle V, E \rangle$.

of t-links are non-negative.

According to [5] and [2], the minimum cost cut on G gives an assignment to variables S_i that minimizes the value of $H_L(S)$ for a given L. The min cut problem is well studied in combinatorial optimization. It can be solved exactly by the max-flow algorithm of Ford and Fulkerson [3] or by the push-relabel algorithm of Goldberg and Tarjan [4]. The theoretical running time is polynomial. In many practical problems its running time is close to linear.

4 Discussion of special cases

In this section we identify some interesting properties of our recognition framework by considering several special cases. We concentrate on the problem of finding an optimal match configuration S_L that minimizes $H_L(S)$ in (8) at a fixed location $L \in \mathcal{L}$. For the examples below this problem can be solved without the graph cut technique of Section 3.

In Section 4.1 we show that Hausdorff matching is a special case of our framework. This provides a probabilistic model for Hausdorff matching and suggests how Hausdorff matching can be improved upon by taking into account spatial dependencies among the features. We then consider models that account for spatial dependencies in Section 4.2, by imposing some neighborhood system over the features. These neighborhood models yield a simple technique which we call *spatially coherent matching* (SCM). This new technique is discussed in Section 4.3, and is a natural generalization of Hausdorff matching.

4.1 Hausdorff Matching

In this section we show that Hausdorff matching is a special case of our framework where the strength of interaction between features of the object is uniform, that is, $\beta_{\{i,j\}} = \beta$ for all $\{i, j\}$ where β is a non-negative constant. The classical Hausdorff distance is a max-min measure for comparing two sets for which there is some underlying distance function on pairs of elements, one from each set. The application of Hausdorff matching in computer vision has used a generalization of this classical measure [6], that computes a distance quantile rather than the maximum distance.

One form of the generalized Hausdorff measure is based on counting the number of features of the object that are within some distance r of the nearest image feature. Let $M_L = \{i \in M : d_I(L \oplus M_i) \leq r\}$ denote the subset of features of the object that are within distance r of features of the image, when the object is positioned at location L. We call M_L a set of *matchable* features for a given location L. Then the Hausdorff approach matches the object at L if and only if $|M_L| > Const$ where $|\cdot|$ denotes the number of elements in the set. The constant usually represents a critical fraction of the total number of object features, m. Thus this measure is often referred to as the *Hausdorff fraction*.

In order to describe this measure using our framework we assume that $g_i(I|1, L) = C_1 \cdot g(d_I(L \oplus M_i))$, as in Example 1. Moreover, we use the particular function

$$g(d) = \begin{cases} \frac{1}{r} & \text{if } d \leq r \\ 0 & \text{if } d > r \end{cases} \tag{10}$$

where r is the distance to the nearest image feature used in Hausdorff matching and in the definition of the set M_L.

We then need the following notation. Any configuration S is uniquely defined by a collection of integers $1_S = \{i \in M : S_i = 1\}$ which is the subset of features of the object assigned a match by S. Consider also $0_S = \{i \in M : S_i = 0\}$. Note that for any configuration S we have $1_S \cup 0_S = M$ and $1_S \cap 0_S = \emptyset$. Therefore, $m = |1_S| + |0_S|$.

Our approach is based on minimizing the function $H_L(S)$ in (8) for a fixed location $L \in \mathcal{L}$. Equation (10) implies that if $d_I(L \oplus M_i) > r$ then $g_i(I|1, L) = 0$. This means that the likelihood of a match for a feature $i \in M$ is zero if the image I does not contain any features near $L \oplus M_i$. Thus, one cannot assign $S_i = 1$ if the ith feature of the object is such that $d_I(L \oplus M_i) > r$, and we must have $1_S \subseteq M_L$. Formally speaking, it is easy to check that $1_S \not\subseteq M_L$ implies $H_L(S) = \infty$. If $1_S \subseteq M_L$ then the second summation in (8) can be rewritten as $|0_S| \cdot (\alpha - \ln C_0) - |1_S| \cdot \ln \frac{C_1}{r}$.

The assumption that $\beta_{\{i,j\}} = \beta$ for all $\{i, j\}$ simplifies the first term of $H_L(S)$ in (8) to $\beta \cdot |1_S| \cdot |0_S|$. Since $|0_S| = m - |1_S|$, $H_L(S)$ can be rewritten as a function of a single scalar

$$H_L(S) = \begin{cases} h(|1_S|) & \text{if } 1_S \subseteq M_L \\ \infty & \text{if } 1_S \not\subseteq M_L \end{cases} \tag{11}$$

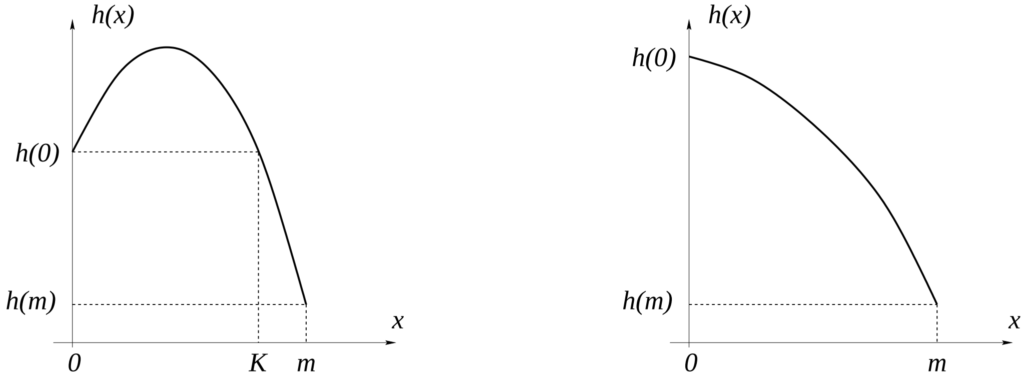
where

$$h(x) = \beta \cdot x \cdot (m - x) \; - \; x \cdot (\alpha + \ln \frac{C_1}{rC_0}) \; + \; m \cdot (\alpha - \ln C_0) \tag{12}$$

is a concave down parabola shown in Figure 2.

Now we can show how to find a configuration S_L that minimizes $H_L(S)$ in (11) for a fixed L. Equation (11) implies that $0 \leq |1_S| \leq |M_L|$. Thus $h(|1_S|)$ is minimized by either $|1_S| = 0$ or $|1_S| = |M_L|$. It is straightforward to check that $h(|M_L|) < h(0)$ if and only if $|M_L| > K$ where

$$K = m - \left(\frac{\alpha + \ln \frac{C_1}{rC_0}}{\beta} \right). \tag{13}$$



a) Strong interaction ($\beta > \beta_o$). b) Weak interaction ($0 \leq \beta \leq \beta_o$).

Figure 2: Two typical cases of $h(x)$ for $\beta > \beta_o$ and $0 \leq \beta \leq \beta_o$ where $\beta_o = (\alpha + \ln \frac{C_1}{rC_0})/m$.

Consequently, $S_L \neq \bar{0}$ if and only if $|M_L| > K$ which is exactly the Hausdorff test described above. That is, using the definition of g_i in Example 1 and g in (10) our framework computes the Hausdorff matching using the Hausdorff fraction.

We close this section by noting that there are two qualitatively different cases depending on whether K is positive or negative. Consider the threshold value β_o defined in Figure 2. If $\beta \leq \beta_o$ then $K \leq 0$. In this case, the inequality $|M_L| > K$ is always true and the configuration S_L is always given by $1_{S_L} = M_L$. Intuitively speaking, the dependencies between the features of the object are so weak for $\beta \leq \beta_o$ that each feature i is matched at L as long as there is some image feature within the distance r from $L \oplus M_i$. This can be described as an *independent* matching of features.

If $\beta > \beta_o$ then $K > 0$. In this case $|M_L|$ can be either greater or smaller than K depending on L and on the observed image. If $|M_L| \leq K$ then $S_L = \bar{0}$ and if $|M_L| > K$ then S_L is given by $1_{S_L} = M_L$. Recall that $|M_L|$ is the number of features of the object that are within distance r of the closest image feature, when the object is positioned at L. Intuitively speaking, if the dependencies between the features of the object are strong enough then they are matched to the image at L if and only if the image fits a sufficiently large group of these features at a given location. This can be described as a dependent matching of features.

4.2 Models with a Local Neighborhood System

In this section we consider another example where the optimal configuration S_L minimizing $H_L(S)$ can be obtained without the general graph-cut technique of Section 3. Having seen that the generalized Hausdorff measure can be viewed in our framework as having equal weights between all pairs of features, we now consider models where features of the object have higher weights connecting them to features within some local neighborhood. This model captures the fact that nearby features of an object will tend to be matched or mismatched together.

We denote by $\mathcal{N}_M$ the set of all pairs of neighboring features for a given object M. We assume that $\beta_{\{i,j\}} = \beta + \beta_N$ if the features $\{i,j\} \in \mathcal{N}_M$ are neighbors and $\beta_{\{i,j\}} = \beta$ if the features $\{i,j\} \notin \mathcal{N}_M$ are not neighbors. The coefficients β and β_N are some nonnegative constants. It is reasonable to expect that two neighboring features are more likely to have the same label than a pair of features isolated from each other.

We assume that the likelihood g_i is defined the same way as in Section 4.1. Then equation (8) can be written as

$$H_L(S) \;=\; \begin{cases} \beta_N \cdot b(S) + h(|1_S|) & \text{if } 1_S \subseteq M_L \\ \infty & \text{if } 1_S \not\subseteq M_L \end{cases} \tag{14}$$

where $b(S) = |\{i,j\} \in \mathcal{N}_M : S_i \neq S_j|$ denotes the number of pairs of neighboring features assigned opposite labels by the configuration S. For simplicity, we will refer to $b(S)$ as the number of N-discontinuities in the configuration S. The rest of notation is borrowed from Section 4.1.

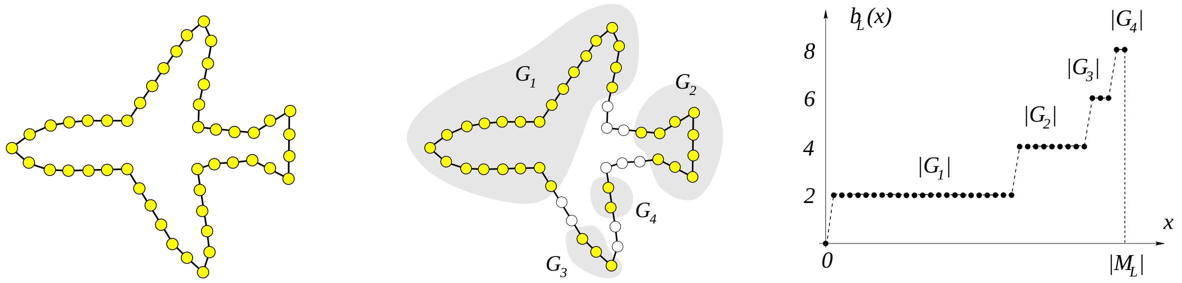
We would like to show how to find a configuration S_L that minimizes $H_L(S)$ for a fixed location L. Note that among all configurations S with a fixed size $|1_S| = x$ and such that $1_S \subseteq M_L$ there exists some configuration that has the smallest number of N-discontinuities. Let

$$b_L(x) \;=\; \min\left\{ b(S) \;:\; |1_S| = x,\; 1_S \subseteq M_L \right\} \qquad \text{for } 0 \leq x \leq |M_L|$$

denote the corresponding minimal number of N-discontinuities for a given location L and size x. In general, $b_L(0) = 0$ and $b_L(|M_L|)$ is a number of N-discontinuities in the configuration S given by $1_S = M_L$. For $0 < x < |M_L|$ the exact value of $b_L(x)$ can be derived analytically only in some simple cases.

For example, consider a model with a "chain" neighborhood system, as illustrated in Figure 3(a). The nontrivial case is when $|M_L| < m$. Let $\mathcal{G}_L = \{G_1, G_2, \ldots, G_{n_L}\}$ denote the set of all connected components (or groups) in a given M_L. Each group $G \in \mathcal{G}_L$ is a subset of features in M_L connected under the neighborhood system $\mathcal{N}_M$. Consider the example in Figure 3(b). The features in M_L are highlighted by shading. In this case $\mathcal{G}_L = \{G_1, \ldots, G_4\}$. In general, we can assume that the groups are indexed according to their size so that $|G_1| \geq |G_2| \geq \ldots \geq |G_{n_L}|$. Then the configuration S that has the smallest number of N-discontinuities given that $1_S \subseteq M_L$ and $|1_S| = x$ can be obtained as follows. The key idea is to assign x matches ($S_i = 1$) to features of M_L so that the matched features form the smallest number of connected components possible. Thus, the matches should fill out the largest groups of M_L. Assume that x satisfies $|G_1| + \ldots + |G_{k-1}| < x \leq |G_1| + \ldots + |G_k|$. Then all features in the groups G_1 through G_{k-1} should be assigned a match while the remaining matches can be assigned to connected features in G_k. The number of N-discontinuities in the obtained configuration S equals $2k$. Clearly, it is the smallest number of N-discontinuities for a given x. Therefore, if $|M_L| < m$ then $b_L(x) = 2k$ for $\sum_{i=1}^{k-1} |G_i| \;<\; x \;\leq\; \sum_{i=1}^{k} |G_i|$. In Figure 3(c) we show the plot of $b_L(x)$ for the example of M_L in part (b) of the same figure.

Minimizing $H_L(S)$ in (14) is equivalent to minimizing $\beta_N \cdot b_L(x) + h(x)$ for $0 \leq x \leq |M_L|$. If the function $b_L(x)$ is known then the exact minimum can be obtained. Similarly to Section 4.1, minimization of $H_L(S)$ is reduced to a one dimensional problem in this case.



a) The neighborhood system $\mathcal{N}_M$ is represented by edges in the feature space of the object.

b) $M_L = G_1 \cup G_2 \cup G_3 \cup G_4$. The sizes of the groups are $|G_1| = 23$, $|G_2| = 9$, $|G_3| = 3$, and $|G_4| = 2$.

c) The plot of $b_L(x)$ corresponding to M_L in (b).

Figure 3: Example of a model with a "chain" neighborhood system.

4.3 Spatially Coherent Matching

The solution in Section 4.2 is primarily of theoretical interest because the closed form of function $b_L(x)$ can be obtained only in a limited number of cases. In this section we introduce another matching method that captures the stronger dependencies between features in a local neighborhood. We call this method spatially coherent matching (SCM) because it takes into account the fact that feature mismatches generally occur in coherent groups (e.g., due to partial occlusion of an object). The SCM method can be implemented as an extension to the Hausdorff matching computation discussed in Section 4.1. We also show that that the SCM approach is a good approximation to the neighborhood scheme presented in the previous section.

First we describe the SCM technique. We use some of the notation from Section 4.1. Recall that $M_L = \{i \in M : d_I(L \oplus M_i) \leq r\}$ is the subset of object features lying within distance r of image features, when the object is positioned at location L. As we saw before, we can think of M_L as a set of *matchable* features of the object for a given location L, because they are object features that are within the critical distance r of some image feature. The Hausdorff fraction is computed from the size of this set, M_L. In addition, we define the complementary subset of *unmatchable* features of the object $U_L = \{i \in M \mid d_I(L \oplus M_i) > r\} = M - M_L$, also corresponding to a fixed location L. The set U_L consists of features of the object that are greater than distance r from any image features. Equation (14) implies that $1_S \subseteq M_L$. Thus, we note that the features in U_L must be mismatched (i.e., $U_L \subseteq 0_S$).

The main idea of the spatially coherent matching scheme is to require that matching features should form large connected groups. There should be no isolated matches. Assume for the moment the chain neighborhood model of Section 4.2, where $\gamma(i, j)$ denotes the number of chains in the shortest sequence $\{i, i_1\}$, $\{i_1, i_2\}$, $\ldots, \{i_{k-1}, j\}$ in $\mathcal{N}_M$ connecting two feature i and j in M. Let B_L denote the subset of features in M_L that are "near" features of U_L. That is, $B_L = \{i \in M_L \mid \exists\, j \in U_L, \gamma(i, j) \leq R\}$, where R is a fixed integer parameter. We will refer to B_L as a *boundary* of the set of matchable features M_L. For example, in

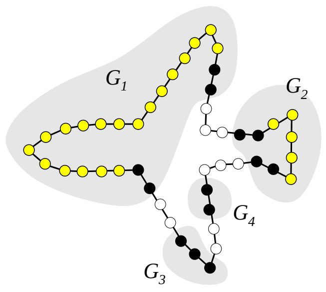


Figure 4: The spatially coherent matching technique. The set of matchable features M_L is the same as in Figure 3(b). The features of M_L (covered by a shade) form four groups $\mathcal{G}_L = \{G_1, \ldots, G_4\}$. The unmatchable features U_L are white. The boundary features B_L for $R = 2$ are shown black. In this case $\tilde{S}$ is given by $1_{\tilde{S}} = G_1 \cup G_2$ since G_1 and G_2 are the only groups in $\mathcal{G}_L$ that contain some non-boundary features.

Figure 4 the black features are the boundary B_L corresponding to M_L in Figure 3(b) for $R = 2$. The locally coherent matching technique works as follows. The main test is

$$|M_L| - |B_L| \quad > \quad K \tag{15}$$

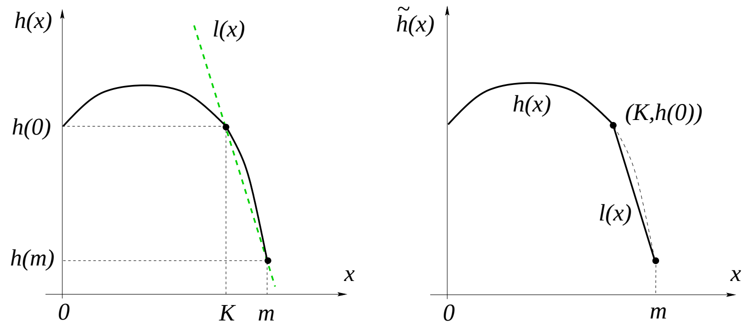
where K is the same as in (13). Note that $|M_L| - |B_L|$ is the number of non-boundary features in M_L. If (15) is false then $S_L = \bar{0}$ and there is no match. If the number of non-boundary features is sufficiently large so that (15) holds then the matching configuration is $S_L = \tilde{S}$ where

$$1_{\tilde{S}} \quad = \bigcup_{G \in \mathcal{G}_L \;:\; G \not\subseteq B_L} G\,. \tag{16}$$

Note that in (16) we include groups G in M_L which have some non-boundary features. Therefore, the match is assigned only to those features which belong to sufficiently large connected components in M_L. For example, in the case of Figure 4 we have $1_{\tilde{S}} = G_1 \cup G_2$. The groups G_3 and G_4 are discarded because they lie completely inside the boundary B_L.

The spatially coherent matching technique is easy to implement using morphological dilation. In practice the boundary set B_L can be estimated by dilating the unmatchable features U_L in the object's feature space by the radius R and then collecting the matchable features in M_L that lie in the dilated area. That is, the boundary features are those matchable features that lie within distance R of some unmatchable feature.

The spatially coherent matching method can be seen as a generalization of Hausdorff matching technique explained in Section 4.1, because the SCM techniques is equivalent to Hausdorff matching when $R = 0$. For $R > 0$ the size of the boundary $|B_L|$ is small if the matchable features M_L are grouped in large connected blobs and $|B_L|$ is large if the matchable features are isolated from each other. Therefore, the SCM technique (for $R > 0$) is reluctant to match if the features in M_L are scattered in small groups even if the size of M_L is large. In contrast, Hausdorff matching cares only about the size of M_L and ignores connectedness.



a) A plot of $h(x)$ and $l(x)$. b) A plot of $\tilde{h}(x)$.

Figure 5: Approximating the function $h(x)$ by $\tilde{h}(x)$.

We now show that the spatially coherent matching technique approximates the exact solution for a model with a chain neighborhood system introduced in Section 4.2. We consider the case when $\beta > \beta_o$ so that $K > 0$. Then $H_L(S)$ in (14) is well approximated by

$$\tilde{H}_L(S) \quad = \quad \begin{cases} \beta_N \cdot b(S) + \tilde{h}(|1_S|) & \text{if } 1_S \subseteq M_L \\ \infty & \text{if } 1_S \not\subseteq M_L. \end{cases}$$

We take $\tilde{h}(x) = h(x)$ for $0 \le x \le K$ and $\tilde{h}(x) = l(x)$ for $K \le x \le m$ where

$$l(x) \quad = \quad \beta \cdot m \cdot (K - x) \; + \; m \cdot (\alpha - \ln C_0)$$

is a line that agrees with $h(x)$ at $x = K$ and $x = m$, as shown in Figure 5(a). Figure 5(b) illustrates that $\tilde{h}(x)$ is a reasonable approximation of $h(x)$ and, therefore, $\tilde{H}_L(S) \approx H_L(S)$.

Theorem 1 *Assume that the neighborhood system forms a chain and that $\tilde{S}$ is defined as in (16). Assume also that $\beta_N = R \cdot \beta \cdot m$. Then the function $\tilde{H}_L(S)$ is minimized by $\tilde{S}$ if (15) holds and by $\bar{0}$ otherwise.*

The proof of Theorem 1 is split into three lemmas. Consider $F(S) = \beta_N \cdot b(S) + l(|1_S|)$.

Lemma 1 *The configuration $\tilde{S}$ minimizes the function $F(S)$ over $1_S \subseteq M_L$.*

PROOF: We need to minimize $\beta_N \cdot b_L(x) + l(x)$. In Section 4.2 we found that for chain neighborhood systems the function $b_L(x)$ is determined by the sizes $|G_1| \ge \ldots \ge |G_{n_L}|$ of the connected components in M_L. Following Figure 6, the function $\beta_N \cdot b_L(x) + l(x)$ is minimized by $\tilde{x} = |G_1| + \ldots + |G_k|$ where k is such that $|G_k| > \frac{2\beta_N}{\beta \cdot m} = 2R$ and $|G_{k+1}| \le \frac{2\beta_N}{\beta \cdot m} = 2R$. Note that the set $1_{\tilde{S}} = G_1 \cup \ldots \cup G_k$ consists of all connected components $G \in \mathcal{G}_L$ which are not inside B_L, that is, which have sizes $> 2R$. Therefore $\tilde{x} = |1_{\tilde{S}}|$ and $b(\tilde{S}) = 2k = b_L(\tilde{x})$. ∎

Lemma 2 *The function $\tilde{H}_L(S)$ is minimized by $\tilde{S}$ if $F(\tilde{S}) < h(0)$ and by $\bar{0}$ otherwise.*

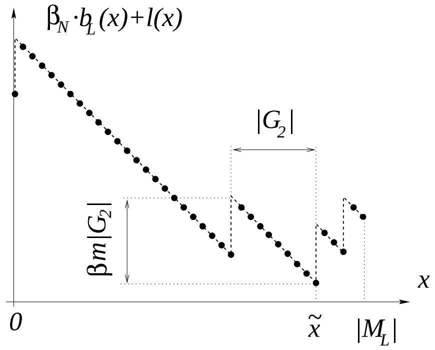


Figure 6: The plot of $\beta_N \cdot b_L(x) + l(x)$ corresponding to $b_L(x)$ in the example of Figure 3(c). All intervals have the same slope $-\beta m$ which is determined by the slope of the line $l(x)$. The vertical jumps between the intervals are all equal to $2\beta_N$. In this example we have $\tilde{x} = |G_1| + |G_2|$. Note that $\beta m|G_2| > 2\beta_N$ and $\beta m|G_3| \leq 2\beta_N$.

PROOF: If $x \in [0, K]$ then $h(x) \geq h(0)$. Thus, $\bar{0}$ minimizes $\tilde{H}_L(S)$ for $0 \leq |1_S| \leq K$. Let S^o be a configuration minimizing $\tilde{H}_L(S)$ for $K \leq |1_S| \leq |M_L|$. If $F(\tilde{S}) < h(0)$ then $l(|1_{\tilde{S}}|) < h(0)$, and as illustrated in Figure 5(a), this implies that $|1_{\tilde{S}}| > K$. Note that $\tilde{H}_L(S) = F(S)$ for $|1_S| > K$ and $1_S \subseteq M_L$. Thus, Lemma 1 gives $S^o = \tilde{S}$ and $\tilde{H}_L(\tilde{S}) = F(\tilde{S}) < h(0) = \tilde{H}_L(\bar{0})$ implies that $\tilde{S}$ is an optimal configuration. If $F(\tilde{S}) \geq h(0)$ then $\tilde{H}_L(S^o) = F(S^o) \geq F(\tilde{S}) \geq h(0) = \tilde{H}_L(\bar{0})$ and, therefore, the optimal configuration is $\bar{0}$. ■

Lemma 3 *The test $F(\tilde{S}) < h(0)$ is equivalent to the inequality in (15).*

PROOF: We have $F(\tilde{S}) = \beta_N \cdot b(\tilde{S}) + \beta \cdot m \cdot (K - |1_{\tilde{S}}|) + h(0)$. By substituting $|1_{\tilde{S}}| = |G_1| + \ldots + |G_k|$ and $\beta_N = R \cdot \beta \cdot m$ we derive that $F(\tilde{S}) < h(0)$ is equivalent to

$$|G_1| + \ldots + |G_k| - R \cdot b(\tilde{S}) \quad > \quad K. \tag{17}$$

Note that $1_{\tilde{S}}$ is a union of all groups $G \in \mathcal{G}_L$ which are larger than $2R$. Then $R \cdot b(\tilde{S}) = 2Rk$ gives the number of boundary features inside $1_{\tilde{S}}$. Since all non-boundary features of M_L are inside $1_{\tilde{S}}$ then the left hand side of (17) gives the total number of non-boundary features, that is $|M_L| - |B_L|$. ■

5 Monte Carlo results

In order to evaluate the recognition measures developed in this paper, we have run a series of experiments using Monte Carlo techniques to estimate Receiver Operating Characteristic (ROC) curves for each measure. A ROC curve plots the probability of detection along the y-axis and the probability of false alarm along the x-axis. Thus, the ideal recognition algorithms would produce results near the top left of the graph (low probability of false alarm and high probability of detection).

We use the experimental procedure reported in [7], where it was shown that Hausdorff matching works better than a number of previous binary image matching methods including correlation and Chamfer matching. For that reason we are mainly interested in comparing the algorithms developed here with Hausdorff matching, because it has already been shown to have better performance than these other techniques. Thus we contrast Hausdorff matching with the graph cut algorithm and the spatially coherent matching techniques. In Section 5.1 we explain some extra details about implementing the new recognition schemes. In 5.2 we discuss the Monte Carlo technique used to estimate the ROC curves and present the results.

5.1 Implementation of Recognition Techniques

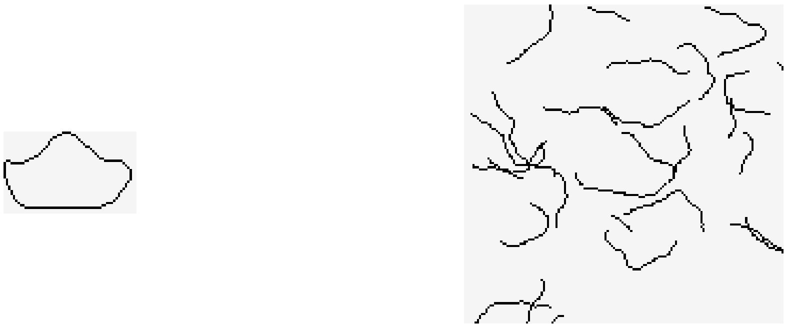
In this section we provide some details of our implementation of the general graph cut solution introduced in Section 3 and the SCM technique from Section 4.3.

Section 3 describes the graph cut technique that can be used to find a configuration S_L minimizing $H_L(S)$ in (8) for a fixed location L. This method is general and applies to any recognition problem that can be expressed within the framework of Section 2.

The method requires finding a min cut on a graph, illustrated in Figure 1. For the experiments presented in this paper we used the push-relabel algorithm [4] to find such a cut. For objects with around 100 features this approach produced an optimal configuration S_L for a fixed location L in milliseconds. Since the best match $\{\hat{S}, \hat{L}\}$ has to be chosen over all $L \in \mathcal{L}$, we have to run the min cut algorithm for all possible locations of the object in the image. In the experiments below the value of L specifies the translation of the object in the image. For an image of size 100×100 pixels, with integral translations at the pixel values, there are 10000 different values of L. Thus, the running time adds up to several seconds per image. This performance of the general graph cut algorithm is acceptable for our purposes. However, given that the Monte Carlo simulation requires processing a large number of images we further accelerate the running time of our experiments by skipping locations L in each image where the number of matchable features is less than 50% of the total number of object features. The number of matchable features, M_L, can be computed quickly using morphological techniques as explained below.

For our experiments we apply the general graph cut approach using the neighborhood system described in Section 4.2. This allows us to evaluate the the exact solution of our framework in this case. As was shown in Section 4.3, the spatially coherent matching is closely related to this exact solution. The Monte Carlo results presented in the next section support this.

The SCM technique is simple to implement using image morphology. Given the set of model features, M, and location, L, the set of matchable features M_L are those within distance r of image features. This can be computed by dilating the set of image features I by radius r (replacing each feature point with a disc of radius r). Now the set M_L is simply the intersection of M with this dilated image. The next step is to compute the boundary B_L which is the subset of features in M_L that are within distance R of some feature in U_L, the set of unmatchable features. Recall that $U_L = M - M_L$. Again, we can find features in one set near the features in some other set using dilation. Dilating the set U_L by R, and taking the intersection with M_L yields B_L, the points of M_L within distance R of points in U_L.



a) Example of an object. b) A simulated image (4% of clutter). The perturbed and partly occluded (30% of occlusion) instance of the object is located in the center.

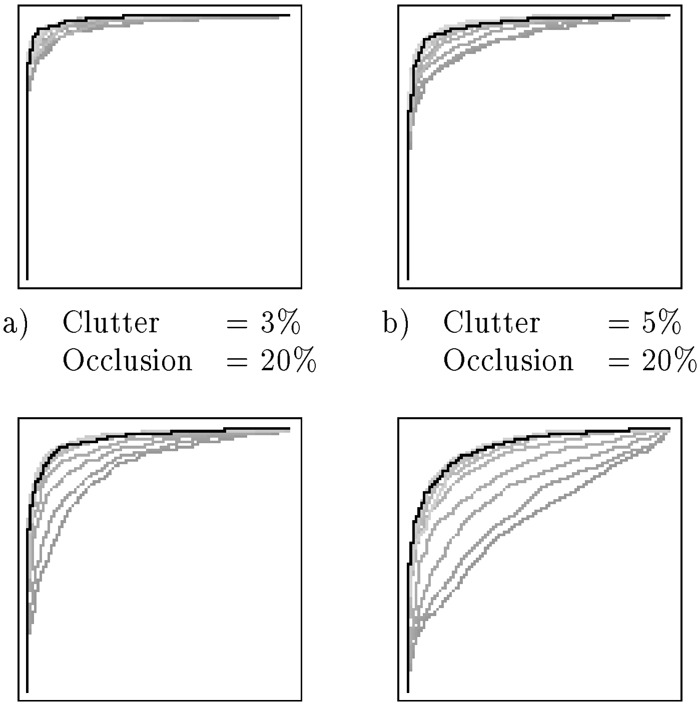
Figure 7: Monte Carlo experiments.

The quality of the match produced by the spatially coherent matching technique at each location L is determined by the number of non-boundary matchable features, that is, by $|M_L| - |B_L|$. Note that the search for the best match over all values of $L \in \mathcal{L}$ can be accelerated using the same pruning techniques that were developed for the Hausdorff measure [11]. This follows from a simple fact that if the Hausdorff measure gives no match at L then the spatially coherent matching technique can not match at L either. It is easy to see that $|M_L| < K$ implies that the test in (15) is necessarily false.

5.2 ROC Curves

We have estimated ROC curves for the measures described above by performing matching in synthetic images and using the matches found in these images to estimate the curve over a range of possible parameter settings. 1000 test images were used in the experiments, and were generated according to the following procedure. Random chains of edge pixels with a uniform distribution of lengths between 20 and 60 pixels were generated in a 150×150 image until a predetermined fraction of the image was covered with such chains. Curved chains were generated by changing the orientation of the chain at each pixel by a value selected from a uniform distribution between $-\frac{\pi}{8}$ and $+\frac{\pi}{8}$. An instance of the object was then placed in the image, after rotating, scaling, and translating the object by random values. The scale change was limited to $\pm 10\%$ and the rotation change was limited to $\pm\frac{\pi}{18}$. Occlusion was simulated by erasing the pixels corresponding to a connected chain of the model image pixels. Gaussian noise was added to the locations of the model image pixels ($\sigma = 0.25$). The pixel coordinates were finally rounded to the closest integer. This procedure was also used in [7].

For the experiments reported here, we performed recognition using the 56×34 object shown in Figure 7(a). This object contains 126 edge features. An example of a synthetic image generated using this object and the procedure described above is shown in Figure 7(b). In each trial, a given matching measure with a given parameter value was used to find all the matches of the object to the image. A trial was said to find the correct object if the



a) Clutter = 3%, Occlusion = 20%
b) Clutter = 5%, Occlusion = 20%
c) Clutter = 3%, Occlusion = 40%
d) Clutter = 5%, Occlusion = 40%

Figure 8: ROC curves for various levels of image clutter and occlusion of the object. The black curve corresponds to the best result obtained by the general graph approach. The gray curves correspond to the spatially coherent matching technique for various values of R: 0, 1, 2, 4, 7, 10, 13, 17, 21, 25. The best curves for the spatially coherent matching correspond to R in the range of 17-25. The Hausdorff matching ($R = 0$) corresponds to the lowest curve in all of the plots shown above.

position (considering only translation) of one of the matches was within three pixels of the correct location of the object in the image. A trial was said to find a false positive if any match was found outside of this range (and that match was not contiguous with a correct match position). Thus note that the test images were formed by slight rotation and scaling of the object model, but the searched was only done under translation. Any non-translational change to the object was not modeled by the matching process.

Figure 8 shows the ROC curves corresponding to experiments with different levels of occlusion and image clutter. The black curve shows the best results we could obtain from the general graph approach. The gray curves correspond to the spatially coherent matching technique for various values of $R \in [0, 25]$. As R gets larger, up to 20 or 21, the results improve, so the curves closer to the top left are for larger values of R. For even larger values of R, which we do not show, the ROC curves rapidly deteriorate. It is interesting to note that given this particular object, a distance of $R = 25$ corresponds approximately to the height of the object. Thus the performance does not deteriorate until the coherence region begins connecting together disconnected pieces of the object.

The case of $R = 0$ corresponds to Hausdorff matching. Thus the spatial coherence approach plays a large role in improving the quality of the match, because $R = 0$ has the worst matching performance. Note that in [7], using the same Monte Carlo framework, it was shown that Hausdorff matching works better than a number of other methods including binary correlation and Chamfer matching. Thus these results indicate that spatially coherent matching is a substantial improvement over several commonly used binary image matching techniques.

It should be noted that the value of R does not make a big difference for lower clutter or occlusion cases (top row of the figure), but makes a very large difference when these are larger (bottom row of the figure). Thus we see that for "easy" recognition problems, the spatial coherence of the matches is less important (though still offers a slight improvement). However as the object becomes more occluded and as there are more distractors, it becomes quite important to consider the spatial coherence of the matches. It should also be noted that in real imaging situations there would likely be small gaps in the instance of an object for which it would be undesirable that the SCM technique penalize such gaps. Recall that the parameter r can be used to cause features of the object model to match across small gaps in the image. Any larger gaps would then be subject to penalty based on the value of R.

6 Conclusion

We have presented a new Bayesian approach to object recognition using Markov random fields (MRF's). The central idea underlying this approach is to explicitly capture dependencies between individual features of an object. Markov random fields provide a good theoretical framework for representing such dependencies between features. These MRF's can be solved efficiently in practice, moreover we present fast approximation methods that do not require solving the MRF estimation problem.

Our approach contrasts with most feature-based object recognition techniques, as they do not explicitly account for dependencies between features of the object. It is desirable to be able to account for such dependencies, because they occur in real imaging situations. For example, a common case occurs with partial occlusion of objects, where features that are near one another in the image are likely to be occluded together. Our framework represents only pairwise dependencies between features of the object, however these are rich enough to model effects such as partial occlusion.

We showed that the generalized Hausdorff matching technique can be viewed as a special case of our approach, where the dependencies between all pairs of features in the object are equally strong. We then suggested a closely related method, which we call *spatially coherent matching* (SCM). This method requires that matching features be more than some critical distance from features that do not match, thus ensuring spatially contiguous sets of matching features. Monte Carlo experiments demonstrate that this SCM approach is a substantial improvement over Hausdorff and other previous matching techniques, in cases where the image is cluttered with many irrelevant features and there is substantial occlusion of the object to be recognized.

References

[1] H.G. Barrow, J.M. Tenenbaum, R.C. Bolles, and H.C. Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. In *The International Joint Conference on Artificial Intelligence*, pages 659–663, 1997.

[2] Y. Boykov, O. Veksler, and R. Zabih. Markov random fields with efficient approximations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 648–655, 1998.

[3] L. Ford and D. Fulkerson. *Flows in Networks.* Princeton University Press, 1962.

[4] A. Goldberg and R. Tarjan. A new approach to the maximum flow problem. *Journal of the Association for Computing Machinery*, 35(4):921–940, October 1988.

[5] D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society, Series B*, 51(2):271–279, 1989.

[6] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the Hausdorf distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, September 1993.

[7] Daniel P. Huttenlocher. Monte carlo comparison of distance transform based matching measures. In *DARPA Image Understanding Workshop*, 1997.

[8] S. Z. Li. *Markov Random Field Modeling in Computer Vision.* Springer-Verlag, 1995.

[9] Clark F. Olson. A probabilistic formulation for Hausdorff matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 150–156, 1998.

[10] Arthur Pope and David G. Lowe. Learning probabilistic appearance models for object recognition. In Shree K. Nayar and Tomaso Poggio, editors, *Early Visual Learning*, pages 67–98. Oxford University Press, 1996.

[11] William Rucklidge. *Efficient Visual Recognition Using the Hausdorff Distance*. Number 1173 in Lecture Notes in Computer Vision. Springer-Verlag, 1996.

[12] Jayashree Subrahmonia, David B. Cooper, and Daniel Keren. Practical reliable bayesian recognition of 2D and 3D objects using implicit polynomials and algebraic invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(5):505–519, May 1996.