Composite Models of Objects and Scenes for Category Recognition

David J. Crandall Daniel P. Huttenlocher
Department of Computer Science, Cornell University
Ithaca, NY 14853
{crandall,dph}@cs.cornell.edu

Abstract

This paper presents a method of learning and recognizing generic object categories using part-based spatial models. The models are multiscale, with a scene component that specifies relationships between the object and surrounding scene context, and an object component that specifies relationships between parts of the object. The underlying graphical model forms a tree structure, with a star topology for both the contextual and object components. A partially supervised paradigm is used for learning the models, where each training image is labeled with bounding boxes indicating the overall location of object instances, but parts or regions of the objects and scene are not specified. The parts, regions and spatial relationships are learned automatically. We demonstrate the method on the detection task on the PASCAL 2006 Visual Object Classes Challenge dataset, where objects must be correctly localized. Our results demonstrate better overall performance than those of previously reported techniques, in terms of the average precision measure used in the PASCAL detection evaluation. Our results also show that incorporating scene context into the models improves performance in comparison with not using such contextual information.

1. Introduction

In this paper we investigate the use of scene context in the recognition and localization of generic object classes in images. The past few years have seen substantial advances in the ability to *classify* images according to whether or not they contain instances of generic categories of objects (see related work in the next section). However many of these techniques do not *localize* where the objects are in the image. Rather these methods use a combination of cues in order to classify an image according to whether it contains instances of a given object, without explicitly determining object location. While image classification is well-suited to some tasks such as image retrieval, accurate object localization is essential for applications that involve interacting with the world such as navigation, surveillance, and visual user interfaces.

There are a number of reasons for the recent focus on image classification instead of localization. First, a broad range of machine learning techniques are directly applicable to the classification problem but less so to the localization problem. Second, large-scale training and test sets that provide object location information have not been available. The latter issue has recently been addressed with the creation of the dataset for the PASCAL 2006 Visual Object Classes Challenge (VOC) [7]. We address the same task using the same data and scoring as in the `comp3` portion of the competition. However we use a slightly different terminology that we believe is less open to confusion, referring to this task as *localization* rather than *detection*. This is because the term detection is often used in the literature to refer to methods that do not actually determine the location of objects in images (i.e., that perform what both we and the VOC refer to as *classification*).

The main focus of this paper is on the use of scene context to improve localization accuracy in comparison with not using contextual information. Recently there have been some impressive demonstrations of the power of scene context for classification and to a lesser extent for localization (e.g., [12, 15, 16, 17]). However, there has been relatively little investigation of the use of scene context for recognizing object categories that are highly confusable with one another, such as motorbikes and bicycles, and on large scale datasets that have also been used for evaluating non-context-based methods. Here we demonstrate that for our method the addition of scene context yields significant improvement in localization results for the bicycle, motorbike, car and bus categories of the PASCAL 2006 VOC dataset. Moreover, our method achieves better overall accuracy on these datasets than previously reported results from entries in the 2006 VOC challenge.

In this work we represent objects using part-based statistical models similar to those developed in [1, 3, 9]. We augment these models with scene context by creating a two-level hierarchical model, where the spatial configuration of

regions in the scene is represented at a coarse level, and the multi-part object is represented at a finer level of detail. We use a partially-supervised learning paradigm where the scene and object components of the model are both learned from training data that is annotated only with bounding boxes of entire object instances. The subparts of the object and the regions composing the scene, as well as spatial relations between them, are determined automatically, as illustrated in Figure 1.

2. Related work

Object recognition using models composed of image patches with pairwise spatial relations between them dates back at least to the Pictorial Structure models of Fischler and Elschlager [11]. Recently there has been a resurgence of such techniques, including constellation models (e.g., [2, 9]) as well as new techniques based on Pictorial Structures (e.g., [3, 10]). These lines of work have found that simple star-shaped graph topologies work well for such spatial models, and thus we employ star models here.

The past few years have also seen a resurgence of work on context-based recognition, which similarly is an area where research dates back to the 1970's. Whereas much of the early work on scene context sought to parse the scene, accounting for all the objects and relations between those objects, more recent work is generally aimed at using contextual information to improve recognition. For instance, information about the locations of sidewalks, roads or the horizon has been shown to improve localization of automobiles and pedestrians (e.g., [12, 17]) and the mutual presence of objects in an office scene can aid in recognition (e.g., [15]).

Thus far there has been relatively little evaluation of context-based recognition techniques on standard object recognition test sets that are also widely used for non-context-based methods. For example [12] consider the task of detecting cars using the PASCAL VOC 2005 data, but do not obtain state of the art results, achieving an average precision (APR) of 0.423 compared to 0.613 and 0.489 for the top two methods reported in that challenge. There has also been relatively little work on the use of scene context for multi-part spatial models. In contrast, previous work has tended to focus on rigid template object models (e.g., [12, 15]) or on pixel-level classification (e.g., [16]).

Finally we note that the dominant paradigm for research in object category recognition over the past several years has been classification – determining whether an image contains instances of specified categories – as opposed to localization. As an illustration, the PASCAL VOC 2006 had 23 entries in the `comp1` classification task and most of those entries gave results for all the object categories. In contrast, there were only 6 entries in the `comp3` localization task and only 2 of those gave results for all the categories. This is in part because the range of techniques applicable to classification is considerably broader than those applicable to localization; for instance bag models (e.g., [5]) do not incorporate spatial relations. While classification is an important task, localization is also highly important but less well studied, and hence the focus of this paper.

3. A Part-Based Object and Scene Model

Our approach is based on the probabilistic part-based models developed in [1, 3], where an object is represented in terms of a set of patches and spatial relations between those patches. As in the Patchwork of Parts (POP) model [1] we explicitly model overlapping patches, whereas most part-based models do not account for such overlap. We also constrain the underlying undirected graphical model defined by the parts and relations, so that the topology of the graph yields factored distributions for efficient computation (following the development in [3] more than [1]).

As our primary interest here is the use of scene context in object localization, we use a part-based model not only to represent an object in terms of constituent parts but also to represent the surrounding scene context. The form of the object and scene components of our model are the same, differing only in the types of parts that are used for the object versus the scene. More formally, a model consists of an object component and a scene component, each of which is composed of a collection of patches and spatial relations between pairs of those patches. Let $V^O = (v_1^O, \ldots, v_n^O)$ be the set of patches in the object model with pairwise spatial relations $E^O \subseteq V^O \times V^O$ among those parts. Similarly let the set of scene patches be $V^S = (v_1^S, \ldots, v_m^S)$ with pairwise spatial relations $E^S \subseteq V^S \times V^S$. In addition to the spatial relations between pairs of object patches and those between pairs of scene patches, there are also spatial relations between pairs that consist of one scene patch and one object patch, which we denote $E^{OS} \subseteq V^O \times V^S$. A model thus consists of the object and scene patches, the spatial relations within those sets of patches and the spatial relations between the two sets of patches $M = (V^O, V^S, E^O, E^S, E^{OS})$.

A *configuration* L of a model M with respect to an image I specifies a location in the image for each of the object and scene patches of the model; that is, $L = (l_1^O, \ldots, l_n^O, l_1^S, \ldots l_m^S)$ where each l_i^S specifies the location of scene patch v_i^S and l_i^O the location of object patch v_i^O. Note that these locations are mappings from a model coordinate frame to the image coordinates, and may involve transformations more complicated than simply a translation. While a configuration L specifies locations for each scene patch as well as each object patch, we consider only the locations of the object patches in localizing an instance of an object in an image (i.e., in placing a bounding box around the instance). The scene patches serve to constrain

the configuration of object patches via the spatial relations of the model, but are not themselves part of the object location.

For the object localization problem, we are interested in finding good configurations L of a particular model M with respect to an image I. That is, we seek configurations L that have high posterior probability, $P(L|I,M)$. A single best configuration can be found by identifying a configuration L^* that maximizes the posterior. However, more generally we seek high probability configurations. By Bayes' rule the posterior is proportional to the product of the prior probability over configurations and the likelihood of seeing the image given a configuration of the model,

$$P(L|I,M) \propto P(L|M)P(I|L,M). \quad (1)$$

Since we are seeking high posterior probability configurations, we do not need to consider the $P(I)$ term which would be in the denominator of Bayes' equality. We thus refer to $P(L|I,M)$ as the posterior even though it is only proportional to the actual posterior distribution.

This general form of distribution is intractable to compute even for a relatively small number of parts, particularly because the number of possible locations per part is quite large. Following the approach taken in [1] and [3] we consider a restricted form of problem such that both the likelihood $P(I|L,M)$ and the prior $P(L|M)$ factor into products of low-dimensional distributions. If the underlying graphical model $G = (V,E)$, with vertices $V = V^O \cup V^S$ and edges $E = E^O \cup E^S \cup E^{OS}$, forms a tree, the prior factors according to

$$P(L|M) = \frac{\prod_{e_{ij} \in E} P(l_i, l_j|M)}{\prod_{v_i \in V} P(l_i|M)^{\deg(v_i)-1}}.$$

In the current setting, we are modeling only relative locations of pairs of parts and not absolute part location. Thus the prior probability of individual part locations, $P(l_i|M)$, is uniform and the denominator can be omitted so that $P(L|M)$ is simply the product of the pairwise priors corresponding to the edges of the tree.

For the image likelihood it is common to simply assume that the appearances of the individual parts are independent from one another, such that the likelihood $P(I|L,M)$ factors into a product over parts. This yields an overall factorization of the posterior in (1) into

$$P(L|I,M) \propto \prod_{e_{ij} \in E} P(l_i, l_j|M) \prod_{v_i \in V} P(I|l_i, M). \quad (2)$$

For the combined object and scene models used here, this factorization applies to both the scene and object components of the model, yielding a factorization of the posterior into

$$P(L^O|M)P(L^S|M)P(L^{OS}|M)P(I|L^O,M)P(I|L^S,M) \quad (3)$$

where

$$\begin{aligned}
P(L^O|M) &= \prod_{e_{ij} \in E^O} P(l_i^O, l_j^O|M) \\
P(L^S|M) &= \prod_{e_{ij} \in E^S} P(l_i^S, l_j^S|M) \\
P(L^{OS}|M) &= \prod_{e_{ij} \in E^{OS}} P(l_i^O, l_j^S|M) \\
P(I|L^O,M) &\propto \prod_{v_i \in V^O} P(I|l_i^O, M) \\
P(I|L^S,M) &\propto \prod_{v_i \in V^S} P(I|l_i^S, M).
\end{aligned}$$

That is, the posterior distribution factors into products over the spatial priors of the scene model, spatial priors of the object model, and spatial priors between the scene and object models, as well as products over the part appearance likelihoods of the scene model and of the object model.

Each pairwise spatial model $P(l_i, l_j)$ is represented as a Gaussian over relative location of the two corresponding parts. These Gaussians can be over more complex configuration spaces than simply relative translation, for example including scale, orientation or other transformations. As discussed in more detail in Section 4 below, the appearance models for the object and the scene patches are different, yielding different forms of image likelihood. The object patches are high-resolution edge-based models whereas the scene patches are low-resolution color, edge and surface orientation models.

In the experiments reported below, we consider a somewhat limited form of the models introduced in this section. We learn object models and scene models that each form a star graph (a tree of depth one) rather than more general tree structures. In practice star graphs have been found to be quite powerful for modeling objects (e.g., [3, 10]). We also limit the form of the spatial relationship between the scene and object components of the model to a single edge between the root part of the object model and a special distinguished part in the scene model. The root part of the object model is learned automatically whereas the distinguished part of the scene model is the bounding box of the object rather than an arbitrary patch in the scene. In the future we plan to investigate other forms of tree structured models for both the object and scene. The simplified form used here was chosen because it is relatively straightforward to learn without much supervision, as discussed further in Section 6 below.

A given object category may consist of more than one model of the form described in this section. For instance, there may be models corresponding to different viewpoints, to distinctive sub-categories of objects, or even to different scene contexts. In the experiments reported in this paper,

we used one scene context model per object category, with one object model per viewpoint as defined by the labels in the training data.

4. Part Appearance Models

We use a simple probabilistic framework for modeling patch appearance that is general enough to handle a broad range of different types of image features. This approach is a generalization of the patch models in [3], which used only edges. In this framework a preprocessor is first run on the image to assign one of a small set of labels to each pixel. The preprocessor is chosen according to the type of image features of interest. For example, one preprocessor might be an edge detector that produces a binary edge map, while another preprocessor might examine local texture features and mark each pixel with one of a small set of possible texture labels. We use both the color quantization method in [14] and the surface orientation method in [13], illustrating the flexibility of this approach for handling a variety of image descriptors.

More formally we assume that the output J of the preprocessor is one of a small set of labels at each pixel, such that $J(p) \in \{1, 2, ..., r\}$ for each pixel location p. The foreground appearance model for object or scene patch i consists of a small template T_i that gives the probability of observing each possible label at each location within the template. That is, for part i the foreground appearance model is a function $f_i(p)[u]$ for all $p \in T_i$ and $u \in \{1, 2, ..., r\}$ specifying the probability of observing label u at location p within the template when the patch is centered over a true part location in an image. Another function $b[u]$ gives the probability of observing label u at a background pixel (i.e. a pixel in an image region not corresponding to an object part). Assuming that background pixels are independent, the probability that an image is composed entirely of background pixels (which we call hypothesis w_0) is,

$$P(J|w_0, M) = \prod_p b[J(p)].$$

For a given configuration of parts L in which no two patches overlap one another, we can write,

$$P(J|L, M) = P(J|w_0, M) \prod_{v_i \in V} \prod_{p \in T_i} h_i(p + l_i, l_i), \quad (4)$$

where

$$h_i(p, l_i) = \frac{f_i(p - l_i)[J(p)]}{b[J(p)]}. \quad (5)$$

Note that this likelihood function is a true probability distribution (it sums to one over all possible images) as long as no patches overlap in the configuration L. For part configurations with overlap, equation (4) overcounts some pixels and thus gives only an approximation. In Section 5 we describe how part overlap can be handled.

In using the appearance models for recognition, we compute the likelihood for each part v_i of the model at every possible configuration l_i. In other words we can think of each appearance model template T_i as a feature *operator* that when applied to an image produces a likelihood map over the entire configuration space. In contrast, many recognition techniques use feature *detectors* that first detect sparse feature or part locations and then only consider those locations. The approach that we take here does not make any such intermediate decisions, rather computing the entire factored posterior distribution in (3) using the full likelihood functions for all the parts. This complete approach was shown in [3, 10] to produce better results than approaches based on intermediate detection of features for models similar to the ones developed here.

We use low-resolution patches for the scene context models and high-resolution patches for the object models. Because the goal for the scene models is to represent general characteristics of the scene and not details of individual objects, the images are downsampled by a factor of 8 for the scene models.

The patch models make use of three different types of features that all fit into the appearance model framework just described. These features are:

- **Oriented edges:** Oriented edges are used both for the object models and for the scene models. In the object models, edges tend to capture local features such as corners and boundaries of object regions, while in scene models they capture features like the horizon, roads, buildings, etc. We use the Canny edge detection algorithm, apply morphological dilation with a radius of 2.5, and then quantize edge direction into four bins: north-south, east-west, northeast-southwest, and northwest-southeast. Thus in the general framework above, oriented edge models have five possible labels per pixel (four edge directions and the absence of an edge).

- **Color:** Color is an important feature for establishing scene context. For examples, bicycles are often observed above an area of green (grass) or gray (pavement) but rarely appear above an area of blue (sky or water). The color quantization algorithm in [14] is applied to label each pixel with one of ten basic color clusters, yielding ten possible labels for each pixel.

- **Surface orientation:** The surface orientation of regions around an object often provides useful contextual cues. For example, many objects like cars and bicycles are usually observed resting on a horizontal support surface such as a road. We use the surface orientation classification algorithm in [13] to classify each image pixel as one of three labels: ground, sky,

or vertical surface, yielding three possible labels for each pixel.

5. Handling overlapping parts

In the last section we assumed that part appearances were independent, which allowed the likelihood function $P(I|L,M)$ to factor into a product over parts. However this assumption is problematic for configurations where parts overlap, because pixels under overlapping templates are overcounted by equation (4). We address this issue with the same strategy used by the POP model [1]. The idea is that when an image pixel is covered by more than one patch, the multiple likelihood ratios from equation (5) corresponding to that pixel are averaged together. More formally, for a pixel p and a configuration L of patches, let $q(L,p)$ be the set of all patches that overlap p,

$$q(L,p) = \{v_i \,|\, p \in (T_i \oplus l_i)\},$$

where $T_i \oplus l_i$ denotes the transformation of the i-th patch T_i to location l_i, and let $Q(L)$ be the set of pixels covered by at least one patch,

$$Q(L) = \{p \,|\, q(L,p) \neq \emptyset\}.$$

Then the likelihood function in (4) can be rewritten in order to average the likelihoods of pixels covered by multiple patches,

$$\hat{P}(J|L,M) = P(J|w_0,M) \prod_{p \in Q(L)} \frac{\sum_{v_i \in q(L,p)} h_i(p,l_i)}{|q(L,p)|}, \tag{6}$$

where $|\cdot|$ denotes set cardinality. For configurations in which no patches overlap, equation (6) simplifies to the factored likelihood function of equation (4).

We use this POP model separately for the object likelihood $P(I|L^O,M)$ and the scene likelihood $P(I|L^S,M)$, thereby accounting for overlap of object parts with one another and of scene parts with one another. We do not, however, consider overlap of scene and object parts, as the scene parts are quite coarse and can have substantial overlap with the finer resolution object parts without measuring the same attributes of the image (i.e., without overcounting).

There is no known way of performing inference efficiently using the POP model. In [1] the maximum a posteriori (MAP) solution is estimated using the form of the posterior in (2) with the simple likelihood function $P(I|L,M)$ that factors. Then a hill climbing technique is used to maximize the form of the posterior that incorporates the POP likelihood $\hat{P}(I|L,M)$. In this paper we instead sample from the simpler factored posterior and then maximize the posterior that uses the POP likelihood over those samples. Sampling from the simpler form of the posterior can be performed efficiently using convolutions and dynamic programming [8].

6. Learning the Models

For learning both the scene and object context models we use a modification of the approach described in [4]. This technique requires only a set of images, each of which contains at least one exemplar of the object category. The patch models and spatial relations between patches are determined automatically.

We now briefly describe the learning procedure of [4]. There are four overall stages: (i) identifying candidate patches, (ii) finding predictable pairwise spatial relations between patches, (iii) forming an initial model from the patches and pairwise spatial relations, and (iv) updating the patch appearance and spatial models using an Expectation Maximization (E-M) algorithm. In the first stage a large number of patches of certain fixed sizes are sampled at random from the training images. These patches are then correlated with all of the training images to find patches that score well at some location in the vast majority of the images (i.e., patches that may be predictive of the category). In the second stage, pairwise Gaussian spatial models are formed from pairs of patches that are identified in the first stage. This is done by considering the best location of each patch in each image, and modeling pairs of relative locations. The quality of each such pairwise model is evaluated using the posterior probability in equation (2) for just the given pair of parts. In this manner each pair of patches is scored according to how predictable its relative pairwise spatial configuration is throughout the training data. In the third stage, a star-shaped object model is formed by choosing non-overlapping, high-likelihood pairs of patches using a greedy procedure. Finally, this initial model is refined using an iterative E-M algorithm that attempts to increase the likelihood of the training data given the model by updating both the appearance and spatial models.

For this paper we made some improvements to the learning procedure just described. First we modified the E-M step to handle part overlap more accurately by sampling from a proposal distribution and then scoring samples using the POP criterion, as discussed in Section 5. Second, in [4] MAP estimation was used as an approximation to computing expectations in several steps of the algorithm. Instead of using MAP estimates we draw samples from the distribution, which we have found gives better results without incurring a substantial extra cost, especially when the number of available training exemplars is small.

For the experiments presented in this paper, the object and scene context models were learned independently of one another using the approach summarized above. In learning the object model (V^O, E^O), only the regions inside object bounding boxes (as given by the ground truth) are considered by the learning algorithm. In learning the scene context model (V^S, E^S), whole images are processed, but the algorithm is constrained to produce a model

that includes the bounding box of the object as one of its patches. Finally the model for the edge connecting the object and scene models, E^{OS}, is learned by simply estimating a Gaussian on the relative location of the object model's reference part within the object bounding box. Since object bounding boxes were used during training, we refer to our learning procedure as *partially supervised*.

7. Localization experiments

In this section we present experimental results of our scene and context models.

7.1. Image dataset

For our experiments we used the images and ground truth data from the 2006 PASCAL Visual Object Classes (VOC) challenge [7]. This is probably the largest and most challenging object class recognition database publicly available. The database includes more than 5,000 images, mostly collected from image sharing websites on the Internet, and contains a variety of object categories in a wide variety of scene types. Object scale and viewpoint are completely unconstrained and most images have multiple objects and cluttered backgrounds. Ground truth data is also provided giving labels and bounding boxes for objects of interest in the image database. The ground truth also includes a rough viewpoint label (e.g. left side, right side, front, back) for some object instances. Ground truth is given for ten object classes, but the evaluation here considers four classes in particular: cars, buses, motorbikes, and bicycles.

7.2. Learning

For each object class we learned a single scene model and several object models corresponding to different viewpoints. For the bus, bicycle, and motorbike classes we learned four models each, corresponding to front, back, left and right views. For the car class we learned four additional models for the front-left, front-right, rear-left and rear-right in between views. The images and ground truth annotation data from the VOC `trainval` image set were used as the training set. As permitted by the VOC challenge rules, we added missing viewpoint annotations to the ground truth for some images. Object instances marked as "difficult" or "truncated" in the ground truth were not used during training.

The object and scene context models were learned using the partially-supervised learning algorithm described in Section 6. In learning the object models the training images were first scale-normalized so that the width of the object of interest was fixed. Candidate appearance model patches for the object model were generated by sampling rectangular regions of size 24×24 pixels from the positive training data. For the scene model, patches were sampled for each appearance model type (edges, colors, surface orientations) at three different patch sizes (12×12, 12×36, and 36×12 after subsampling). The background model for each type of appearance model was learned from the negative images (those not containing the object of interest). Figure 1 presents some sample models learned by our algorithm.

7.3. Localization

As previously noted the likelihood function in equation (4) is only an approximation for configurations in which the appearance model patches overlap. The POP criterion in equation (6) gives a better approximation $\hat{P}(I|L, M)$ of the likelihood function but as noted above no efficient inference algorithm is known. We therefore treat the factored posterior distribution in (3) as a proposal distribution and sample likely configurations from that distribution. Sampling from this distribution can be performed efficiently in time $O((m + n)h + (m + n)s)$ where n is number of object parts, m is the number of scene parts, h is the number of pixels in I and s is the number of samples to be drawn [8]. For each sampled configuration we then compute the posterior using the POP versions of the likelihood $\hat{P}(I|L^O, M)$ and $\hat{P}(I|L^S, M)$ rather than the factored versions $P(I|L^O, M)$ and $P(I|L^S, M)$ in (3). We use the resulting posterior probabilities, which account for patch overlap, in assessing the quality of each configuration.

Performing object localization on a given image I proceeds as follows. Recall that there are several object models for each object category, corresponding to different viewpoints. For each of these models M_i, we score samples according to the posterior using the POP model as just described. For the experiments reported here we used 20,000 samples per model and retained the top 10%. Because object scale is unconstrained, this process is repeated at multiple image scales; we used 32 scales in our experiments. For each image this produces a large set of candidate object configurations for each object category. To avoid duplicate and near-duplicate localizations, only the higher-likelihood configuration is retained when configurations significantly overlap. Note that this highlights an important advantage of our probabilistic models: localization confidences across different object and scene models are directly comparable. Finally object bounding boxes are computed for localizations with likelihoods above a threshold, which for the VOC competition is varied in order to compute a precision-recall curve.

7.4. Results

Figure 2 shows some sample localizations produced by our object and scene context models on the VOC challenge `test` dataset. In the figure, the configuration of the object parts is shown with the green rectangles while the object

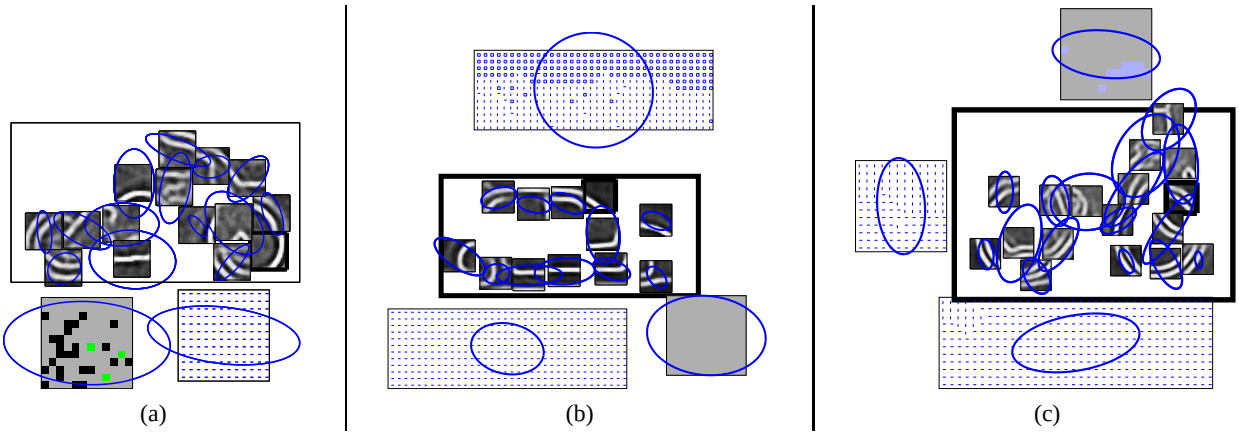


Figure 1. Sample models learned by our partially-supervised process: (a) motorbike side view, (b) car side view, (c) bicycle side view. Patches are drawn at the mean configuration with ellipses showing spatial covariance (at a 2σ level set). Thick outlines designate the root patches of the scene and object models. Simple illustrations of the appearance models are also shown. For the edge appearance models the probability of an edge is shown, with brighter pixels indicating higher probabilities. For the color and surface orientation features the mode (most probable label) is shown. For surface orientation patches, horizontal dashes represent ground, vertical lines represent vertical surfaces and boxes represent sky. Note that the actual appearance models have full probability distributions at each pixel, not a single label as is shown here for illustrative purposes.

bounding box is also shown. The color of the object bounding box indicates the object class that was detected. The images also illustrate the difficulty of the dataset, including very small objects like the bicycle in image (e) and the car in image (f) that contribute to the false negative rate.

We also conducted a quantitative evaluation of object localization performance using the rules of the 2006 PASCAL VOC competition's `comp3` localization task. Under these rules, a localization is considered correct if the detected object category label matches the ground truth and

$$\frac{\text{area}(B_p \cap B_{gt})}{\text{area}(B_p \cup B_{gt})} > 0.5,$$

where B_p is the localized bounding box and B_{gt} is the bounding box specified in the ground truth.

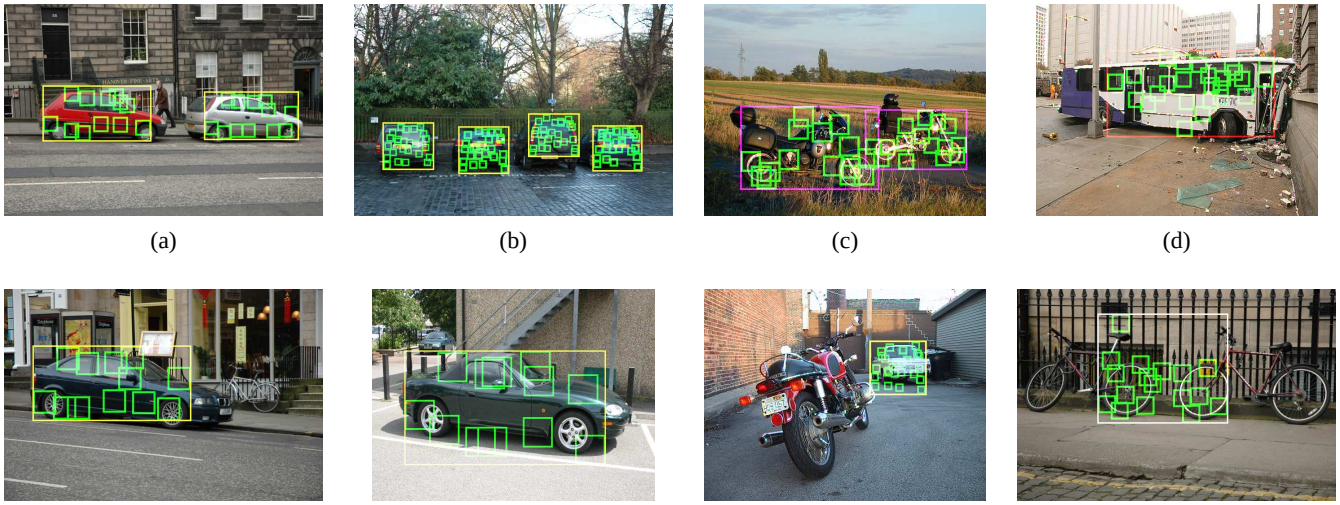
The VOC image test set contains several thousand images that have instances of objects from a number of categories. Unlike other common recognition test datasets, there are not separate sets for each object category. Each image may have several instances of one or more different object categories or may have no instances at all. Performance is measured using the "average precision" metric defined in the competition, which is the mean precision at the operating points corresponding to 11 prespecified recall values. As in the competition, object instances marked as "difficult" in the ground truth were ignored (neither contributing to false negatives nor false positives).

Table 1 presents average precision statistics on the challenge test data. As the table shows, including scene context models improved localization performance for all four classes when compared to using object models alone. To judge the statistical significance of this difference we ran the statistical test of DeLong et al. [6] on the area under the ROC curve (AUC) statistics. The improvements when adding scene context were statistically significant at a 99% confidence level.

Table 1 also shows the best average precision obtained by any of the entries in the 2006 PASCAL VOC challenge for each object class as reported in [7]. The results from our experiments are comparable to the VOC results because we used the same training and test dataset and conducted our experiments according to the rules of the competition. Our combined scene context and object models outperformed the best VOC results by a substantial margin for the bus, car, and bicycle classes. For the motorbike class our average precision was slightly lower, but the difference is probably not statistically significant. A common error mode was detecting a bicycle instead of a motorbike and vice-versa, as the visual difference between these two classes is often quite subtle. Moreover, unlike our method which performed uniformly well, none of the other methods entered in the competition performed well on all categories. For example, the algorithm that performed best on buses gave relatively poor results on bicycles and motorbikes.

Acknowledgments

This work was supported in part by a National Science Foundation grant (IIS 0629447) and a National Science Foundation Graduate Research Fellowship.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2. Sample output of our method on the VOC test dataset using object and scene context models: correct localizations in (a)-(d), correct localizations and false negatives in (e)-(g) (small bicycle in (e), distant car in (f) and motorbike at an unusual perspective in (g)), and a false positive and two false negatives in (h). Individual part localizations are shown by small green boxes and object bounding boxes are color-coded according to the localized object class: yellow for cars, red for buses, white for bicycles, purple for motorbikes.

Object class	Obj. model only	Scene + obj. model	Best VOC result
Bicycle	0.421	0.498	0.440
Bus	0.172	0.185	0.169
Car	0.429	0.458	0.444
Motorbike	0.342	0.388	0.390

Table 1. Results of localization experiments in terms of average precision (higher is better) on test images from the 2006 Pascal VOC challenge. For comparison, the third column shows the highest performance reported by any participant in the challenge [7].

References

[1] Y. Amit and A. Trouve. POP: Patchwork of parts models for object recognition. Technical report, The University of Chicago, April 2005.

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

[3] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *CVPR*, pages 10–17, 2005.

[4] D. Crandall and D. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *ECCV*, 2006.

[5] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *ECCV*, 2004.

[6] E. DeLong, D. DeLong, and D. Clarke-Pearson. Comparing the areas under two or more correlated ROC curves: a nonparametric approach. *Biometrics*, 44(3), 1998.

[7] M. Everingham, A. Zisserman, C. Williams, and L. V. Gool. The PASCAL Visual Object Classes challenge 2006 results. Technical report, September 2006.

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

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

[10] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *CVPR*, pages 380–387, 2005.

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

[12] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. In *CVPR*, pages 2137–2144, 2006.

[13] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, pages 654–661, 2005.

[14] J. Luo and D. Crandall. Robust color object detection using spatial-color joint probability functions. *IEEE Transactions on Image Processing*, 15(6):1443–1453, 2006.

[15] K. Murphy, A. Torralba, and W. Freeman. Graphical model for recognizing scenes and objects. In *Proceedings of NIPS*, 2003.

[16] J. Shotton, J. M. Winn, C. Rother, and A. Criminisi. *extonBoost*: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV (1)*, pages 1–15, 2006.

[17] A. B. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *ICCV*, pages 273–280, 2003.