Weakly Supervised Learning of Part-Based Spatial Models for Visual Object Recognition

David J. Crandall and Daniel P. Huttenlocher

Cornell University, Ithaca, NY 14850, USA,
{crandall,dph}@cs.cornell.edu

Abstract. In this paper we investigate a new method of learning part-based models for visual object recognition, from training data that only provides information about class membership (and not object location or configuration). This method learns both a model of local part appearance and a model of the spatial relations between those parts. In contrast, other work using such a weakly supervised learning paradigm has not considered the problem of simultaneously learning appearance and spatial models. Some of these methods use a "bag" model where only part appearance is considered whereas other methods learn spatial models but only given the output of a particular feature detector. Previous techniques for learning both part appearance and spatial relations have instead used a highly supervised learning process that provides substantial information about object part location. We show that our weakly supervised technique produces better results than these previous highly supervised methods. Moreover, we investigate the degree to which both richer spatial models and richer appearance models are helpful in improving recognition performance. Our results show that while both spatial and appearance information can be useful, the effect on performance depends substantially on the particular object class and on the difficulty of the test dataset.

1 Introduction

We consider the weakly supervised learning problem for object class recognition, in which we are given a set of positive exemplars that each contain at least one instance of a given object class, and a set of negative exemplars that generally do not contain instances of that class. We use an undirected graphical model (or Markov random field) representation scheme, where nodes of the graph correspond to local image regions that represent object parts, and edges connect pairs of nodes whose relative locations are constrained using a Gaussian model. This type of graphical model has recently been used for object class recognition by a number of researchers including [2, 5, 7, 8]. We use graphical structures that have small maximal clique sizes thus allowing for efficient exact discrete inference. Such structures include trees, star graphs and low tree-width fans (a generalized form of star graph with a central clique rather than a single central node).

We develop a new weakly supervised learning procedure for such models and demonstrate its performance for star-graph models (used in [7]) and fan models (used in [2]). Our learning method achieves better detection performance than these previous techniques on some common datasets. We formulate the learning problem as that of simultaneously estimating models of part appearance and spatial relationships between parts. This type of combined estimation approach has been used in previous *supervised* learning methods, where training data is labeled with part locations (e.g. [2, 5, 11]). However previous work on weakly supervised learning has generally solved a *data association* problem, where a feature detector is first run and then detected features are selected in order to form spatial relational models (e.g., [6–8]). We briefly discuss this related work in the following section. In contrast, our approach uses an EM procedure that iteratively improves both the appearance and spatial models. This procedure is computationally feasible due to the form of the underlying graphical models, which have small cliques and Gaussian spatial relational terms.

1.1 Related Work

The work presented here most closely relates to two current lines of research, both of which are concerned with learning probabilistic models of part appearance and spatial relations. The first line of research involves approaches that simultaneously estimate appearance and spatial parameters from training data using a maximum likelihood formulation (e.g, [2, 5, 11]). However these methods all rely on supervised learning procedures for which individual part locations are marked in the training data. The second line of related research involves approaches that require only weak supervision, where part locations are not provided in training (e.g, [6–8, 14]). However these methods can be viewed as learning spatial models given fixed appearance models, because particular feature detectors are first run to locate interest points. The subsequent learning process then involves forming a model that provides a consistent association to these detected features.

A number of other recent object class recognition techniques are also relevant to our approach, especially work on learning bag models. These models are collections of features or parts that do not explicitly include spatial information (e.g., [13, 3, 12]). Such models can still capture limited spatial information such as relative sizes of parts, and some fragment-based models encode information about overlap of parts at different scales [10]. Among these learning techniques there again is a dichotomy between those that are highly supervised but do not require feature detection (e.g., [13]) and those that rely on feature detectors to solve a data association problem (e.g., [3]). Both [10] and [12] are weakly supervised and do not use feature detectors, making them most similar to the approach we take here, although they do not explicitly model spatial relations.

2 Form of the Model

We use the undirected graphical model (Markov random field) framework in [2, 5, 7] where an object model $\Theta = (A, S)$ consists of appearance templates $A = (a_1, \ldots, a_m)$ for each part, and Gaussian spatial constraints $S = \{s_{ij}\}$ defined between certain pairs of parts. One can think of an underlying graph $G = (V, E)$ with a node $v_i \in V$ for each part and a corresponding appearance template a_i. A random variable l_i specifies the location of each part in some configuration space, and $L = (l_1, \ldots, l_m)$ denotes the overall configuration of an object with m parts (i.e., locations for all of the parts). An undirected edge $e_{ij} \in E$ corresponds to each pair of parts v_i and v_j for which there is a Gaussian constraint s_{ij} on the relative locations of those parts. The particular form of the appearance models a_i and the pairwise spatial constraints s_{ij} are described further below. Examples of some learned models are shown in Figure 2.

We now briefly turn to two important properties of these models. First, the likelihood of seeing an image given a configuration L of the model factors into a term for the background and a product over the individual parts of the model. That is, we assume the appearance of the parts is independent. Second, the prior probability of a configuration L, for a given model Θ, factors into a product of functions over maximal cliques (recall that a clique is a fully connected subset of nodes) of the graph,

$$P(L|\Theta) = \prod_C \Psi_C(L_C), \tag{1}$$

where each $C \subset V$ is a maximal clique, L_C denotes the location parameters corresponding to the vertices $v_i \in C$, and Ψ_C is some (non-negative) function of the location parameters. The utility of this factorization depends on the maximal cliques being small, as it allows the prior to be factored into a product of terms that are each over relatively small state spaces L_C rather than the full state space L. For instance in the case of trees (or star-graphs) the cliques are only size 2.

Taken together these two properties make it possible to efficiently compute the exact likelihood of an image x_n for a given model Θ, with a discrete set of possible locations L,

$$P(x_n|\Theta) = \sum_L P(x_n|\Theta, L)P(L|\Theta). \tag{2}$$

The precise running time is $O(mh^c)$ for a model with m parts, h possible locations per part, and where c is the size of the largest subset C. For Gaussian models this time can be reduced to $O(mh^{c-1})$ using the approximation methods in [5].

It is also common to use the maximizing configuration L^* to approximate (2) rather than summing over all values of L. This configuration is the maximum *a posteriori* location for a given model and image (the MAP estimate),

$$\arg\max_L P(L|x_n, \Theta) \tag{3}$$

Using the distance transform techniques introduced in [5] this MAP estimate can also be computed in $O(mh^{c-1})$ time. For clique sizes $c \leq 3$ these algorithms are quite fast in practice, using conservative pruning heuristics that guarantee the correct answer. While these fast inference procedures have previously been used for detection and localization, [2, 5, 7] here we use them as part of an unsupervised learning procedure that simultaneously estimates appearance and spatial parameters from training data.

2.1 Appearance Model

We use a simple oriented edge appearance template (as in [2]). Let I be the output of an oriented edge detector, so that at each pixel p, $I(p)$ has a value u indicating that either no edge is present or that there is an edge at one of a small fixed number of possible orientations. We model the appearance of the part i by an appearance template a_i. Let $f_i(p)[u]$ denote the probability that pixel $p \in a_i$ has value u. We assume these probabilities are independent given the location of the template.

As is common, we assume that the likelihood of an image given a particular model, as a function of location, is the product of two terms: one for absence of the model and one for presence of the model. When the model is absent we simply assume an independent background probability $b[u]$ for each pixel, yielding $\prod_p b[I(p)]$. When the model is present we assume that the individual part appearances are independent. Thus for a configuration L where the templates do not overlap,

$$p(I|\Theta, L) = \prod_p b[I(p)] \prod_{v_i \in V} g_i(I, l_i), \tag{4}$$

where

$$g_i(I, l_i) = \prod_{p \in T} \frac{f_i(p)[I(p + l_i)]}{b[I(p + l_i)]}. \tag{5}$$

Each term in g_i is the ratio of the foreground and background probabilities for a pixel that is covered by template a_i. In equation (4) the denominator of g_i cancels out the background model contribution for pixels that are under a part.

As long as we only consider configurations L without overlapping parts this likelihood is a true probability distribution over images (i.e., it integrates to one). When parts overlap it becomes an approximation, since evidence is overcounted for pixels under multiple templates. In [1] a patchworking operation was used that averages the probabilities of overlapping templates in computing $P(I|\Theta, L)$ in order to eliminate overcounting. We follow that approach here. However, to make computation tractable, we only apply this more accurate method to evaluating the likelihood of the best configuration L^* and not to the optimization used to estimate L^*.

2.2 Spatial Model

We use the fan models proposed in [2] because they include both bag models (e.g., [13]) and star-graph models (e.g., [7]) as special cases. A k-fan is a graph

with a central clique of k reference nodes, with the remaining $m-k$ non-reference nodes connected to all k reference nodes but to none of the other non-reference nodes. Figure 1 illustrates the structure of 1- and 2-fan models. A 1-fan has a single reference node, with all other nodes connected to that node but not to one another. In other words a 1-fan is a star-graph with a single central node, or equivalently, a tree of depth 1. A 2-fan replaces the single node in the center with a pair of nodes. These two reference nodes are connected to one another and to all the non-reference nodes, but there are no edges between non-reference nodes. When $k = m - 1$ the fan structure is a complete graph. At the other extreme, when $k = 0$ there are no edges, corresponding to a bag model with no spatial constraints between the parts.

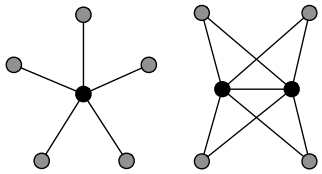

Fig. 1. Example 1- and 2-fans with reference nodes shown in black.

Let $R = \{v_1, \ldots, v_k\}$ be the reference parts of a k-fan and $L_R = (l_1, \ldots, l_k)$ be a particular configuration of these reference parts. Let $\overline{R}$ be the non-reference parts, $\overline{R} = V - R$. The spatial prior for a k-fan can be written in terms of conditional distributions as,

$$P(L|\Theta) = P(L_R|\Theta) \prod_{v_i \in \overline{R}} P(l_i|\Theta, L_R). \tag{6}$$

In this form it is apparent that the location of each non-reference part is independent when conditioned on the root parts.

For small k, this factorization meets our criterion in equation (1) of being a product over small cliques. This can be seen explicitly in the joint form,

$$P(L|\Theta) = \frac{\prod_{v_i \in \overline{R}} P(l_i, L_R|\Theta)}{P(L_R|\Theta)^{n-(k+1)}}. \tag{7}$$

where the denominator can be viewed as a normalization term based on the choice of reference set R.

For a Gaussian model the marginal distribution of any subset of variables is itself Gaussian. If μ_R and Σ_R are the mean and covariance for the locations of the reference parts then the marginal distribution of the reference parts together with one non-reference part v_i is given by the Gaussian with mean and covariance,

$$\mu_{i,R} = \begin{bmatrix} \mu_i \\ \mu_R \end{bmatrix}, \quad \Sigma_{i,R} = \begin{bmatrix} \Sigma_i & \Sigma_{iR} \\ \Sigma_{Ri} & \Sigma_R \end{bmatrix}. \tag{8}$$

These can be used to define the full spatial prior in terms of the above expressions for $P(L|\Theta)$.

3 Weakly Supervised Learning

Given a set of positive exemplar images, $D = (x_1, \ldots, x_N)$, each of which contains at least one instance of the object, it is customary to find a model Θ that maximizes the likelihood of the data,

$$\Theta^* = \arg\max_{\Theta} p(D|\Theta) = \arg\max_{\Theta} \prod_{n=1}^{N} P(x_n|\Theta).$$

In searching over possible models, evaluating $P(x_n|\Theta^t)$ for a particular model Θ^t and image x_n involves summing over the discrete space of possible model configurations for that image,

$$P(x_n|\Theta^t) = \sum_{L} P(x_n|\Theta^t, L)P(L|\Theta^t).$$

As we saw in Section 2, this can be solved efficiently because the model factors according to equations (1) and (4).

Maximum likelihood estimation problems that involve such hidden parameters can be solved using an expectation maximization (EM) algorithm, where a given model Θ^t is used to estimate values of the hidden variables L, which are then used to estimate an improved model Θ^{t+1}. In the current setting, there are two important characteristics that make EM particularly simple. First, both $P(x_n|\Theta^t)$ and corresponding optimal values of the location variable L_n^{t*} can be computed efficiently. In other settings such computations are often intractable, and much effort is devoted to finding good approximations that can be computed efficiently. Second, in our case we do not have a prior for the parameters Θ of the model (i.e., we are using a uniform prior over these parameters). In many applications of EM the prior over the parameters plays an important role in the optimization.

For a given model Θ^t, an optimal set of location parameters L_n^{t*} can be estimated for each image x_n either by computing the expected value of the location parameters or by computing the MAP estimate, as described for equations (2) and (3) above. For a given model Θ^t and image x_n, the MAP estimate of location can be interpreted as the best configuration of the model in the image. On the other hand the expectation might not correspond to any good configuration, if for example there are several instances of an object in the image. Given this natural interpretation of the maximizing configuration, we use the MAP estimate rather than the expected value for L_n^{t*}.

Given the set D of positive exemplar training images and a candidate model Θ^t we estimate the likelihood of the data given the model $P(D|\Theta^t)$ using the MAP location parameters L_n^{t*} for each image $x_n \in D$. Using these best locations, a new maximum likelihood model Θ^{t*} can easily be estimated using

the supervised training procedures in [5, 2]. To summarize, we have described a straightforward EM procedure for estimating the model Θ^* that maximizes the likelihood of the training data D, given some initial model $\Theta^0 = (A^0, S^0)$. We now discuss how to learn an initial model from weakly supervised training data.

4 Learning an Initial Model

The EM approach to learning object models described in the previous section requires an initial model $\Theta^0 = (A^0, S^0)$. Since EM is a local search technique, it is important to start with a reasonable initial model. Our approach is to compute a large set of candidate appearance templates that seem promising based only upon how well they individually discriminate between the positive and negative training data. Then we examine the configurations of those templates in the positive training data to both choose which candidates to include in the initial appearance model A^0 and to define an initial spatial model S^0.

4.1 Candidate Patch Models

As in [10, 12], our approach is to first generate a large set of potential appearance template models and then determine how well each such patch predicts the positive training examples compared to how well it predicts the negative training examples. Thus in addition to the positive exemplars D used for training the overall model, we also consider negative exemplars $\overline{D} = (\overline{x}_1, \ldots, \overline{x}_M)$, and we rank a given template a_i by the ratio of the likelihoods of the positive and negative training data,

$$\frac{P(D|a_i)}{P(\overline{D}|a_i)}. \tag{9}$$

We use the appearance templates discussed in Section 2.1 that specify the probability of an edge at each of several orientations at each pixel in the template. For the experiments in this paper, four quantized edge orientations were used: north-south, east-west, northeast-southwest, and northwest-southeast. Our initial set of candidate templates consists of patches drawn at random from the positive training images, sampled uniformly from several patch sizes and from all image locations such that the patches are contained within the image boundaries. We use three patch sizes: 12×12, 24×24, and 48×48 pixels. For the experiments reported in this paper we use approximately 100,000 initial patches. The edges in each patch are dilated in both the spatial and edge orientation dimensions in order to generalize the initial template from a single training example. We use a dilation radius of 2.5 pixels in the spatial dimension and 45 degrees in the orientation dimension.

To improve the quality of the templates, we employ a simple EM procedure, similar to the one discussed in Section 3 for learning the overall models. This procedure only maximizes the likelihood of the positive training data $P(D|a_i)$ rather than the ratio in (9), however in practice we observe that this also increases the ratio (and halts the optimization loop if it does not).

More formally, we are interested in maximizing

$$P(D|a_i) = \prod_n P(x_n|a_i)$$

where

$$P(x_n|a_i) = \sum_l P(x_n|a_i, l) \approx \max_l P(x_n|a_i, l).$$

As above, we use the maximizing location because it specifies the best location of the template in each image, whereas computing an expected location might not correspond to any one particular good match. For a given model a_i^t at iteration t we compute the maximizing location $l_{i,n}^{t*}$ for each image x_n. The resulting set of locations for the positive exemplars D can then be used to estimate an optimal template using the supervised learning procedure discussed above. The process is iterated until the likelihood ratio for the patch stops improving.

This optimization procedure is performed for each initial template. Due to the redundancy in the selection of the initial patches, there is generally considerable similarity between many of the resulting templates. However we do not attempt to cluster or otherwise collapse templates at this stage. All the resulting templates are ranked according to the likelihood ratio in (9). Templates with a low ratio are poor predictors of the positive exemplars over the negative exemplars, so patches with ratios below a threshold (e.g. 1.0) are discarded. All other templates are retained as candidate parts.

4.2 Pairwise Location of Candidate Patches

The previous step generates a very large set (e.g. tens of thousands) of candidate patches. It still remains to select some subset of these patches and to build an initial spatial model. In both selecting among patches and in modeling spatial relationships we want to take location information into account. For instance, it could be that a given template appears in both the positive and negative training examples, but in the positive examples it always appears at a particular location relative to other templates, making it a potentially predictive part of a model.

The simplest spatial relations are between pairs of patches, so we consider all pairs of candidate patches and form a Gaussian model of the relative patch locations $s_{ij} = (\mu_{ij}, \Sigma_{ij})$ for each pair. This is again a simple supervised learning procedure because for each template a_i and image x_n we have previously estimated the best location $l_{i,n}^*$. The mean and covariance of a Gaussian model of relative pairwise location are readily estimated by considering these locations for all the positive training images D and a given pair of templates.

Together with the appearance templates these spatial models yield a pairwise model $\theta_{ij} = (a_i, a_j, s_{ij})$ for each (unordered) pair of templates. These are just simple two-node instances of our more general models. For instance the likelihood of the training data given a model is just

$$P(D|\theta_{ij}) = \prod_n \sum_{l_i, l_j} P(x_n|a_i, l_i) P(x_n|a_j, l_j) P(l_i, l_j|s_{ij}). \tag{10}$$

This serves as a natural measure of the quality of a pair. As we have before, we approximate this using the maximizing parameter values l_i^*, l_j^* rather than summing over the parameters. In practice, we have found that the estimated locations for the parts can be noisy. In order to prevent the disproportionate influence of far outliers, we consider only the 90% of samples that best fit the spatial model $s_{i,j}$ (i.e., we compute a trimmed mean and covariance).

Some objects have two or more distinct parts that are similar in appearance. Examples include the two wheels of a bicycle and the two eyes of a human face. For these objects, the maximizing locations $l_{i,n}^*$ for a given patch a_i may correspond to one part in some images and another part in other images. As a result, the relative displacement between two patches may be a multimodal distribution to which it makes little sense to fit a Gaussian model. In these cases we have found it is better to fit a model to the strongest mode and ignore the rest of the distribution. The underlying idea is that with the high degree of redundancy in the patches, it is not necessary at this stage to explicitly handle patches that match at multiple locations. In practice, we handle this case by fitting a mixture of Gaussian model with a small number of mixture components when a single Gaussian is not a good a good fit. We then choose the highest-likelihood mixture component and use the mean and covariance of that component as the model of the pairwise relative location.

4.3 Initial *k*-fan model

We use a greedy procedure to construct an initial k-fan model for a given k. First consider the case of a 1-fan, in which cliques of the model are just the pairs constructed in the previous section. We exhaustively consider all the candidate patches identified in Section 4.1 as possible root parts (a 1-fan has just a single root part). For a given such choice of root part, a_r, we consider all other parts a_i, $i \neq r$, in order of their quality, ranked by the likelihood of the data given the pairwise model θ_{ri} in (10). Considering the pairs in this order, if a given part a_i does not overlap any of the other parts thus far in the model, then that part is added to the model. In practice a small degree of overlap is allowed. This greedy process continues until there are either no more parts left to add, or until some pre-determined maximal number of parts is reached. The result of this process is a set of parts for a potential model with root a_r. When repeated for each possible root part, a large set of candidate 1-fan models is generated.

This process differs only slightly for references sets of size $k > 1$. We consider all k-tuples of candidate patches rather than all singletons as possible reference sets. For each such reference set we as above greedily form a single model, where for a fixed reference set R all non-reference patches are considered in order, and added only if they do not overlap patches already in the model. The ordering in this case is determined according to the product of the pairwise scores in (10) for all the pairs of a reference patches with the current candidate patch, rather than just a single such score. Let θ_R denote the best model selected in this greedy fashion for each reference set R.

Each potential model θ_R (one corresponding to each possible choice of reference node) is scored in order to select one as the best initial model. Ideally we would like to use the likelihood of the positive exemplars given the model $P(D|\theta_R)$. However it is costly to evaluate this for the tens of thousands of candidate models. Instead we use a simple approximation: the product of all the individual part likelihoods and the product of the spatial priors for each connected pair of parts,

$$\prod_n \left(\prod_{v_i \in V_R} P(x_n|a_i, l^*_{i,n}) \prod_{(v_i,v_j)\in E_R} P(l^*_{i,n}, l^*_{j,n}|s_{ij}) \right), \tag{11}$$

where V_R and E_R are the set of nodes and edges of the model θ_R. Note that in the case of a 1-fan this quantity is the same as the true likelihood, because all the cliques are pairs of nodes. For other fan models, however, the true spatial prior is approximated as a product of pairwise spatial priors.

Finally we choose the model that maximizes (11). While the parts of this model form the initial appearance templates A^0, it is still necessary to create the initial Gaussian spatial models S^0 because the greedy model formation process considers only pairwise spatial models. This is done using the same simple supervised learning procedure by which the pairwise models were formed in Section 4.2 only now the true cliques of the k-fan model are considered rather than just pairs. This results in the initial model $\Theta^0 = (A^0, S^0)$ that is then improved using the EM procedure described previously in Section 3.

5 Experimental Results

In our first set of experiments, we applied our weakly supervised learning method to the image sets of the Caltech database [6]. Each of these image sets consists of 800 positive images and 800 negative (background) images, except for the faces set which contains 435 positive images. The positive and negative datasets were partitioned so that half of the images were used for training and the other half were held out for testing. Positive images were scaled so that object size was approximately uniform across the set of images. In these experiments we learned models that were limited to six parts, to facilitate comparison with earlier methods that also used six parts. Table 1 presents the results of these experiments, and compares the equal-ROC detection accuracy of our method to other recently reported results. The results are directly comparable because the image data and experimental protocol are identical across all of these tests. Figure 2 shows examples of the models that were learned for some of the Caltech object classes.

These results show that our weakly supervised learning method performs substantially better than the supervised results presented in [2], in which models were learned using hand-labeled locations for each part in each image. This is an encouraging result, as one might expect that carefully hand- labeled data would yield better performance. The results also show that our unsupervised learning

		0-fan	1-fan	2-fan	Results from literature
Motorbikes	unsupervised	**96.7%**	**98.6%**	**98.6%**	92.5% [6], 97.3% [7]
	supervised [2]	96.5%	97.0%	97.0%	
Airplanes	unsupervised	**90.3%**	**94.3%**	**95.0%**	90.2% [6], 93.6% [7]
	supervised [2]	90.5%	91.3%	93.3%	
Faces	unsupervised	**86.0%**	**98.0%**	**98.2%**	96.4% [6], 90.3% [7]
	supervised [2]	98.2%	98.2%	98.2%	
Cars (rear view)	unsupervised	**88.9%**	**94.4%**	**94.4%**	90.3% [6], 87.7% [7]

Table 1. Results of detection experiment on CalTech image set.

method produces better results than previous techniques that use a fixed set of feature detectors rather than simultaneously learning part appearance and spatial models [6, 7].

The results in Table 1 show that the detection accuracy for all classes increases substantially between the 0-fan models and the 1-fan models. There was also some improvement as k increased from 1 to 2 for the airplanes, but little or no improvement for the other classes. This suggests that for some objects and image sets, increasing the degree of spatial constraint (i.e. increasing k) in the object model improves detection performance whereas for other objects and image sets additional spatial information provides little or no benefit. In part this is due to the fact that the positive versus negative images in this database are highly different from one another, making it unnecessary to use spatial relationships to distinguish positive from negative.

We also tested the detection accuracy of the models learned by our unsupervised algorithm on the non-normalized version of the Caltech imageset, in which scale is not known. As in [2], we did this by applying the models at several different scales on each image and choosing the scale having the highest-likelihood detection. The equal-ROC points for the 0-fan and 1-fan models in this setting were 94.3% and 97.0% for motorbikes, 88.3% and 90.7% for airplanes, 85.7% and 98.0% for faces, and 86.0% and 93.5% for cars, respectively.

We also considered two more challenging datasets. The first of these is the Graz bicycle image set [9], consisting of 150 images with bicycles and 150 negative images. Unlike the Caltech data, many of the negative images in this set are similar to the positive images. The second is a hybrid set using the Caltech motorbike images as the positive images and the Graz bicycle data as the negative images. This is particularly challenging because many of the local features such as wheels and handlebars are quite similar between these two classes. As before, the images were partitioned into separate training and testing sets, and positive images were rescaled so that the object width was approximately uniform.

Table 2 presents the results of the experiments using the Graz bicycles data, showing equal-ROC detection results for 0-, 1- and 2-fan models consisting of 6 and of 25 parts. We considered the effect of adding more parts to the model because approaches that use a bag model generally use large numbers of features

or "parts" (e.g., [4, 13, 3]). The results show that both increasing the number of parts and increasing the degree of spatial constraint improve the performance. These results still do not quite achieve the accuracy of bag models on this dataset; for instance [4] report an equal-ROC rate of 88.0% and [9] report one of 86.5%, but they come closer than any other spatial models we are aware of.

	0-fan	1-fan	2-fan
6 parts	79.0%	81.0%	81.0%
25 parts	80.0%	84.0%	84.0%

Table 2. Results of detection experiment on Graz bicycle image set.

Table 3 presents the results of the experiments using the Caltech motorbike data with the Graz bicycle images as the negative test images. Again this table shows equal-ROC detection results for 0-, 1- and 2-fan models consisting of 6 and 25 parts. The most pronounced result is that for this data increasing k from 0 to 2 increased the equal-ROC detection results by about 6 percentage points (i.e., in going from a bag model with no explicit spatial constraints to a model with a moderate amount of spatial constraint).

	0-fan	1-fan	2-fan
6 parts	83.3%	88.1%	88.8%
25 parts	84.3%	89.3%	90.1%

Table 3. Results of detection experiment on motorbikes, with bicycles as background images.

The running time of the entire unsupervised learning process is approximately 24 hours on a small cluster of 20 Pentium III nodes. Note that the majority of this processing time is spent performing the correlations between the training images and the tens of thousands of candidate part templates. The results of this part of the process can be cached and reused when learning models for different values of k. Once the correlation computation has been performed, learning a new model requires approximately 1 hour on a single Pentium III node. Once a model has been learned, the average time required to localize an object in an image is approximately 0.1 seconds for a 0-fan, 0.3 seconds for a 1-fan, and 2.5 seconds for a 2-fan.

6 Summary

We have introduced a weakly supervised method of learning undirected graphical models for object class recognition. This method simultaneously estimates both

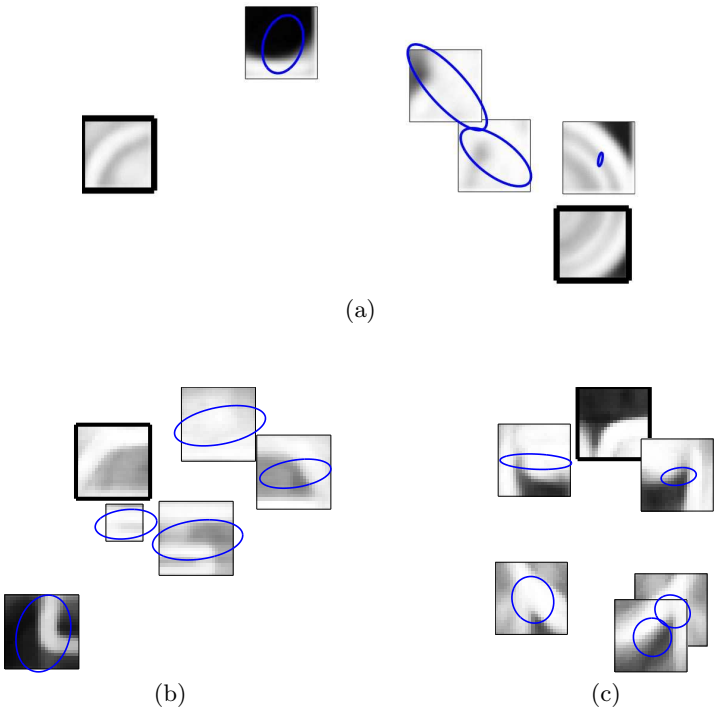

Fig. 2. Some models produced by our weakly supervised learning technique: (a) 2-fan motorbike model, (b) 1-fan rear-view car model, and (c) 1-fan face model. Reference parts are shown with a thick border. The spatial covariance with respect to these reference parts is illustrated with an ellipse. For simplicity, each template shows only the overall probability of an edge rather than the probability of each orientation.

part appearance and spatial relations between parts. In contrast, existing weakly supervised methods for learning these kinds of models rely on feature detectors rather than learning both appearance and spatial models from the data. Our method uses previously developed efficient inference and supervised learning algorithms to develop a simple and effective EM procedure. We have shown that our method produces better detection results on some standard datasets than are obtained by state-of-the art methods for learning such spatial models. We have also shown that for some problems, spatial information seems to be quite important in achieving high accuracy.

Our results, together with results of some other recent research, raise interesting questions about the role of feature detection in object class recognition. For bag models, with no explicit spatial information, very good detection performance is obtained both using feature detection (e.g., [4]) and by methods that do not use features (e.g., [13]). On the other hand, for spatial models such as the one used here, better results seem to be obtained by methods that do not

use feature detection. Two recent papers have demonstrated improved object detection results by not using feature detectors [2, 7]. In this paper we further demonstrate that better object detection results can be obtained by also not using features in the learning process, and instead learning appearance models together with spatial models. Another interesting set of open questions is raised by the fact that bag models currently perform better than spatial models for most common datasets. Our results suggest that this may partly be due to the datasets, but it remains to better characterize what aspects of bag models versus spatial models seem to account for these differences.

References

1. Y. Amit and A. Trouve. Pop: Patchwork of parts models for object recognition. Technical report, The University of Chicago, April 2005.
2. D.J. Crandall, P.F. Felzenszwalb, and D.P. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10–17, 2005.
3. C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *European Conference on Computer Vision*, 2004.
4. Gyuri Dorko and Cordelia Schmid. Object class recognition using discriminative local features. Technical report, INRIA Grenoble, September 2005.
5. P.F. Felzenszwalb and D.P. Huttenlocher. Efficient matching of pictorial structures. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages II:66–73, 2000.
6. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2003.
7. R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 380–387, 2005.
8. S. Ioffe and D.A. Forsyth. Probabilistic methods for finding people. *International Journal of Computer Vision*, 43(1):45–68, 2001.
9. A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *European Conference on Computer Vision*, pages 71–84, 2004.
10. E. Sali S. Ullman and M. Vidal-Naquet. A fragment-based approach to object representation and classification. In *4th International Workshop on Visual Form, IWVF4*, 2001.
11. H. Schneiderman and T. Kanade. Probabilistic formulation for object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1998.
12. T. Serre, L. Wolf, and T. Poggio. A new biologically motivated framework for robust object recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
13. J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *IEEE International Conference on Computer Vision*, 2005.
14. W. Zhang, B. Yu, D. Samaras, and G. Zelinsky. Object class recognition using multiple layer boosting with heterogeneous features. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.