Spatial Priors for Part-Based Recognition using Statistical Models

David Crandall[1]
Cornell University
crandall@cs.cornell.edu

Pedro Felzenszwalb
The Univerisity of Chicago
pff@cs.uchicago.edu

Daniel Huttenlocher
Cornell Univerisity
dph@cs.cornell.edu

Abstract

We present a class of statistical models for part-based object recognition that are explicitly parameterized according to the degree of spatial structure they can represent. These models provide a way of relating different spatial priors that have been used for recognizing generic classes of objects, including joint Gaussian models and tree-structured models. By providing explicit control over the degree of spatial structure, our models make it possible to study the extent to which additional spatial constraints among parts are actually helpful in detection and localization, and to consider the tradeoff in representational power and computational cost. We consider these questions for object classes that have substantial geometric structure, such as airplanes, faces and motorbikes, using datasets employed by other researchers to facilitate evaluation. We find that for these classes of objects, a relatively small amount of spatial structure in the model can provide statistically indistinguishable recognition performance from more powerful models, and at a substantially lower computational cost.

1. Introduction

Since the 1970's it has been observed that many objects can be represented in terms of a set of parts arranged in a deformable configuration (e.g., [1, 2, 3, 6, 7, 8, 9, 10, 11]). In such models, each part is generally represented by a small template, or other local image information, and the spatial relationships between parts are represented by statistical models or spring-like connections between pairs of parts. Recently there has been a considerable resurgence in the use of these models for object recognition – both for detection and localization – and in the learning of such models from example images. Particular emphasis has been on the use of these models for recognizing generic *classes* of objects that are learned from specific examples.

The forms of spatial priors that have been used to capture geometric relationships between parts of an object differ substantially in their representational power. In general there is a tradeoff between representational power and computational complexity. On one hand, joint Gaussian models (e.g., [3, 7]) have been used to capture explicit spatial dependencies between all pairs of parts, but detection and localization algorithms using these models have relied on search heuristics in order to be computationally tractable. On the other hand, tree-structured graphical models (e.g., [6, 9]) have been used to efficiently detect and localize certain kinds of objects such as humans and faces, but are only able to explicitly capture a small fraction of the spatial dependencies between parts of an object. The main goal of this paper is to improve our understanding of such tradeoffs between representational power and computational complexity for part-based recognition. We do this by introducing a parameterized family of spatial priors that vary in the degree of spatial structure that they can capture.

We use a problem formulation similar to the one in [3, 6, 8], where for detection or localization a single overall problem is solved that takes into account how well individual parts match each image location and the global spatial arrangement of parts. This approach has been referred to as a "soft detection" strategy, because rather than first detecting features and then using detected locations to find good configurations, both problems are solved together. Often soft detection approaches have been avoided due to computational cost. In [6] an efficient method was developed for soft detection in the case of tree-structured models. However tree-structured models may not always be appropriate because of their relative lack of spatial structure.

In this paper we introduce a class of graphs that we call k-fans. Graphical models defined by k-fans provide a natural family of spatial priors for part-based recognition. The parameter k controls both the representational power of the models and the computational cost of doing inference with them. At one extreme, $k = 0$, there is no dependence between the locations of different object parts. When $k = 1$ the structure is that of a star graph. For $k = n - 1$ (where n is the number of parts in the model), there are dependencies between all pairs of parts, as in the case of a joint Gaussian model. Not only does this family of models allow us to study computational issues, but it also provides a natural way of investigating the degree to which additional

spatial constraints improve recognition performance. Using more powerful models (in this context, models with larger k) does not necessarily improve classification, as it can lead to over-fitting during learning.

For certain object classes that have been used recently in the literature, such as motorbikes, faces and airplanes, a relatively small amount of spatial structure provides nearly the same recognition accuracy as is obtained using more powerful models. The models with less spatial structure can be used for detection and localization of these objects with a substantially lower computational cost.

2. Statistical Models

Consider a model with n parts $V = (v_1, \ldots, v_n)$. The location of the object in an image is given by a configuration of its parts $L = (l_1, \ldots, l_n)$, where l_i is the location of the ith part. Throughout this paper we assume that the location of a part is given by a point in the image, $l_i = (x_i, y_i)$.

The spatial relationships between parts in a model are captured by a set of parameters S, while the appearance of each part is characterized by a set of parameters A. The pair $M = (S, A)$ defines an object model. Using Bayes' law the probability that the object is at a particular location given an image and fixed model parameters can be written as,

$$p_M(L|I) \propto p_M(I|L)p_M(L). \tag{1}$$

Here $p_M(I|L)$ is the likelihood of seeing image I given a particular configuration for the object parts and $p_M(L)$ is the prior probability that the object would assume the spatial configuration L. There are three fundamental problems that can be formulated in terms of these distributions:

Detection The detection problem is to decide if the image has an instance of the object (hypothesis w_1) or if the image is background-only (hypothesis w_0). It is natural to consider the likelihood ratio,

$$q = \frac{p_M(I|w_1)}{p_M(I|w_0)}. \tag{2}$$

Where the numerator is usually computed by summing over possible configurations L (see Section 3.4).

Localization Assuming the object is present in an image, the location that is most likely its true position is one with maximum posterior probability,

$$L^* = \arg\max_L p_M(L|I).$$

Supervised learning The maximum-likelihood estimate of the model parameters $M = (S, A)$ given a set of T labeled training images $\{(I_1, L_1), \ldots, (I_T, L_T)\}$ is,

$$\begin{aligned} S^* &= \arg\max_S \prod_i p_M(L_i), \\ A^* &= \arg\max_A \prod_i p_M(I_i|L_i). \end{aligned}$$

The algorithmic complexity of solving these three problems is highly dependent on the form of the likelihood model $p_M(I|L)$ and the spatial prior $p_M(L)$. Most approaches use similar restrictions on the form of the likelihood. The focus of this paper is primarily on the form of the spatial prior.

2.1. Appearance

For computational purposes the most important property of the appearance model is that $p_M(I|L)$ factors into a product of functions, each dependent on the location of a single part, and one extra term which does not depend on the object configuration. The majority of the recent work on part-based recognition has used a similar factorization.

In our models the appearance of a part is given by a template. Let I be the output of an oriented edge detector. For each pixel p, $I(p)$ is either 0 indicating that there is no edge at p or a value in $\{1, \ldots, r\}$ indicating that there is an edge in one of r possible orientations at p. We assume that the values of each pixel in the image are independent given the location of the object. Let $\mathcal{T}_i$ be the set of pixels in the ith template. The probability that a pixel $p \in \mathcal{T}_i$ has value u is defined by a foreground model for that part $f_i(p)[u]$. Each pixel in the background has value u with probability $b[u]$.

Let w_0 be the hypothesis that the object is not present in the image. By our independence assumption we have,

$$p_M(I|w_0) = \prod_p b[I(p)].$$

We say that parts i and j do not overlap if $(\mathcal{T}_i \oplus l_i) \cap (\mathcal{T}_j \oplus l_j) = \emptyset$. Here $\oplus$ denotes Minkowsky addition, which is used to translate the templates according the the locations of the parts. For a configuration L without overlap we have,

$$p_M(I|L) = p_M(I|w_0) \prod_{v_i \in V} g_i(I, l_i), \tag{3}$$

where

$$g_i(I, l_i) = \prod_{p \in \mathcal{T}} \frac{f_i(p)[I(p + l_i)]}{b[I(p + l_i)]}.$$

Each term in g_i is the ratio of the foreground and background probabilities for a pixel that is covered by template $\mathcal{T}_i$. In equation (3) the denominator of g_i cancels out the contribution of $p_M(I|w_0)$ for those pixels that are under some part. As long as we only consider configurations L without overlapping parts the likelihood function defined above is a true probability distribution over images (it integrates to one). When parts overlap this is an approximation. Note that for many objects the spatial prior $p_M(L)$ enforces that parts in the model do not overlap.

2.2. Spatial Prior

The simplest approach is to assume that the part locations are independent (a naive Bayes assumption),

$$p_M(L) = \prod_{v_i \in V} p_M(l_i).$$

The localization problem is particularly easy with this prior. To maximize $p_M(L|I)$ we just need to maximize $g_i(I, l_i)p_M(l_i)$ independently for each l_i. If there are n parts and h locations in the image this can be done in $O(nh)$ time. While this model yields tractable inference and learning procedures, it encodes only weak spatial information and is unable to accurately represent multi-part objects.

The other extreme is to make no independence assumption on the locations of different parts by, for example, using a joint Gaussian model for the spatial distribution $p_M(L)$ as in [3]. Learning this distribution from labeled images is easy. However it is not known how to perform inference (localization or detection) using this spatial prior efficiently, and various heuristics have been employed when using these models. Most methods rely on hard feature detection to constrain the possible locations of each part.

Spatial models between the two extremes just described can be defined by making certain conditional independence assumptions. These assumptions are commonly represented using an undirected graphical model (or a Markov random field). Let $G = (V, E)$ be an undirected graph. The graph is used to define a distribution for the random variables $(l_1, \ldots, l_n)$ in the following way. The value for the location of v_i is independent of the values of all other nodes, conditioned on the values of the neighbors of v_i in the graph.

There are efficient learning and inference procedures for models with tree-structured spatial priors, where the detection and localization problems can be solved in $O(nh^2)$ time using dynamic programming. In many cases one can solve these problems in $O(nh)$ time – the same asymptotic time as the naive Bayes case where there are no dependencies between part locations (see [6]).

3. k-fans

Now we consider a class of spatial priors that lie between the two extremes of a naive Bayes assumption and a full joint Gaussian model. Our goal is to find models with performance comparable to the joint Gaussian but that support fast procedures for exact (discrete) inference and learning. We start by considering a restricted form of tree model and then extend that model. A star graph is a tree with a central node that is connected to all other nodes. Let $G = (V, E)$ be a star graph with central node v_r. Graphical models with a star structure have a particularly simple interpretation in terms of conditional distributions. The values of random variables associated with nodes $v_i \neq v_r$ are independent when conditioned on the value of v_r. This leads to the following factorization of the prior distribution,

$$p_M(L) = p_M(l_r) \prod_{v_i \neq v_r} p_M(l_i|l_r).$$

We can think of the central node v_r as a *reference* part. The position of other parts in the model are evaluated relative to the position of this reference part.

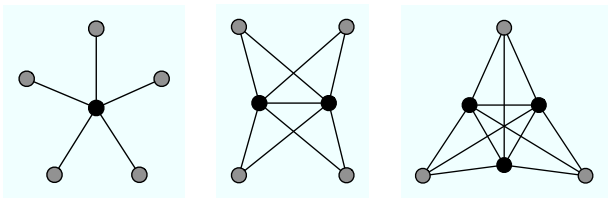


Figure 1. Some k-fans on 6 nodes. The reference nodes are shown in black while the regular nodes are shown in gray.

More generally let $R \subseteq V$ be a set of reference parts, and $\overline{R} = V - R$ be the remaining parts in a model. The set R can be used to define a graph, which we call a k-fan for $k = |R|$. This graph consists of a complete subgraph over the nodes in R, and each node in $\overline{R}$ is connected to every node in R (and nothing else). Recall that in a clique there is an edge connecting each pair of nodes. A k-fan can be seen as a collection of cliques of size $k + 1$ glued together along a common clique of size k. The k nodes in the common clique are the reference parts R. Some examples of k-fans on 6 nodes are shown in Figure 1.

We claim that k-fans form an important class of graphical models for spatial priors. These are exactly the models where the locations of the non-reference parts are conditionally independent given the locations of the reference parts. Let $R = \{v_1, \ldots, v_k\}$ be the reference parts in a k-fan. We denote by $l_R = (l_1, \ldots, l_k)$ a particular configuration of the reference parts. The spatial prior defined by a k-fan can be written in conditional form as,

$$p_M(L) = p_M(l_R) \prod_{v_i \in \overline{R}} p_M(l_i|l_R). \quad (4)$$

The set of k-fans on n nodes form a family of graphs that lie between the completely disconnected graph ($k = 0$) and the complete graph ($k = n - 1$). In general both the localization and detection problems for models with spatial priors based on k-fans can be solved in $O(nh^{k+1})$ time, where as before n is the number of parts in the model and h is the number of locations in the image. Thus k controls the computational complexity of inference with these models. With the additional assumption that $p_M(L)$ is Gaussian we can use distance transforms and convolutions to solve the inference problems in $O(nh^k)$, as described below. In practice

the running time can be further improved using conservative pruning heuristics.

For learning k-fan models it will be useful to write down the spatial prior in terms of marginal distributions,

$$p_M(L) = \frac{\prod_{v_i \in \overline{R}} p_M(l_i, l_R)}{p_M(l_R)^{n-(k+1)}}. \quad (5)$$

The numerator is the product of marginal probabilities for the $n-k$ maximal cliques and the denominator involves the marginal probability for the nodes shared by all maximal cliques (the so-called separator set which in this case is R). This is a special form of the factorization for a triangulated graph, which is the ratio of a product over maximal cliques and a product over separators.

3.1. Geometric Interpretation

There is a natural connection between k-fan models and object recognition using geometric invariants. Each maximal clique in a k-fan consists of exactly $k+1$ parts, and the location of these parts can be used to define shape constraints that are invariant to certain geometric transformations (see [4]). The number of reference parts controls the type of geometric invariants that can be represented.

In a k-fan the location of a non-reference part can be described in a reference frame defined by the locations of the k reference parts. For example, when $k = 1$ the location of a non-reference part can be described relative to the location of the single reference part. The values $l'_i = l_i - l_r$ are invariant under translations, so 1-fans can be used to define translation invariant models. For the case of $k = 2$ the two reference parts can be used to define models that are invariant to rigid motions and global scaling. When $k = 3$ we can use the three reference parts to define an affine basis in the image plane, if the location of every non-reference part is described in this basis we obtain affine invariant models. These models are important because they capture arbitrary views of planar objects under orthographic projection.

To enforce geometric invariants over $k+1$ parts we can define $p_M(l_i|l_R)$ to be one if the $k+1$ locations satisfy a geometric constraint and zero otherwise. In general our models capture soft geometric constraints, giving preference to configurations that satisfy relationships on $k+1$ features as much as possible. The distribution over the reference part locations $p_M(l_R)$ could be uniform in the case where all geometric constrains are defined in terms of $k+1$ parts.

3.2. Gaussian k-fans

We now consider k-fan models with the additional constraint that $p_M(L)$ is a Gaussian distribution. For a Gaussian model the marginal distribution of any subset of variables is itself Gaussian. Let μ_R and Σ_R be the mean and covariance for the locations of the reference parts. The marginal distribution of the reference parts together with one non-reference part is given by a Gaussian with mean and covariance,

$$\mu_{i,R} = \begin{bmatrix} \mu_i \\ \mu_R \end{bmatrix}, \quad \Sigma_{i,R} = \begin{bmatrix} \Sigma_i & \Sigma_{iR} \\ \Sigma_{Ri} & \Sigma_R \end{bmatrix}. \quad (6)$$

These can be used to define the spatial prior in terms of equation (5). We will use this for learning Gaussian k-fans. For inference we use the conditional form of the prior in equation (4). For a Gaussian distribution, conditioning on a set of variables preserves the Gaussian property. The conditional distribution of a non-reference part location given particular locations for the reference parts $p_M(l_i|l_R)$ has mean and covariance,

$$\mu_{i|R}(l_R) = \mu_i + \Sigma_{iR}{\Sigma_R}^{-1}(l_R - \mu_R), \quad (7)$$

$$\Sigma_{i|R} = \Sigma_i - \Sigma_{iR}{\Sigma_R}^{-1}\Sigma_{Ri}, \quad (8)$$

Note how the covariance $\Sigma_{i|R}$ is independent of the location of the reference parts. This is a non-trivial property that enables the use of distance transforms and convolutions to obtain faster inference algorithms than is possible with non-Gaussian models.

3.3. Learning

We can learn the spatial prior for Gaussian k-fan models from labeled images using a maximum likelihood criterion. For a fixed set of reference parts, estimating the maximum likelihood parameters S^* involves estimating the mean and covariances in (6). These can be obtained from the sample mean and covariance of the labeled configurations.

The more interesting case is when the reference parts are not fixed. In this situation all possible reference sets of size k can be considered in order to find the set R that yields the best possible model. There are $\binom{n}{k}$ possible reference sets, and this is not very large for small values of k. For each reference set we compute the maximum likelihood model parameters using the simple procedure described above. We can select the best reference set based on the likelihood of the data under the maximum likelihood model for each R.

Learning the appearance parameters A^* for the models described in Section 2.1 using labeled training data is simple. To estimate f_i the position of the ith part in each training example is used to align the training images. The maximum likelihood estimate for $f_i(p)[v]$ is simply the frequency that pixel p has value v on the aligned data. The only parameter that is not learned from the data is the size and shape of the template $\mathcal{T}_i$. For the experiments shown in this paper we used rectangular windows of a fixed size.

3.4. Detection

For detection we consider the likelihood ratio in (2). The numerator can be expressed as a sum over all possible object

configurations,

$$p_M(I|w_1) = \sum_L p_M(L) p_M(I|L).$$

Using the likelihood function (3) we see that

$$\frac{p_M(I|w_1)}{p_M(I|w_0)} = \sum_L p_M(L) \prod_{v_i \in V} g_i(I, l_i).$$

For a k-fan model the sum over all configurations L can be factored using the conditional form of the spatial prior in (4). For each $v_i \in \overline{R}$ we define

$$\alpha_i(l_R) = \sum_{l_i} p_M(l_i|l_R) g_i(I, l_i).$$

Now the likelihood ratio can be computed as,

$$\frac{p_M(I|w_1)}{p_M(I|w_0)} = \sum_{l_R} \left(p_M(l_R) \prod_{v_i \in R} g_i(I, l_i) \prod_{v_i \in \overline{R}} \alpha_i(l_R) \right).$$

Note that each α_i can be computed by brute force in $O(h^{k+1})$ time, while the likelihood ratio can be computed using the α_i in $O(nh^k)$ time. This procedure gives an $O(nh^{k+1})$ algorithm for computing the likelihood ratio.

For the case of a Gaussian k-fan we can compute the likelihood ratio even faster, using convolutions. For each non-reference part v_i we have,

$$p_M(l_i|l_R) = \mathcal{N}(l_i, \mu_{i|R}(l_R), \Sigma_{i|R}),$$

a Gaussian distribution with mean and covariance given by equations (7) and (8). Let $\alpha'_i(l_i)$ be the convolution of $g_i(I, l_i)$ with a Gaussian kernel of covariance $\Sigma_{i|R}$. It is not hard to see that,

$$\alpha_i(l_R) = \alpha'_i(\mu_{i|R}(l_R)).$$

So each α_i can be implicitly computed by a convolution in the space of possible locations in the image. This can be done in $O(h \log h)$ time instead of $O(h^{k+1})$.

The overall running time of the likelihood ratio computation for the case of a Gaussian k-fan model is $O(nh^k + nh \log h)$. Note that for a 1-fan model this is almost the same as $O(nh)$, the time that it would take to compute the likelihood ratio if the locations of the parts were completely independent. The $\log h$ dependency can be removed by using linear time methods that approximate Gaussian convolutions, such as the box-filter technique in [12].

3.5. Localization

For localization we look for an object configuration L^* with maximum posterior probability. Using Bayes law the posterior distribution for a k-fan model can be written in terms of the likelihood function (3) and the spatial prior (4). By manipulating the terms we get,

$$p_M(L|I) \propto p_M(l_R) \prod_{v_i \in R} g_i(I, l_i) \prod_{v_i \in \overline{R}} p_M(l_i|l_R) g_i(I, l_i).$$

For any $v_i \in \overline{R}$ the quality of an optimal location for the ith part can be expressed as a function of the reference locations,

$$\alpha_i^*(l_R) = \max_{l_i} p_M(l_i|l_R) g_i(I, l_i).$$

Using the α_i^* we can express the posterior probability of an optimal configuration for the object with particular reference locations l_R as,

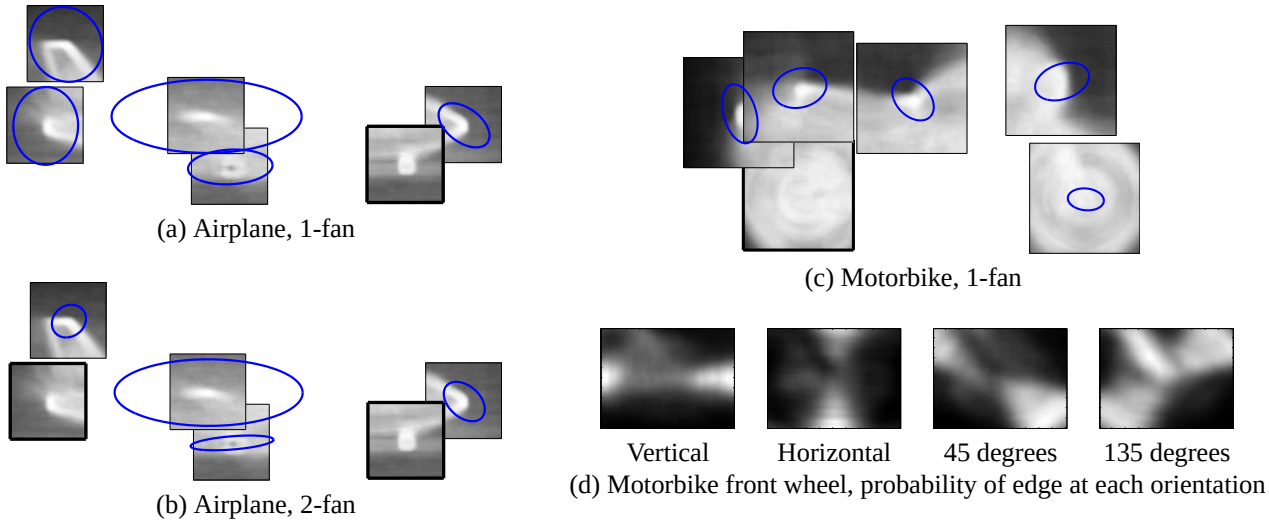
$$\beta^*(l_R) = p_M(l_R) \prod_{v_i \in R} g_i(I, l_i) \prod_{v_i \in \overline{R}} \alpha_i^*(l_R).$$

These functions can be used to compute an optimal configuration for the object in time polynomial in the number of parts n and the number of locations for each part h (but exponential in k). Each α_i^* can be computed by brute force in $O(h^{k+1})$ time, while β^* can be computed in $O(nh^k)$ time. An optimal configuration for the reference parts l_R^* is one maximizing β^*. Finally, for each non-reference part we select l_i^* maximizing $p_M(l_i|l_R^*) g_i(I, l_i)$. This can be done in $O(h)$ time. The overall running time of this procedure is $O(nh^{k+1})$, which is reasonable for very small k.

As in the case of detection we can speed up the localization procedure for Gaussian k-fans. For localization the role of convolutions is played by generalized distance transforms [6]. In this case the running time of the localization algorithm is reduced to $O(nh^k)$.

4. Experiments

We implemented the learning and localization methods for k-fan models and carried out experiments to investigate how increasing the degree of spatial constraints (i.e. increasing k) affects object detection and localization accuracy. To facilitate comparison of these results with previous work we used some of the datasets from [7]: airplanes (800 images), faces (435 images), motorbikes (800 images), and background (800 images). To further facilitate evaluation, we considered only the case of Gaussian k-fans (that is, we did not use the reference parts to define a geometric basis as described in Section 3.1). We tried to reproduce the experimental protocol of [7] as closely as possible, including using the same partitioning of the data into training and test images and using the same ground truth bounding boxes to normalize for the object size across images.



(a) Airplane, 1-fan

(b) Airplane, 2-fan

(c) Motorbike, 1-fan

(d) Motorbike front wheel, probability of edge at each orientation

Figure 2. Illustration of some of the learned models. Images (a) through (c) show part appearance models positioned at their mean configuration. The reference parts have a black border around them. The ellipses illustrate the location variances for a non-reference part given the locations of the references. High intensity pixels represent high edge probabilities. For clarity, just the probability of an edge is shown, although the actual models capture probabilities of each individual edge orientation. In (d), the probability map template for each edge orientation is shown for a sample part (the front wheel of the motorbike model). Note how the locations of parts in the 2-fan airplane model are more constrained than in the 1-fan model.

4.1. Learning the Models

As in [7], six parts were used to model each object. For airplanes we used the front and back landing gear, nose, wing tip, rear-most point of plane, and tail. For faces we used the two eyes, nose, two corners of the mouth, and chin. For motorbikes, the front and back wheel, headlight and tail light, and the front and back of the seat were used. Ground truth was collected by hand-labeling the training images. Note that [7] used an unsupervised training method but we should not expect supervised learning to necessarily give better results than unsupervised learning – a supervised approach is limited by the quality of the parts chosen and the accuracy of the hand-labeled ground truth.

The models were learned from labeled examples using the procedure described in Section 3.3. To learn the appearance model for a given part, a fixed-size patch surrounding the labeled part location was extracted from each training image. Canny edge detection was used to generate edge maps. Edge orientation was quantized into four directions (north/south, east/west, northeast/southwest, northwest/southeast) and represented as four separate binary edge maps. Morphological dilation was applied on each map independently. The four maps were then combined to form a single map with 16 possible values (corresponding to the 2^4 possible edge orientation combinations) at each pixel location. Foreground model probabilities were estimated by computing the frequency of each of the 16 edge values at each position in the template across the patches extracted from the training images. The background model probabilities were estimated from the observed density of edges in background images.

Figure 2 illustrates some of the models we have learned. Note that in each case the configuration of parts is readily recognizable.

4.2. Detection Results

For detection we found an optimal configuration for the object in each test image, using the procedure described in Section 3.5, and then used that location to approximate the likelihood ratio. With this approach each positive detection comes with a particular localization.

In the first set of detection experiments, we pre-scaled all images so that object width was roughly uniform, and all parameters were kept exactly the same over different object classes (template size = 50×50, dilation radius = 2.5 pixels). To prevent biases related to image size, we padded out all images to a large, uniform size.

Figure 3 shows the ROC curves generated from these experiments. For each object class, the figure compares ROC curves for k-fans with k ranging from 0 (no structure) to 2. We observe that for motorbikes, high accuracy is achieved using 0-fans, and adding spatial constraints gives little improvement. On the other hand, for airplanes, 1-fans perform significantly better than 0-fans, and 2-fans perform significantly better than 1-fans, indicating that increasing degrees of spatial constraints give better performance. We conclude

	Planes	Bikes	Faces
0-fans	90.5%	96.5%	98.2%
1-fans	**91.3%**	**97.0%**	98.2%
2-fans	**93.3%**	97.0%	98.2%

Table 1. Equal ROC performance for the detection experiments. A boldface number for a k-fan indicates a statistically significant difference between the areas under the ROC curves of the $k-1$ and k-fan models (with 95% confidence).

that the appropriate amount of spatial structure in the model varies from object to object.

Table 1 summarizes the recognition accuracy at the equal ROC points (point at which the true positive rate equals one minus the false positive rate). We note that our equal ROC results compare favorably with those obtained using full multivariate Gaussian structural models in [7]. They report 90.2%, 92.5% and 96.4% for airplanes, motorbikes and faces respectively, under the same experimental conditions. We applied the statistical test of DeLong et al [5] to judge the differences in areas under the ROC curves of the various models. These results are also shown in Table 1. For each object class we computed the probability that the area under the ROC curve for the k-fan model is significantly different from the area under the ROC curve for the model with one less reference part. Differences significant at a greater than 95% confidence level are shown in boldface in the table.

Finally, we conducted multi-class detection experiments, in order to test the ability of the models to differentiate between the three different object classes and the background images. For each test image, the three object detectors were applied, and the object class with the highest likelihood was chosen. That likelihood was compared to the threshold at the equal ROC point to decide between that object class and the background class. The results are shown in Table 2. The performance of multi-class recognition is similar to the single class case. The use of relatively accurate probabilistic models allows for direct comparison between the scores of each object class without tunning weighting parameters.

As in [7], we tested the detectors in a setting where the object scale was not known. The object widths varied between about 200 and 700 pixels for the motorbike and plane categories, while the face dataset has very little scale variation. We applied the detectors at four different scales to each image and chose the scale having the highest-likelihood detection. Recognition performance in this experiment was comparable to the case of pre-scaled images.

The average running time per image of the detection algorithm on these datasets on a 3GHz Pentium IV is approximately 0.1 seconds for a 1-fan model, 3.3 seconds for a 2-fan model, and 37.6 seconds for a 3-fan model.

[1]David Crandall is supported by an NSF graduate research fellowship.

4.3. Localization Accuracy

Figure 4 illustrates some localization results produced by our system on the motorbike dataset, showing precise localization of the parts despite substantial variability in their appearances and locations. Recent work has generally focused on evaluating detection performance but we believe it is also important to evaluate the accuracy of localization. For each object class, for the subset of images that were correctly classified during the detection task at the equal ROC point, the part locations produced by our system were compared to hand-labeled ground truth. We computed the trimmed means (at 75% and 90%) of the Euclidean distances (in pixels) between estimated locations and the ground truth. For the motorbike models the localization errors are reasonably small (less than 10 pixels) for most parts when $k > 0$, while the errors for faces are less than 2 pixels. Table 3 summarizes the results for the motorbikes models. In this case the localization accuracy is high for most parts when using a model without spatial structure. The accuracy increases as we add spatial constraints even when recognition performance does not increase.

References

[1] Y. Amit. *2D Object Detection and Recognition, Models, Algorithms, and Networks*. MIT Press, 2002.

[2] M.C. Burl and P. Perona. Recognition of planar object classes. In *CVPR*, 1996.

[3] M.C. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *ECCV*, 1998.

[4] S. Carlsson. Geometric structure and view invariant recognition. *Phil. Trans. R. Soc. Lond. A*, 359(1740), 1998.

[5] DeLong E.R., DeLong D.M., and Clarke-Pearson D.L. Comparing the areas under two or more correlated roc curves: a non-parametric approach. *Biometrics*, 44(3), 1998.

[6] P.F. Felzenszwalb and D.P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1), 2005.

[7] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003.

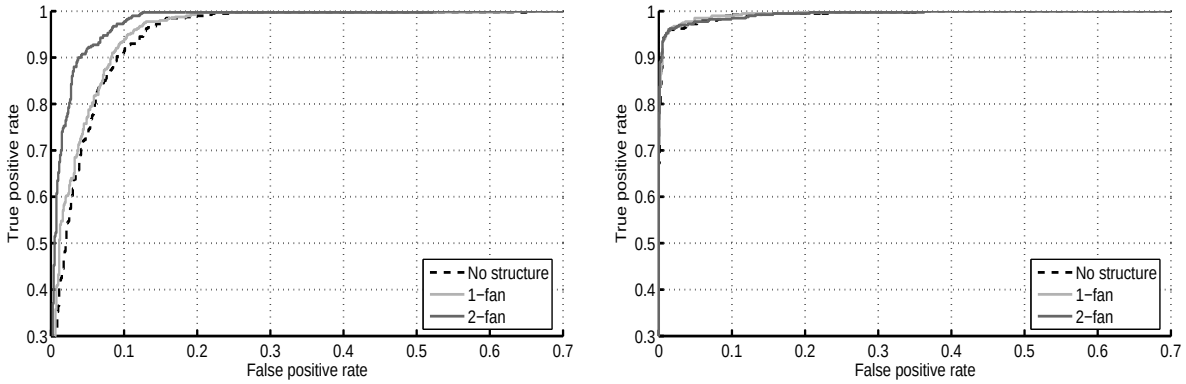
[8] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computer*, 22(1), 1973.

[9] S. Ioffe and D.A. Forsyth. Probabilistic methods for finding people. *IJCV*, 43(1), 2001.

[10] P. Lipson, E. Grimson, and P. Sinha. Configuration based scene classification and image indexing. In *CVPR*, 1997.

[11] H. Schneiderman and T. Kanade. Probabilistic formulation for object recognition. In *CVPR*, 1998.

[12] W.M. Wells, III. Efficient synthesis of Gaussian filters by cascaded uniform filters. *PAMI*, 8(2), 1986.



(a) (b)

Figure 3. Detection results for (a) airplanes and (b) motorbikes. Note that the ROC curves are truncated at a false positive rate of 0.7 and a true positive rate of 0.3.

	0-fan				1-fan				2-fan			
	Planes	Bikes	Faces	BG	Planes	Bikes	Faces	BG	Planes	Bikes	Faces	BG
Planes	357	10	0	33	362	5	0	33	370	8	0	22
Bikes	4	382	0	14	4	384	0	12	4	384	0	12
Faces	3	9	205	0	3	8	206	0	1	9	207	0
Background	72	28	0	700	68	24	0	708	53	23	0	724

Table 2. Confusion matrices for the multi-class detection experiments. Rows correspond to actual classes, while columns correspond to predicted classes.

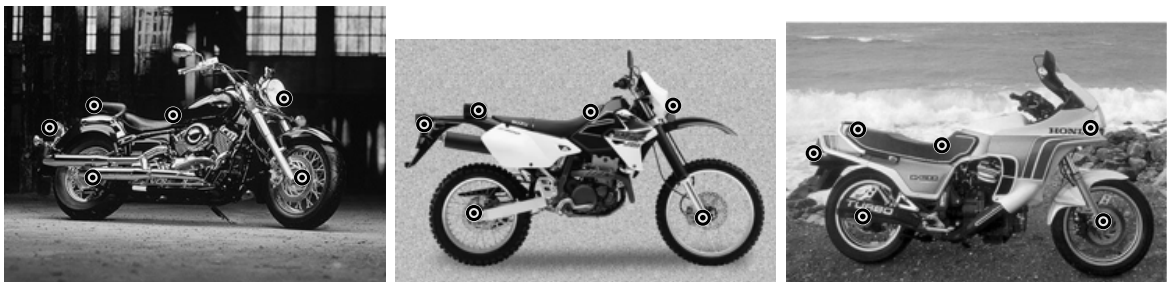


Figure 4. Sample localization results. In each of these cases all parts were localized correctly.

Model	Rear wheel		Front wheel		Headlight		Tail light		Back of seat		Front of seat	
	75%	90%	75%	90%	75%	90%	75%	90%	75%	90%	75%	90%
No structure	15.6	34.4	1.9	2.3	10.9	18.8	12.0	19.3	21.6	33.9	6.3	12.2
1-fan	2.1	12.5	1.9	2.3	10.9	18.6	11.4	18.7	20.6	32.9	6.3	12.0
2-fan	1.9	2.4	1.9	2.3	10.1	16.6	11.0	18.3	17.2	28.5	5.4	9.3

Table 3. Part localization errors for the correctly detected motorbike images, showing 75% and 90% trimmed means of Euclidean distance between estimated part locations and ground truth.