Scene Recognition and Weakly Supervised Object Localization with Deformable Part-Based Models

Megha Pandey and Svetlana Lazebnik
Dept. of Computer Science, University of North Carolina at Chapel Hill
{megha,lazebnik}@cs.unc.edu

Abstract

Weakly supervised discovery of common visual structure in highly variable, cluttered images is a key problem in recognition. We address this problem using deformable part-based models (DPM's) with latent SVM training [6]. These models have been introduced for fully supervised training of object detectors, but we demonstrate that they are also capable of more open-ended learning of latent structure for such tasks as scene recognition and weakly supervised object localization. For scene recognition, DPM's can capture recurring visual elements and salient objects; in combination with standard global image features, they obtain state-of-the-art results on the MIT 67-category indoor scene dataset. For weakly supervised object localization, optimization over latent DPM parameters can discover the spatial extent of objects in cluttered training images without ground-truth bounding boxes. The resulting method outperforms a recent state-of-the-art weakly supervised object localization approach on the PASCAL-07 dataset.

1. Introduction

Weakly supervised discovery of common visual structure among a set of highly variable, cluttered images is one of the key problems in recognition. Consider, for example, the task of learning scene category models. While a few scene types ("beach," "mountain") can be well described by the statistics of low-level features, models for more complex and subtle categories ("nursery," "laundromat") should capture the appearance and spatial configuration of key scene elements – without being told what these elements might be or where they might be located. Another example is weakly supervised object localization, where we are given a set of images containing instances from the same category ("horse," "bus") and told to build a model for that category without knowing exactly where these instances are.

In this paper, we propose to represent the latent common structure of scenes and objects for the above tasks using deformable part-based models (DPM's) and to learn this structure using the *latent SVM* (LSVM) formulation of Felzenszwalb et al. [6]. DPM's currently constitute the state of the art for sliding-window object detection. A DPM represents an object by a lower-resolution root filter and a set of higher-resolution part filters arrenged in a flexible spatial configuration. In the standard (fully supervised) framework for training of an object detector, positive images are annotated with the locations of object bounding boxes, but the part locations are treated as latent information. The LSVM learning procedure acquires part appearance and layout parameters by alternating between making assignments to latent variables (part locations in training images) given the model parameters, and re-optimizing the model parameters given the latent variable assignments. This optimization framework has been very successful at discovering useful latent part structure in highly deformable categories with large intra-class appearance variability. In this paper, we push the limits of LSVM training by applying it to imagery with even more clutter and visual variability, and a significantly larger latent search space.

The first task we consider is scene recognition. Strictly speaking, scene categories do not have "parts" as objects do. However, as argued by Quattoni and Torralba [16], the structure of a scene may be described by a constellation model with a fixed "root" encompassing the entire image and moveable "regions of interest" (ROI's). The root captures the holistic perceptual properties of the entire scene, while the ROI's correspond to the most important objects. DPM's have exactly the right expressive power to implement this kind of model; moreover, the LSVM training process can be used to discover the ROI's automatically, whereas the method of [16] relies on manual annotations. The resulting scene representation, when combined with standard global image features such as GIST [14] and spatial pyramids [11] obtains state-of-the-art results on the MIT 67-category indoor scene dataset [16].

Our second target task is learning to localize objects from images that are annotated with category labels, but not with bounding boxes. In the fully supervised DPM training setup, root filters are initialized based on ground truth bounding boxes, though their locations are treated as

"partially latent" and allowed to move in a small neighborhood of the initial position to compensate for noisy annotation [6]. To deal with training images not having ground truth bounding boxes, we make the root filter locations fully latent and harness LSVM optimization to conduct a multi-stage global search for possible object locations. The resulting approach outperforms a state-of-the-art recent method [4] for weakly supervised object discovery on the challenging PASCAL-07 dataset.

2. Model Description

This section summarizes the DPM framework of [6], which we adapt to scene classification in Section 3 and weakly supervised object localization in Section 4.[1]

An image is represented by a multiscale feature pyramid. Specifically, a variation of histogram-of-gradient (HOG) [6] features is used. In our experiments, we partition the image at each pyramid level into cells of 8×8 pixels and use nine orientation bins per HOG cell. We use pyramids of eight and sixteen levels per octave for scene classification and object localization, respectively.

A DPM consists of a root filter, a set of part filters, and deformation parameters penalizing the deviation of the parts from their default locations relative to the root. Each filter defines a HOG window of a given size. The filter response at a given location and scale in the image is given by the dot product of the vector of filter weights and the HOG features of the corresponding window in the feature pyramid. The part filters are applied to features at twice the spatial resolution of the root. An object detection hypothesis x specifies the location of the root in the feature pyramid, and the positions of the parts relative to it are treated as latent variables z. The hypothesis is scored by the LSVM function

$$f_\beta(x) = \max_z \beta \cdot \Phi(x, z)\,, \tag{1}$$

where β is the vector of DPM parameters, i.e., a concatenation of all the filter and deformation weights, and $\Phi(x, z)$ is the concatenation of the HOG features of the root and part windows, as well as the part displacements. Note that it is the maximization over the latent variables z that makes the LSVM classifier response nonlinear. At detection time, the model score (1) has to be evaluated at every location and scale in the test image. To do this efficiently, the code of [6] relies on dynamic programming and generalized distance transforms [7, 8].

DPM's can be further extended to a mixture of multiple components. In this case, the component label of each hypothesis becomes an additional latent variable, and the model score is computed by maximizing over the scores of all the components.

[1] We use the code made available by the authors of [6] at http://people.cs.uchicago.edu/~pff/latent-release3/.

During training of the object models, the part locations and components are not labeled and hence are treated as latent (hidden) variables. The latent SVM training procedure alternates between two steps until convergence. In the first step, the parameters β are fixed and maximization over the latent variables of all the positive examples is carried out. In the second step, the latent variables are fixed and maximization over β is carried out by solving the margin-based SVM objective function.

Due to the presence of the latent variables, the LSVM training objective is not convex, and the model needs to have a good initialization in order to avoid local minima. In the implementation of [6], components are initialized by sorting ground-truth bounding boxes based on aspect ratio, root filters are initialized by training a standard SVM on the features inside the bounding boxes, and part filters are initialized by successively covering the highest-energy parts of the root filter (see [6] for details).

3. Scene Classification

Scene recognition approaches based on low-level appearance information [11, 14, 18] work poorly on categories that are characterized not by global perceptual characteristics, but by the identities and composition of constituent objects. To cope with such categories, Quattoni and Torralba [16] have proposed a representation composed of a root node capturing global scene properties and a set of ROI's capturing more fine-grained object-level properties. In this section, we use DPM's to obtain a representation with a similar expressive power but much higher performance than that of [16]. Moreover, while the method of [16] requires ground-truth ROI annotations to get the best performance, ours is able to discover them automatically.

3.1. Our Approach

We wish to adapt DPM's for multi-class scene classification in a one-vs-all framework, where we train a binary LSVM classifier for each class using images from all the other classes as negative data. At test time, we label the test image with the class getting the highest response.

At first glance, if we want the LSVM model to behave like a global image classifier, it would not seem to make sense to evaluate (1) at multiple location hypotheses per image. The root filter, which represents global scene characteristics, should be fixed to cover as much of the image as possible, and only the part filters should be allowed to move around to capture finer-scale deformable structure. In this scheme, when training the LSVM model for each scene type, each positive (resp. negative) image would generate a single positive (resp. negative) hypothesis.

Perhaps surprisingly, we have found that scene models trained in the above way do not perform well (they get only 17.6% accuracy), and that to get better results, we need to

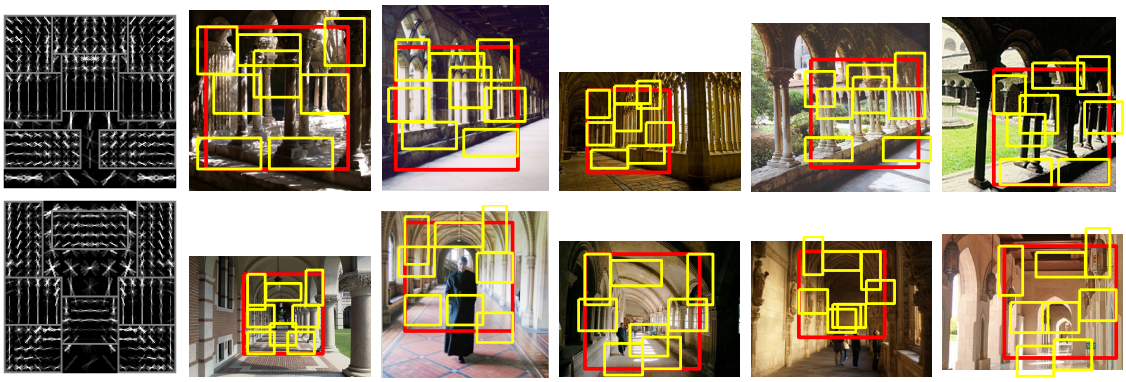


Figure 1. Use of a two-component model to represent different aspects of the category *cloister*. Left column: visualization of the root and part filters for both components. Right five columns: test images with "winning" root (resp. part) filter positions in red (resp. yellow).

allow the root filter to move around, albeit less freely than in an object detector. Specifically, we use a square root filter and restrict it to have at least 40% overlap with the image (this means that for a square image, the root filter covers over 60% of each dimension). In addition, we force the root filter to stay completely inside the image boundaries (in [6], the root filter can go outside to detect partially visible objects). At test time, we compute the classifier score for the image by maximizing (1) over all possible root filter hypotheses. Likewise, during training, we fix latent variable assignments in the positive images to the combination of root and part positions giving the highest response for the current model. We initialize the root filter weights by learning a standard linear SVM on the HOG features covering the entire training images. Part filters and placements are initialized using the same heuristics as in [6].

To get the best possible performance for scene classification, we have also found it necessary to sample multiple negative example windows from every negative training image, just as is done for object detector training in [6]. We sample all negative windows satisfying the same 40% overlap constraint as above. To make training with a large number of negative windows more efficient, the code of [6] takes a "data mining" approach of learning the model on a small subset of "hard" negatives. However, our negative window selection scheme is much more restrictive than a full sampling of windows in the HOG pyramid, so the overhead of the data mining outweighs its potential benefit. Thus, we turn off the data mining and use all the negative examples at once, reducing the training time by at least a factor of two.

The next question is how many part filters to use. Table 1 lists classification performance on the MIT indoor scene dataset [16] as the number of parts is varied from zero to ten. When we go from zero to two parts, we get a big leap in the classification performance from 15.00% to 25.37%, confirming that having a multi-scale latent structure is indeed key to the success of DPM's. We get the best performance with eight parts, so we use that number in all the subsequent experiments.

0	2	4	6	8	10
15.00	25.37	27.99	26.94	30.37	25.22

Table 1. Average classification rates (in %) for different numbers of parts on the MIT indoor scene database.

The final implementation choice concerns the number of mixture components in the model. We have found that two-component models are better able to deal with intra-class variability (see Figure 1 for an illustration). To initialize the corresponding model components, we cluster the training set into two groups based on GIST features [14]. During the training, the images are adaptively re-grouped depending on which component scores higher for that image. The two-component model achieves an average classification rate of 30.37%, compared to 28.43% of a single-component model.

3.2. Experiments

In this section, we evaluate our approach on the 67-category MIT indoor scene dataset [16] using the same training/test split as in [16], where each scene category has about 80 training and 20 test images.

Figure 2 shows the learned models for a few categories. DPM's do extremely well on categories with a stable global structure, such as *church_inside*, *cloister*, and *corridor*. They also do well on categories that can be distinguished on the basis of prominent objects. An obvious example of this is *movietheater*, whose DPM is essentially a screen detector. More interestingly, the model for *nursery* detects cribs with their characteristic vertical bars, the one for *laundromat* detects the round portholes on the doors of washers and dryers, the one for *meeting_room* detects a large table, and the one for *buffet* detects the curved edges of plates.

Table 2 compares our performance with a number of baselines and state-of-the-art approaches [12, 16, 19, 22]. By themselves, DPM's outperform a few recent approaches such as [16, 22], are competitive with GIST features [14] computed on the three color channels of the image, but do not do as well as spatial pyramids (SP) [11]. However, overall classification rates do not tell the whole story, as DPM's

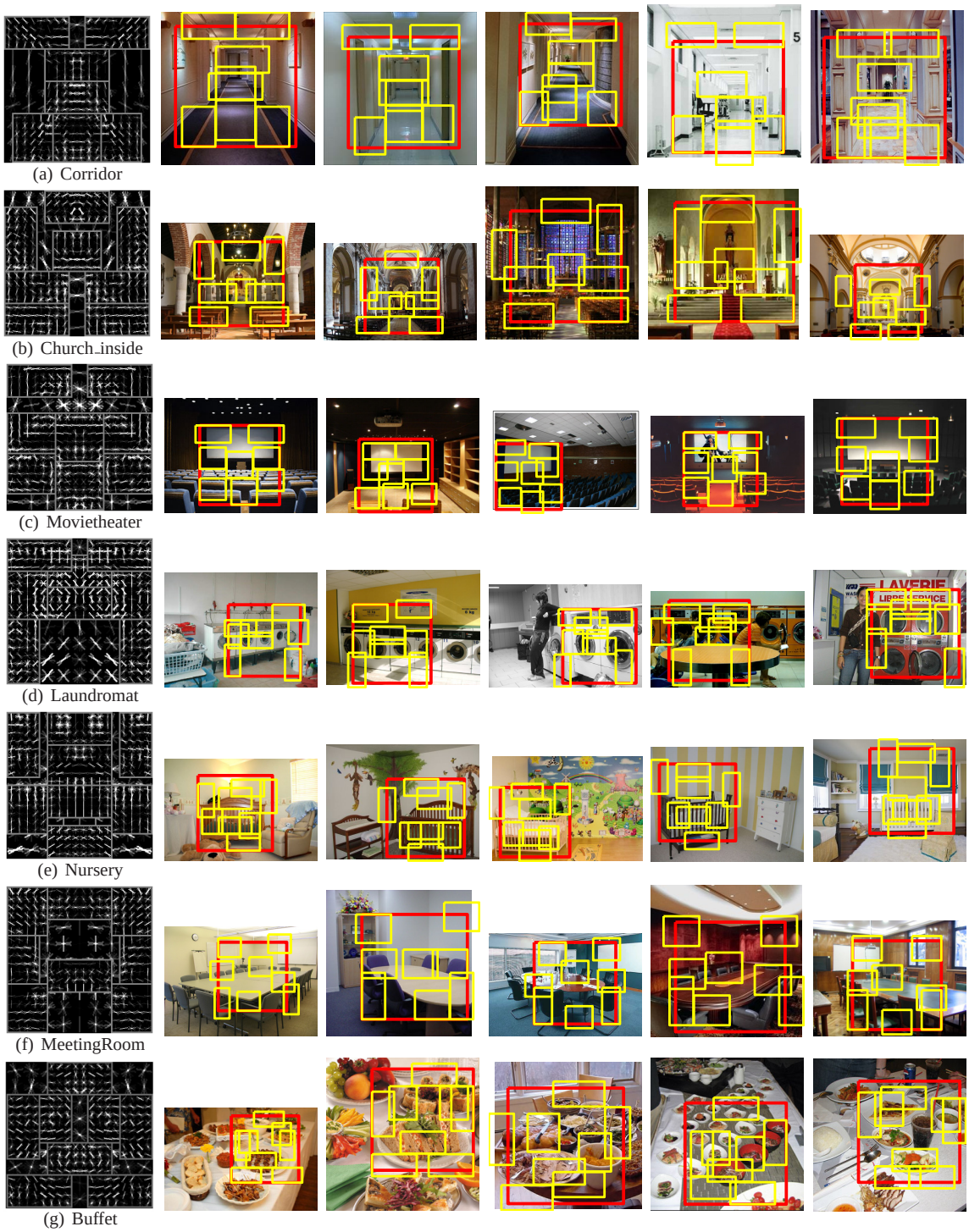


Figure 2. Scene models (only the dominant component) and example test images with the highest scoring filter positions. Detected root filter is displayed in red, and part filters are shown in yellow.

appear to complement existing feature representations in interesting ways. Table 3 lists the performance of our method, GIST-color, and SP on each of the 67 categories. There are quite a few classes such as *florist*, *bookstore*, *classroom*, *meeting_room*, *laundromat*, *nursery*, etc., where DPM's decisively outperform both SP and GIST-color, and for the most part these are the DPM's that also have the best qualitative structure. On the other hand, DPM's are relatively weaker on a few categories such as *poolinside*, *grocerystore*, and *winecellar*, for which color or local texture is more discriminative than global structure.

In order to benefit from the complementarity of DPM's

Baselines	HOG	22.8
	GIST-grayscale	22.0
	GIST-color	29.7
	Spatial Pyramid (SP)	34.4
	GIST-color + SP	38.5
State of the art	ROI+gist [16]	26.5
	MM-scene [22]	28.0
	CENTRIST [19]**	36.9
	Object Bank [12]	37.6
This paper	DPM	30.4
	DPM + GIST-color	39.0
	DPM + SP	40.5
	DPM + GIST-color + SP	43.1

Table 2. Average classification rates for MIT indoor scene dataset. **CENTRIST result is averaged over five random train-test splits, but all the other results use the split from [16]. For HOG, we use the dimension-reduced variant from [6], which for a 9×9 grid comes out to be 1395-dimensional. GIST-grayscale is a 320-dimensional descriptor [14] computed on the grayscale image. GIST-color is formed by concatenating the GIST descriptors of RGB color channels. SVM with a Gaussian kernel is used for the HOG and GIST baselines. For SP [11], we use vocabulary size 200 and three levels, and histogram intersection for the kernel. Multiple features (as in GIST-color + SP) are combined by multiplying softmax-transformed classifier outputs (see text).

and the other features, we use a very simple method to combine their respective classifier scores. Specifically, each feature gives us a bank of n one-vs-all classifiers for each of the n scene classes. If a test image gets scores $(a_1, \ldots, a_n)$ from one of these classifier banks, then the corresponding "confidence" that the image belongs to category i is given by the softmax transformation $\exp(a_i)/(\sum_{k=1}^{n} \exp(a_k))$. To get the combined "confidence" for class i based on all the available features, we multiply the respective softmax-transformed scores. As shown in the last line of Table 2, combining DPM, SP, and GIST-color in this way gives us an average classification performance of 43.08%, which, to our knowledge, is the best number on this dataset to date.

4. Weakly Supervised Object Localization

In this section, we present our approach for using DPM's to perform weakly supervised object localization. Most existing weakly supervised localization techniques have been applied to relatively simple datasets such as Caltech04 [1, 3, 13, 15, 21] or Weizmann horses [20], or one or two PASCAL-VOC categories [20, 21]. Fewer attempts have been made to learn models for a larger number of categories on more challenging datasets. Among these are [17] on the LabelMe dataset, [2] and [10] on PASCAL-VOC06, and [4] on PASCAL-VOC07. We compare our results to the state-of-the-art approach of [4], which has outperformed [2, 17]. This approach incorporates a "generic object model" that scores image windows according to their likelihood of being object bounding boxes, and that has to be learned from a set of "meta-training" images with ground truth object annotations. By contrast, the method we present does not use ground truth annotations at all.

4.1. Our Approach

The starting point for our method is the standard fully supervised training procedure for DPM detectors, which attempts to compensate for noisy or imprecise bounding box annotations by treating root filter positions in training images as "partially latent." Each root filter hypothesis in a positive training image is initialized based on the corresponding bounding box, but it is subsequently allowed to slide around in the neighborhood of that box to maximize the model score. In the weakly supervised scenario, we attempt to turn the root filter placements into full-blown latent variables and see if the LSVM optimization can successfully search the much larger latent space of potential object locations in the positive training images.

In order to avoid bad local minima in that space, we need to have a sensible starting point. In particular, we have found it difficult to learn a good model without initially constraining the root filter size. In the absence of a size constraint, the model tends to latch on to small regions that do not correspond to objects at all. To obtain initial estimates of object bounding boxes in positive training images of a given class, we essentially use the scene recognition approach of Section 3. Specifically, we begin by learning root filter weights from the HOG features of the entire training images, then we constrain root filters to have at least 40% overlap with the image and alternate between updating latent variable assignments (root and part locations) and DPM parameters that maximize the LSVM score.

Note that in Section 3 we only cared about root filter positioning to the extent that it improved the accuracy of scene classification. For that purpose, square root filters worked well. However, to achieve good performance for object localization, the estimated root filter positions have to closely match the ground truth bounding boxes. According to the PASCAL evaluation scheme, a localization is considered correct if the area of the intersection of the estimated and the ground truth bounding boxes divided by the area of their union is at least 0.5 [5]. It is hard to do well according to this criterion if the estimated root filter has the wrong aspect ratio. To date, we have not found a good method for determining this ratio from weakly annotated training data, so we simply initialize it to the average of the aspect ratios of the positive images.

At the end of the initial training stage, the single highest-scoring root filter placement in each positive image serves as the initial bounding box estimate. The 40% overlap threshold between the image and the root filter serves to reduce the latent search space, but it also poses a limitation for localizing smaller objects. In some cases, the poor lo-

	DPM	SP	GC	All		DPM	SP	GC	All		DPM	SP	GC	All		DPM	SP	GC	All
cloister	90	90	80	95	movietheater	45	50	25	55	dentaloffice	24	48	33	48	toystore	9	14	14	18
florist	79	63	63	89	closet	44	72	50	72	warehouse	24	14	24	29	children_room	6	11	17	11
buffet	75	70	50	80	inside_bus	43	57	48	57	computerroom	22	22	28	44	tv_studio	6	44	33	50
pantry	75	40	40	75	hairsalon	43	29	29	52	gym	22	22	11	33	deli	5	0	16	5
meeting_room	75	32	45	77	gameroom	40	20	10	35	livingroom	20	15	10	20	operating_room	5	21	26	26
classroom	67	56	39	61	prisoncell	40	35	35	50	grocerystore	19	48	43	48	airport_inside	5	10	5	10
concert_hall	65	55	60	80	subway	38	38	38	62	locker_room	19	38	5	38	artstudio	5	15	10	15
greenhouse	65	75	55	75	bowling	35	55	45	55	videostore	18	14	18	23	hospital_room	5	25	15	20
church_inside	63	68	74	79	stairscase	35	35	35	55	shoeshop	16	21	11	16	restaurant	5	25	0	10
inside_subway	62	43	10	52	trainstation	35	60	55	70	kindergarden	15	25	25	40	bedroom	5	14	0	10
nursery	60	45	50	65	clothingstore	33	33	11	33	winecellar	14	38	43	38	waitingrom	5	14	14	33
corridor	57	52	48	67	casino	32	47	32	47	museum	13	22	4	17	jewelleryshop	5	5	5	5
garage	56	50	28	56	studiomusic	32	58	42	63	fastfood_restaurant	12	12	18	24	laboratorywet	5	14	9	14
elevator	52	62	67	86	lobby	30	25	30	35	auditorium	11	44	22	33	restaurant_kitchen	4	22	17	13
bathroom	50	39	33	56	kitchen	29	24	43	52	bar	11	39	11	33	library	0	45	35	35
laundromat	45	23	18	50	dining_room	28	17	50	56	bakery	11	26	37	26	pool_inside	0	15	55	45
bookstore	45	25	20	35	mall	25	15	20	20	office	10	10	10	10					

Table 3. Per-class classification rates for our approach (DPM), spatial pyramid (SP), GIST-color (GC) and the combination of DPM + SP + GIST-color (All). The categories are listed in decreasing order of their DPM performance. All results in %.

calization is "obvious," in that a large bounding box ends up enclosing a mostly blank background region with a very small object instance in the middle (Figure 3).

To improve the localization in such "easy" examples and to obtain a more accurate estimate of the bounding box aspect ratio, we re-crop each bounding box by finding the area enclosing 99.9% of its edge energy using a modification of the technique from [9]. Briefly, we compute a low-resolution gradient magnitude image over the bounding box and set the values that are less the 10% of the maximum to zero. Starting from the centroid (center of mass) of the magnitude image, we expand the bounding box in four directions until the gradient magnitude inside it adds up to 99.9% of the total. This simple technique crops out plain background regions, allowing the bounding box to be a tighter fit around the object. However, it does not help for the images where the background is cluttred or textured. Figure 3 shows the result of bounding box re-cropping on a few images, and Table 4 shows the effect of this simple procedure on the accuracy of object localization.

Clearly, any correct localizations we manage at this stage are on the large, prominent, centered object instances – not just because of the overlap constraint, but also because root filter weights are initialized based on the HOG features of the entire images. Nevertheless, we hope that the "signal" in these instances overcomes the "noise" of the incorrect localizations to give us a reasonable starting model that can be subjected to iterative refinement. We re-train the model using the standard fully-supervised scheme of [6] with "partially latent" root filter positions. The only difference is that in [6] object bounding boxes come from the ground truth, while we use the re-cropped bounding box estimates from the automatic initialization step. We allow the root filter positions to move as long as they maintain at least 40% overlap with the input bounding box estimates. However, unlike the initialization, the re-training does not impose any constraint on the root filter size. In this manner, we can improve our localization of smaller object instances.

We repeat the re-training step several times using the bounding box estimates from the previous iteration as input, and re-crop the bounding boxes each time. With better bounding box estimates, the trained model improves further, giving higher localization results. Table 4 shows the localization performance at the end of each stage on two PASCAL07 subsets (see next section for details). After three rounds of re-training, the models converge to a stable level of performance.

4.2. Experiments

We follow the protocol of [4] by evaluating localization performance on two subsets from the training + validation set (*trainval*) of PASCAL07: PASCAL07-6x2 and PASCAL07-all [4]. The PASCAL07-6x2 subset consists of images from 6 classes (*aeroplane, bicycle, boat, bus, horse* and *motorbike*) for *Left* and *Right* aspects of each class, resulting in a total of 12 class/aspect combinations. The PASCAL07-all subset consists of 42 class/aspect combinations covering 14 classes and 5 aspects (*Left, Right, Frontal, Rear, Unspecified*). Just as in [4], for every class, the images labeled as either *difficult* and/or *truncated* were excluded from training and evaluation. To train a model for each aspect/class combination, we use the images from that aspect/class as positive training data, and images outside of that class as negative training data. For these models, we use only a single component, since separating the aspects reduces the amount of intra-class variability as well as the amount of positive training images.

Similarly to [4], we evaluate the accuracy of localizing instances of the target class in the *training* images. Note that our approach can localize multiple instances per image by applying the learned DPM model to the image in the usual sliding window fashion. However, the approach of [4] is restricted to a single detection per image, so to compare with them, we consider only the single highest-scoring window per image. The perfomance is measured as the percentage of training images in which an instance was correctly localized according to the PASCAL criterion (window-intersection-over-union ≥ 0.50). A breakdown of the results for each training iteration is given in Table 4. Our average performance on the PASCAL07-6x2 and PASCAL07-all subsets is 61.05% and 30.31% respectively, versus 50% and 26% for [4]. One should note that [4] uses

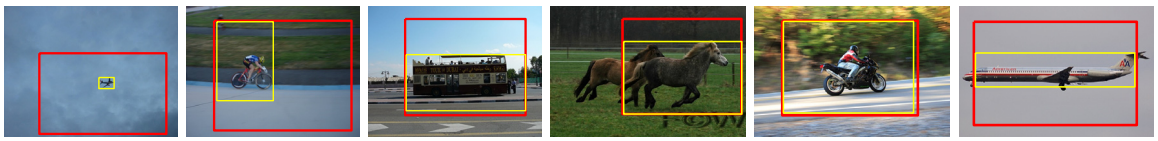


Figure 3. Bounding box re-cropping. Boxes before (resp. after) re-cropping are shown in red (resp. yellow).

a set of 799 images with bounding box annotations as meta-training data in order to learn the parameters of the generic object model, while we do not use ground truth annotations at all. On the other hand, once the generic object model is trained, the formulation of [4] learns the model for each class generatively (i.e., ignoring the images from all the other classes), while our approach trains the DPM models discriminatively (using the images outside the target class as negative data).

	PASCAL07-6x2		PASCAL07-all	
	Before cropping	After cropping	Before cropping	After cropping
Initialization	36.72	43.73	19.98	23.00
Refinement 1	51.63	53.11	25.11	26.38
Refinement 2	56.99	59.31	27.69	29.39
Refinement 3	59.32	61.05	28.98	30.31
Result from [4]	50.00		26.00	

Table 4. Average localization results (in %) for every stage of our iterative procedure.

Figure 4 visually compares the initial and final models obtained by our method for three classes. Both Figure 4 and Table 4 confirm that iteratively re-training the models and re-cropping the bounding boxes significantly improves the model quality and localization performance.

We have also experimented with weakly supervised learning of a model using as positive examples all the images of a given object regardless of their aspect. The images labeled as *difficult* and/or *truncated* are excluded in this case as well. Since we now need to model a mixture of viewpoints, we use a two-component model for this test. The components are initialized by sorting the training images according to their aspect ratio. The average localization performance of the resulting models for the fourteen PASCAL07-all classes is 29.98%, which is almost the same as that of the per-aspect models. Thus, the multi-component LSVM formulation is strong enough that we do not actually need to separate the aspects manually during training.

Finally, we apply the DPM's obtained through weakly supervised learning to detect objects in previously unseen test images. Table 5 compares the object detection performance for the PASCAL07-6x2 models to those of [4]. The performance is measured by the average precision (AP) on the entire PASCAL 2007 test set (4952 images). Our mean AP (mAP) is 0.208, compared to 0.160 from [4]. For reference, the mAP performance of DPM's learned with full supervision is 0.330 [4].

	Ours	[4]		Ours	[4]
aeroplane-left	0.075	0.091	aeroplane-right	0.211	0.236
bicycle-left	0.385	0.334	bicycle-right	0.448	0.494
boat-left	0.003	0.000	boat-right	0.005	0.000
bus-left	0.000	0.000	bus-right	0.030	0.164
horse-left	0.459	0.096	horse-right	0.173	0.091
motorbike-left	0.438	0.209	motorbike-right	0.272	0.161

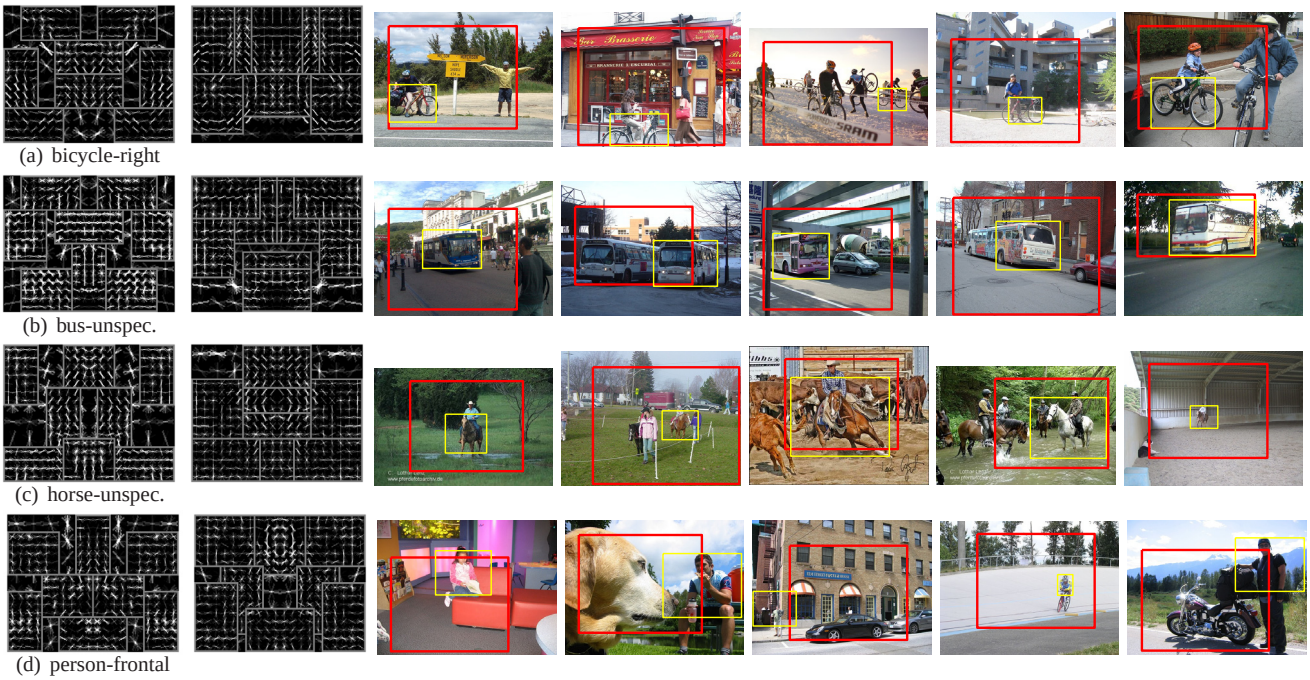
Table 5. Comparison of average precision for object detection on the PASCAL07-6x2 test set for our method vs. [4].

Even though the initial results presented in this section are encouraging, there remain glaring limitations and obvious avenues for improvement. One of the main limitations is the lack of a good method for initializing the aspect ratio of the root filter. We currently initialize it with the average aspect ratio of the positive images for the given class. However, the aspect ratio of the input images may not be a good indication of the object shape. One such example is our learned model for the *person-frontal* class (Figure 4 (d)), which is actually pretty good at locating people, but happens to have the wrong (horizontal) aspect ratio. For this reason, the bounding box estimate it returns often fails to satisfy the correct localization criterion.

5. Discussion

In Section 3, we used DPM's to learn the structural properties of indoor scenes in order to perform scene classification. By evaluating a multi-component model at different positions and scales, we were able to deal with changes in aspect and framing. Further, DPM models trained without any detailed object-level or ROI annotation can sometimes learn to identify common objects in the scenes. This ability makes them suitable for the problem of weakly supervised object localization as well. With a rather straightforward iterative refinement approach presented in Section 4, we were able to outperform a more complex state-of-the-art method [4] on the PASCAL-VOC07 dataset.

To summarize our contributions, we have demonstrated how the strengths of the DPM framework can be exploited to advance the state of the art in challenging recognition problems involving the discovery of latent correspondence among a set of cluttered, highly variable images. Another contribution is that, in showing the success of DPM's outside of their originally intended setting, for problems with a higher intra-class variability and a larger latent search space, we are able to give a better idea of their representational power and make an argument that they belong in the toolbox of the most effective *general-purpose* recognition methods available to date.



(a) bicycle-right
(b) bus-unspec.
(c) horse-unspec.
(d) person-frontal

Figure 4. Comparison of initial model (first column) with the final one (second column). The images compare the bounding boxes corresponding to these two models. Initial bounding box estimate is shown in red and the final one is shown in yellow. Re-cropping has been applied to the bounding boxes in both cases.

Acknowledgments. We would like to thank Joe Tighe for initially adapting the LSVM code to scene classification, and for continuing to lend his help throughout the project. This work was partially supported by NSF CAREER Award IIS 0845629, Microsoft Research Faculty Fellowship, Xerox, and the DARPA Computer Science Study Group.

References

[1] H. Arora, N. Loeff, D. A. Forsyth, and N. Ahuja. Unsupervised segmentation of objects using efficient learning. In *CVPR*, 2007. 5

[2] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *CVPR*, 2007. 5

[3] D. J. Crandall and D. P. Huttenlocher. Weakly supervised learning of part-based spatial models for visual object recognition. In *ECCV*, 2006. 5

[4] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *ECCV*, 2010. 2, 5, 6, 7

[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 5

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 2010. 1, 2, 3, 5, 6

[7] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. Technical report, Cornell Computing and Information Science, 2004. 2

[8] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005. 2

[9] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006. 6

[10] G. Kim and A. Torralba. Unsupervised Detection of Regions of Interest using Iterative Link Analysis. In *NIPS*, 2009. 5

[11] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 1, 2, 3, 5

[12] E. P. X. Li-Jia Li, Hao Su and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *NIPS*, 2010. 3, 5

[13] M. H. Nguyen, L. Torresani, F. De la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *ICCV*, 2009. 5

[14] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001. 1, 2, 3, 5

[15] A. Opelt and A. Pinz. Object localization with boosting and weak supervision for generic object recognition. In *SCIA*, 2005. 5

[16] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009. 1, 2, 3, 5

[17] B. C. Russell, W. T. Freeman, A. A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 5

[18] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *IEEE Workshop on Content-Based Access of Image and Video Databases*, 1998. 2

[19] J. Wu and J. M. Rehg. CENTRIST: a visual descriptor for scene categorization. *PAMI*, 2010. 3, 5

[20] X. Yang and L. J. Latecki. Weakly supervised shape based object detection with particle filter. In *ECCV*, 2010. 5

[21] Y. Zhang and T. Chen. Weakly supervised object recognition and localization with invariant high order features. In *BMVC*, 2010. 5

[22] J. Zhu, L.-J. Li, L. Fei-Fei, and E. P. Xing. Large margin learning of upstream scene understanding models. In *NIPS*, 2010. 3, 5