Feature selection, data association and spatial integration of cues for object recognition

Paper No.: 1315

Abstract

This paper presents an approach for object categorization and localization in a semi-supervised framework. It shows that sparse kernel methods and Bayesian learning techniques identify the most relevant sparse sets of local features for recognizing object classes. By adopting data association with constraints, our approach achieves performance comparable to the fully supervised setting. We extend our semi-supervised approach to a conditional random field that integrates multiple cues — superpixels and local interest regions — enabling robust localization of objects. Experimental results for object categorization show that our model outperforms existing methods. Experiments also report excellent quantitative results in localization.

1 Introduction

The proliferation of methods for extracting and describing salient and repeatable features [1, 2, 3, 4, 5] combined with recent advances in machine learning has fostered new and robust representations of object classes [6, 7, 8, 9, 10]. In this paper, we show that significant improvements in state-of-the-art local feature recognition systems can be achieved by adopting principled probabilistic models and semi-supervised learning techniques for automatically selecting and combining features from multiple interest region detectors. Using Bayesian learning techniques applied to kernel machines and conditional random fields, we select the most discriminative scale-invariant feature vectors and combine them with super-pixels. The resulting representations accurately detect and localize objects found in a wide variety of scenes at different poses and scales. Moreover, we learn with very little supervision from the user.

The first step in our recognition approach is to obtain a sparse set of *a priori* informative regions of the scene, thereby reducing the complexity of subsequent learning and inference. Local features bring tolerance to clutter, occlusions and deformations. Good detectors extract a sparse set of interest regions without sacrificing information content of the scene, and select the same regions when observed at different viewpoints and scales. In current literature, there are many definitions as to what constitutes a good scale-invariant interest region, grounded on maximizing disparate criteria. Therefore, we expect that using multiple detectors will provide complementary information, and hence improve recognition.

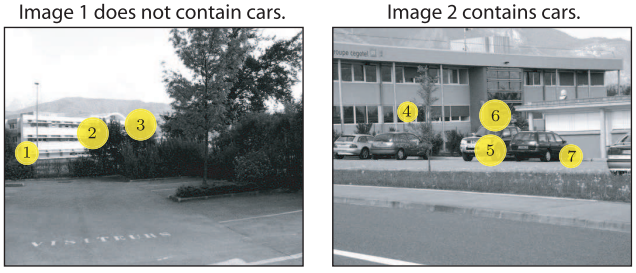


Figure 1: A couple annotated images selected from our car training set. The circles depict some of the features, along with their characteristic scale. The feature labels y_1 to y_3 in the first image are known. In the second image, we don't know the correspondence between the features and the labels, as indicated by the question marks on the y_i's. In this case, the correct correspondence would be $y_4 = -1, y_5 = 1, y_6 = -1, y_7 = 1$.

Interest point detectors extract regions based on local information, which is not enough to determine whether or not they are genuinely useful for distinguishing classes of objects. We need to learn which features are most discriminative based on some form of supervision from the user. Complete supervision requires the user to label individual features by segmenting the object from the background. Not only is this a time-consuming and poorly defined task, since people tend to segment scenes differently, it also inhibits us from exploiting large quantities of captioned images (available on the Internet in the form of news photos, for example [11]). Rather, we request labels on the images, which indicate the presence or absence of an object in each scene. Consider the case when we are learning to recognize cars provided the toy training set in Fig. 1. Each label on an interest region is a binary variable indicating whether it belongs to a car or to the background. The first caption indicates a car is not present in the scene, so the labels are set to the negative class. The second caption says there is

at least one car present, so learning requires determining the labels of the individual regions. This approach is called *semi-supervised learning by data association*, and was first introduced in [12]. When an object is present, we need to introduce constraints stating that we expect some of the features to belong to the positive class. The constraints introduce a new set of parameters, but we show that our model's performance is reasonably insensitive to these parameters, and in fact performs quite well with weak constraints. Semi-supervised learning with constraints can be considered a generalization of the multiple instance learning problem, as in [13]. Our model does not discretize the feature space, and handles large "bags" and constraints. Note we can achieve multi-category classification by combining responses from multiple binary classifier [14].

One might be skeptical that one can successfully learn to recognize objects in scenes from such weakly labeled data, given the high dimension of the features, the wide variability exhibited in the training images, and the fact that there are hundreds of unlabeled points per image! Previous work has demonstrated good performance in similar but less ambitious tasks using unsupervised [15, 16, 17] and semi-supervised frameworks [12]. Other publications use captioned images in order to learn representations of objects [6, 7, 9], but none of them explicitly perform data association. We hypothesize that exemplar selection is more effective when explicitly modeling the labels, since it allows the learning algorithm to exploit unlabeled background features instead of treating them as an inconvenience (i.e. noise). Experiments in Sec. 4.2 show that our data association scheme mostly compensates for the lack of unknown correspondences in the positive images.

We employ the augmented Bayesian classification model developed in [12] with an efficient Markov Chain Monte Carlo (MCMC) algorithm to simultaneously learn the unobserved labels and select a sparse object class representation from the high-dimensional descriptors. We introduce a generalized Gibbs sampler to explore the space of labels that satisfy the constraints. Bayesian learning comprehends approximation of the posterior distribution through integration of multiple hypotheses, not only a crucial ingredient for robust performance in noisy environments, but also resolves sensitivity to initialization. Unlike other methods, we do not reduce the dimension of the features through unsupervised techniques which may purge valuable information.

We propose a two-step approach that integrates different types of visual cues for localization of objects in Sec. 4.3. The first step involves training the Bayesian model to construct sparse representations based on interest regions and *superpixels*, segments that induce a low image compression [18]. Small segments are suitable for object-background separation, but the extracted features are hardly sufficient for the difficult problem of locating object classes in cluttered scenes. Second, it builds a probabilistic graphical model for each image which takes into account input from neighbouring superpixels and SIFT features. Spatial integration is achieved in a generic fashion, so we expect our localization applies to a variety of object classes.

2 Model specification

The training data consists of the set of labeled images $d = \{d_1, d_2, \ldots, d_D\}$. We represent image d_j by a set of feature vectors $\{x_i \,|\, i \in d_j\}$, and denote the entire set of N training features by $\boldsymbol{x} \triangleq \{x_1, x_2, \ldots, x_N\}$. We denote an observed label by y_i^k and an unknown label by y_i^u, where $y_i \in \{-1, 1\}$. When the image is labeled as being negative, the labels of the points are set to $y_i^k = -1$.

We implement the semi-supervised strategy of [12] to sample the labels y_i^u keeping in mind additional constraints on the number of observations per class. Such constraints play a crucial role in learning the unknown labels when we do not have any observed labels of a certain class. We define n_0 and n_1 to be the constraints on the minimum number of negative points and minimum number of positive points in a positively labeled document, respectively.

We use three detectors for generating observations $\boldsymbol{x}$ in a scene: the Harris-Laplace detector [2] finds corner-like features, the Kadir-Brady detector [4] proposes circular regions with maximum grey-level entropy, and the Laplacian method [3] focuses on regions of uniform intensity. Based on earlier studies [19], we choose the Scale Invariant Feature Transform (SIFT) [1] to describe the normalized regions extracted by the detectors. We compute each SIFT description using 8 orientations and a 4×4 grid, resulting in a 128-dimension feature vector x_i.

We also generate observations using superpixels, small segments that largely preserve image structure [18]. First, the image is segmented into 100 regions using the procedure detailed in [18], then 9-d features describing simple colour and texture properties are computed for each superpixel.

We use a sparse kernel machine to classify the features. The classification output depends on the feature being classified, x_i, and its relation to a subset of relevant features. The outputs of the classifier are mapped to class probabilities using a probit model [20]. That is,

$$p(y_i = 1 \,|\, x_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \Phi\Big(\beta_0 + \sum_{j=1}^{N} \gamma_j \beta_j k(x_i, x_j)\Big), \quad (1)$$

where the probit link $\Phi(\cdot)$ is the cumulative density function of the standard Normal distribution. By convention, researchers tend to adopt a logistic (sigmoidal) link function, but from a Bayesian computational point of view, the probit link has many advantages and is equally valid. $k(\cdot, \cdot)$ denotes the kernel. We used Gaussian kernels as they worked

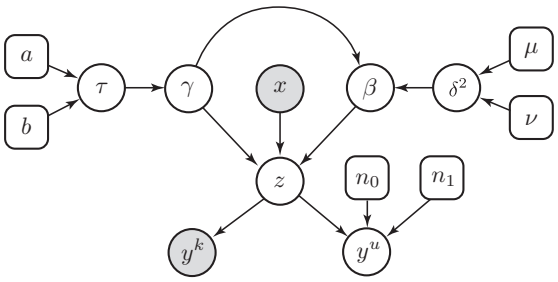


Figure 2: The directed graphical representation of the semi-supervised classification model. Shaded nodes are observed and square nodes are fixed hyperparameters. The figure does not show the dependencies between the y_i^u's.

well in our experiments, but other choices are possible. We introduce sparsity through a set of feature selection parameters $\boldsymbol{\gamma} \triangleq [\gamma_1, \gamma_2, \ldots, \gamma_N]$, where $\gamma_j \in \{0, 1\}$. Most of these binary variables will be zero and so the classification probability for feature vector x_i will only depend on a small subset of features. That is, by learning γ, we learn the relevant set of feature vectors, or prototypes, for each class. The vector of coefficients for the kernels is denoted by $\boldsymbol{\beta} \triangleq [\beta_0, \beta_1, \ldots, \beta_N]^T$. It is convenient to express (1) in matrix notation:

$$p(y_i = 1 \,|\, x_i, \beta, \gamma) = \Phi\left(\boldsymbol{K}_{i,\gamma}\boldsymbol{\beta}_\gamma\right),$$

where $\boldsymbol{K} \in \mathbb{R}^{N\times N}$ is the kernel matrix with entries $\boldsymbol{K}_{i,j} = k(x_i, x_j)$, $\boldsymbol{K}_{i,\gamma}$ is the ith row of the kernel matrix, with zeroed columns corresponding to inactive entries of $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}_\gamma$ is the reduced version of $\boldsymbol{\beta}$ containing only the coefficients of the active kernels.

We follow a hierarchical Bayesian strategy, where the unknown parameters $\{\boldsymbol{\gamma}, \boldsymbol{\beta}\}$ and missing labels $\boldsymbol{y}^u$ are drawn from appropriate prior distributions. We place a regularized maximum entropy g-prior on the regression coefficients $p(\boldsymbol{\beta} \,|\, \delta, \boldsymbol{\gamma}) = \mathcal{N}(0, \delta^2(\boldsymbol{K}_\gamma^T\boldsymbol{K}_\gamma + \epsilon I_K)^{-1})$, where ϵ is a small value that helps maintain a prior covariance with full rank. We assign an inverse Gamma prior to δ^2 with fixed hyperparameters μ, ν typically set to near-uninformative values. Each γ_j follows a Bernoulli distribution with success rate $\tau \in [0, 1]$, which in turn follows a Beta distribution with parameters $a, b \geq 1$. This allows the model to adapt to the data while allowing the user some control over the prior. Setting $b \gg a$ on large data sets initializes the Gibbs sampler to a reasonable number of active kernels. The model is summarized in Fig. 2. The intuition behind this hierarchical Bayesian approach is that by increasing the levels of inference, we can make the higher level priors increasingly more diffuse, or vague. That is, we avoid having to specify sensitive parameters and, therefore, are more likely to obtain results that are independent of parameter tuning.

The model is highly intractable. In particular, it is nonlinear and the posterior of the coefficients $\boldsymbol{\beta} \in \mathbb{R}^N$ is a correlated, hard to sample, high-dimensional distribution. However, we can simplify the problem enormously by augmenting the model with easy to sample low-dimensional variables $\boldsymbol{z}$. Then, by conditioning on the samples of these variables, we can solve for the posterior of $\boldsymbol{\beta}$ analytically. This is accomplished by ensuring that the variables $\boldsymbol{z} \triangleq \{z_1, z_2, \ldots, z_N\}$ have distribution $p(z_i \,|\, \boldsymbol{\gamma}, \boldsymbol{\beta}, x_i) = \mathcal{N}(\boldsymbol{K}_{i,\gamma}\boldsymbol{\beta}_\gamma, 1)$. It then follows that, conditioned on $\boldsymbol{z}$, the posterior of the high-dimensional coefficients $\boldsymbol{\beta}$ is a Gaussian distribution that can be obtained analytically. This augmentation strategy, which enables us to replace the original problem with the much simpler problem of sampling independent low-dimensional variables, was first introduced by Nobel Laureate Daniel McFadden [21].

Finally, to obtain the binary labels, we truncate $p(z_i \,|\, \gamma, \beta, x_i)$ appropriately, as in [20, 21]. The z_i^u's in a positively labeled document d_j must satisfy constraints on the minimum number of features of each class. The prior for document j, containing F features, is:

$$p(z_j^u | \boldsymbol{\gamma}, \boldsymbol{\beta}, \{x_i\}) \propto \left(\prod_{i=1}^{F} \mathcal{N}(\boldsymbol{K}_{i,\gamma}\boldsymbol{\beta}, 1)\right) \mathbb{I}_{C_0}(z_j^u)\, \mathbb{I}_{C_1}(z_j^u),$$

where C_0 is the set of assignments to y_i^u (and accordingly z_i^u) that obey the negative labels constraint n_0, C_1 is the set of assignments to y_i^u that satisfy the constraint n_1, and $\mathbb{I}_\Omega(\omega)$ is the set indicator: 1 if $\omega \in \Omega$, and 0 otherwise.

Discrete constraints in non-convex continuous optimization problems can be highly problematic. However, they are easily handled by MCMC algorithms.

3 Model computation

The classification objective is to estimate the density $p(y_{N+1} = 1 \,|\, x_{N+1}, \boldsymbol{x}, y^k)$ for an unseen point x_{N+1}, given the training data $\{\boldsymbol{x}, y^k\}$. Obtaining this probability requires a solution to an intractable integral, so we approximate it with the Monte Carlo point-mass estimate

$$p(y_{N+1} = 1 | x_{N+1}, \boldsymbol{x}, y^k) \approx 1 - \frac{1}{n_s}\sum_{s=1}^{n_s} \Phi\big(-\boldsymbol{K}_{N+1,\gamma^{(s)}}\boldsymbol{\beta}^{(s)}\big),$$

where n_s is the number of samples, and each sample $\{\boldsymbol{\gamma}^{(s)}, \boldsymbol{\beta}^{(s)}\}$ is distributed according to the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta} \,|\, \boldsymbol{x}, y^k)$. Kück *et al.* [12] develop an MCMC algorithm for sampling from the posterior by augmenting the original blocked Gibbs sampler [20] to the data association scenario. We follow their strategy for sampling these variables efficiently using Rao-Blackwellisation for variance reduction and the Morrison-Sherman lemma for fast matrix updates. One key difference is that [12] use rejection sampling to sample the unknown labels subject to the constraints, while we adopt a more efficient MCMC scheme and sample from the full conditionals in each document.

Figure 3: Images from the car data set. Notice no image contains only cars, and the cars vary considerably in size.

The full details of the MCMC algorithm can be found in our technical report [22].

4 Experiments

We conduct three sets of experiments, divided into three subsections. First, we measure the model's ability to detect the presence or absence of objects in scenes, comparing performance with previously proposed models. Second, we assess the model's capacity for learning the correct associations between local features and class labels by training the model with varying levels of supervision. Third, by integrating local feature and segmentation labels in a principled manner, we demonstrate reliable localization of objects.

4.1 Image classification

The experiments in this section quantify our model's capacity for identifying the presence or absence of objects in images. One should take caution, however, in generalizing the results to recognition: unless the image data is well-constructed, one cannot legitimately make the case that image classification is equivalent to object recognition. It is important to ensure the model learns to recognize cars, not objects associated with cars, such as stop signs. We address these concerns by proposing new experiment data consisting of images arising from the same environment, parking lots with and without cars. The outdoor scenes exhibit a significant amount of variation in scale, pose and lighting conditions. In addition, the new data set poses a challenge to learning with weak supervision, since the cars often occupy a small portion of the scene. See Fig. 3 for some example images. For purposes of comparison with other methods, we first present results on some existing data sets. The experiment data is summarized in Table 1.

We adopt a simple voting scheme for image classification by summing over the feature label probabilities assigned by the model.

Results of the image classification experiments are shown in Tables 2 and 3. The tables report performance with the Receiver Operating Characteristic (ROC) equal error rate, a standard evaluation criterion [8, 9]. It is defined to be the point on the ROC curve — obtained by varying the classification threshold — when the proportion of true positives is equal to the proportion of true negatives. Observe that our model in combination with the three detectors always produces the best image classification (when comparisons with other methods are available). Moreover, our model does very well in detecting cars in spite of the aforementioned challenges posed by the training examples. We omit error bars because independent MCMC trials with fixed priors exhibit little variance.

	Training images		Test images	
class	with object	without	with object	without
airplanes	400	450	400	450
motorbikes	400	450	400	450
wildcats	100	450	100	450
bikes	100	100	50	50
people	100	100	50	50
cars	50	50	29	21

Table 1: Summary of experiment data. The sources are the Caltech motorcycles (side) and airplanes (side) categories (http://www.vision.caltech.edu/html-files/archive.html), the Corel Image database for the Wildcats, and the Graz bikes and people data sets (http://www.emt.tugraz.at/∼pinz/data/GRAZ_01). The car data is available upon request.

data set	H-L	K-B	LoG	Combo	Fergus	Opelt
airplanes	0.985	0.993	0.938	0.998	0.902	0.889
motorbikes	0.988	0.998	0.983	1.000	0.925	0.922
wildcats	0.960	0.980	0.930	0.990	0.900	—
bikes	0.920	0.880	0.840	0.900	—	0.865
people	0.800	0.740	0.840	0.820	—	0.808
cars	0.966	0.897	0.897	0.931	—	—

Table 2: Image classification performance on test sets measured using the ROC equal error rate. The last two columns refer to the performance reported by Fergus *et al.* [9] and Opelt *et al.* [8]. All the other columns state the performance obtained using the proposed Bayesian model with regions extracted by various detectors (from left to right): Harris-Laplace [2], entropy detector [4], Laplacian of Gaussians [3], and combination of the three detectors.

We now explain how we obtained the results. For fair comparison, we adjusted the thresholds of all the detectors in order to obtain an average of 100 interest regions per training image. The combination scenario has an average of 300 detections per image. We compared performance on the airplanes, motorbikes, wildcats, bikes and people with the results obtained in [8, 9]. Note Fergus *et al.* [9] extract only 20 features per image on average, owing in part to the expense of training, while Opelt *et al.* [8] learn from several hundred regions per image.

In all our experiments, we fixed the label constraint n_0 to 0 and set n_1 between 15 and 30, depending on the ob-

data set	Combo	Superpixels
bikes	0.931	0.840
people	0.900	0.720
cars	0.820	0.966

Table 3: Image classification performance on test sets using the combination (see Table 2) and superpixel features.

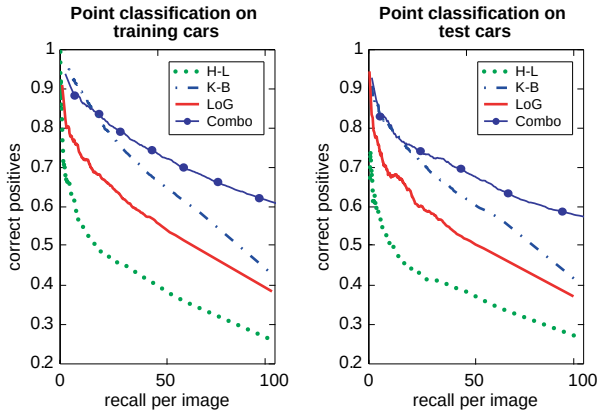


Figure 4: Plots of precision (percentage of correct positives) versus average recall per image for the task of labeling individual features as belonging to cars. Our algorithm learns which features are best in the combination, but this performance does not necessarily translate to better image classification (shown in Table 2). The Harris-Laplace detector is overly penalized in this performance measure because it often selects corner features that are near, but not on, cars.

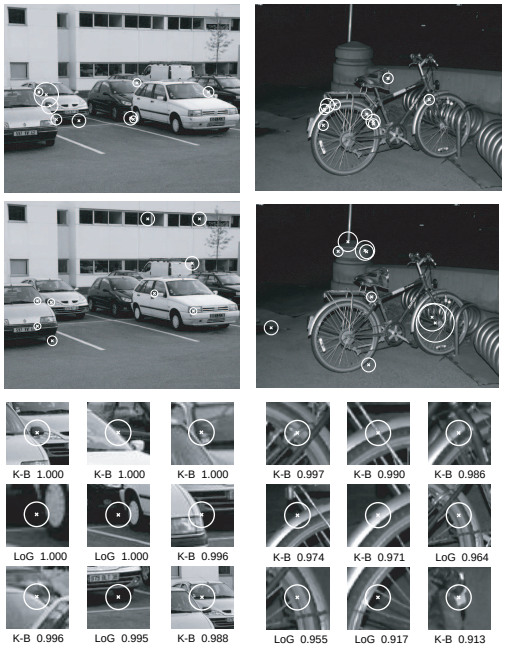


Figure 5: Two examples in which the combination of detectors (top row) results in improved detection over the Harris-Laplace detector (middle row). The circles represent the 9 interest regions that are most likely to belong to cars or bikes. The bottom row shows the top features along with feature type and probability of positive classification. The combination is an improvement precisely because the Harris-Laplace detector fails to select good features in these two images.

ject in question. Our constraints tend to be conservative, the advantage being that they do not force too many points to belong to objects that occupy only a small portion of the scene.[1] We set $a = 1$ and b according to a feature selection prior of approximately 200 active kernel centres, and bestowed near uninformative priors on the rest of the model parameters. In all our experiments, we set λ to $1/100$ because our MCMC algorithm reliably converged to a good solution. (Scale selection is an unsolved problem.) We found that 2000 MCMC samples with a burn-in period of 100 was sufficient for a stable approximation of the model posterior. Prediction by integrating the samples is fast: it takes about 1 s per image on a 2 GHz Pentium machine.

One of the more interesting results of Tables 2 and 3 is that no single detector dominates over the rest, highlighting the importance of having a wide variety of feature types for general object recognition. The colour and texture features extracted from superpixel segments are highly discriminative for cars, although it remains to be seen whether they would perform as well in a more diverse setting.

[1] In a series of separate trials on the cars, we varied n_1 between 10 and 100 and did not identify a relationship between image classification performance and constraint strength.

Training with the combination of the Harris-Laplace, Kadir-Brady and LoG detectors often, albeit inconsistently, improves the equal error rate. Further analysis of the combination scenario in Fig. 4 shows that it picks the best individual features first (according to hand segmentations of the scenes) so, clearly, performance at this task does not necessarily translate to good image classification. Fig. 5 shows a couple examples where the combination does result in an improved image classification.

We show examples of correctly and incorrectly classified images, along with the interest regions extracted by the detectors, in Fig. 6. Notice the different detection behaviour with a superpixel-trained model in Fig. 7. In both cases, incorrectly classified images were apart from those observed during training, such as the van and child's bike in Fig. 6. Problematic images also tended to exhibit challenging illumination conditions.

4.2 Investigation of data association

In this section, we ask to what extent our proposed scheme for data association correctly selects labels on the individ-

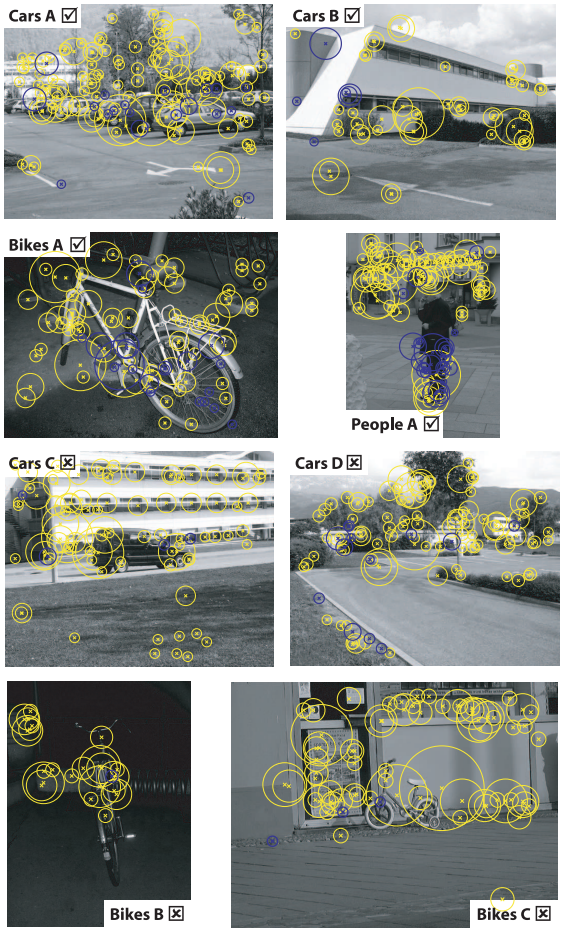


Figure 6: Test images correctly (top four images) and incorrectly (bottom four) classified using interest regions extracted by the Harris-Laplace (for cars and bikes) or LoG detector (for people). Dark blue circles represent local interest regions that are more likely to belong to the object, while yellow circles more probably belong to the background.

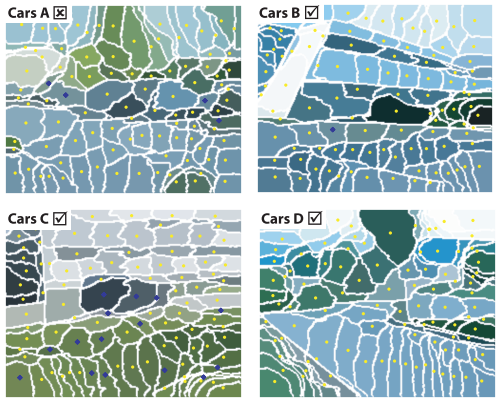


Figure 7: Car test examples classified using superpixels. A dark blue diamond indicates the segments is more likely than not to belong to the object, while a small yellow circle indicates that it probably belongs to the background. The misclassified car images are different than in Fig. 6.

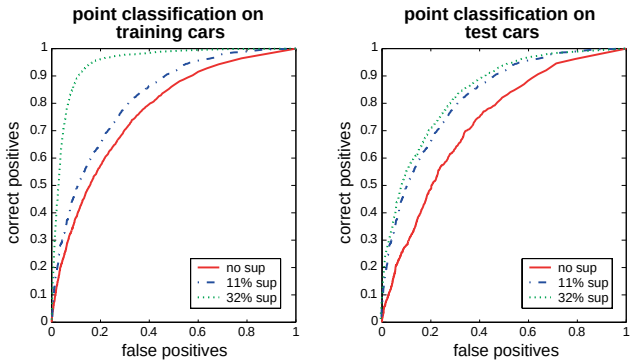


Figure 8: The ROC plots demonstrate how learning with different proportions of hand-labeled points affects performance on labeling individual car features. The Harris-Laplace detector is used for these experiments.

ual features, given that it is provided very little information. Another way to frame the question: if one were to provide partial manual segmentation, how would that influence performance? The answer depends on the nature and quality of the data, but the car data set is appropriate for such an inquiry since it appears to exhibit sufficient information to delineate positive and negative instances.

Experiment results on cars using the Harris-Laplace detector are shown in Fig. 8 — the ROC curves quantify accuracy in labeling individual features. As expected, the addition of a few hand-labeled points improves recognition on the car training set. However, further upgrades in supervision result in almost no gains in generalization to the test images. Fig. 9 demonstrates the effect of supervision.

4.3 Object localization

The previous section showed that the SIFT features lie mostly on the object, but they are inadequate for separating the object from the background. Here we add an additional layer to localize the objects in the image, motivated by the observation that regions marked by poorly discriminative superpixels can still capture useful information from SIFT descriptors (as demonstrated by comparing Figures 6 and 7). First, the model described in Sec. 2 is trained independently on the SIFT descriptors and the superpixels. The novel step is the construction a conditional random field [23] (CRF) that fuses label classification probabilities from both feature types for a final object localization.

Denoting $e_i^s, e_k^f \in \{0, 1\}$ to be the original label estimates, where the s superscript refers to superpixels and the f superscript refers to SIFT descriptors, and denoting each

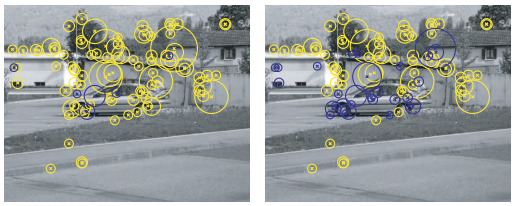
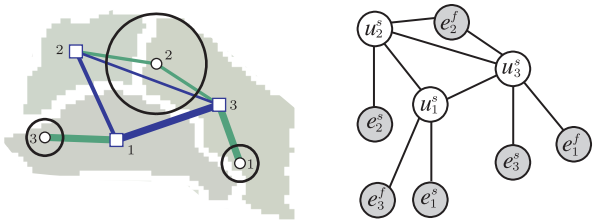


Figure 9: The figure shows the classification of individual interest regions along with their characteristic scale: dark blue circles are more likely than not to belong to the object, while light yellow circles more probably belong to the background. trained with various levels of supervision (see Fig. 8). *Left:* Car test image, no observed car labels during training. *Right:* The same image, except that the model was trained with an additional 11% observed feature labels.



(a) (b)

Figure 10: Conditional random field that integrates SIFT descriptors and superpixel label estimates. *(a)* The width of the blue lines depict the ϕ_{ij} edge potential strengths, which are determined according to the relative size of the shared border between the segments. Likewise, the width of a green line shows the strength of μ_{ik}, determined by the normalized area shared between interest region i (black circle) and segment k. *(b)* The undirected graphical model.

u_i^s to be the hidden (true) label of a segment, the joint of the labels u and e in an image is given by the probability model

$$p(u|e) = \frac{1}{Z(e)} \prod_i \phi(u_i^s, e_i^s) \prod_{(i,j)} \psi_{ij}(u_i^s, u_j^s) \prod_{(i,k)} \mu_{ik}(u_i^s, e_k^f),$$

where i ranges over the segments in the image, (i,j) ranges over neighbouring pairs of segments, (i,k) ranges over each interest region k overlaping with superpixel i, and $Z(e)$ is the partition function. The pairwise potential ψ_{ij} is the prior compatibility of the labels of neighbouring segments, μ_{ik} is the label compatibility for an overlapping segment and interest region, and ϕ describes the confidence in the original superpixel labels e_i^s. The potentials are given by

$$\phi(u_i^s, e_i^s) = \exp\{\delta(u_i^s = e_i^s)\theta_\phi + \delta(u_i^s \neq e_i^s)\}$$
$$\psi_{ij}(u_i^s, u_j^s) = \exp\left\{\left(1 + \tilde{b}_{ij}\right)\delta(u_i^s = u_j^s)\theta_\psi + \delta(u_i^s \neq u_j^s)\right\}$$
$$\mu_{ik}(u_i^s, e_k^f) = \exp\left\{\left(1 + \frac{a_{ik}}{a_k}\right)\delta(u_i^s = e_k^f)\theta_\mu + \delta(u_i^s \neq e_k^f)\right\},$$

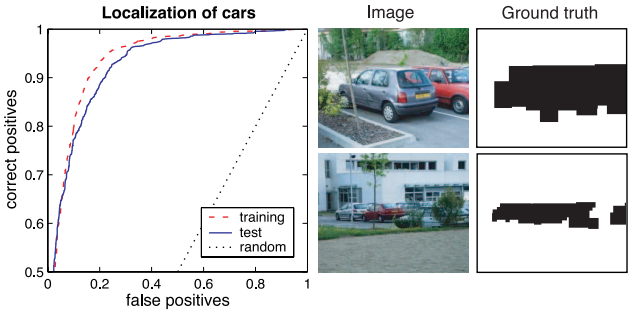


Figure 11: ROC plot evaluating localization of cars in training and test images by comparing estimated and ground truth segmentations. On the right are examples of ground truth segmentations.

where $\delta(\cdot)$ is the delta-Dirac indicator, $\tilde{b}_{ij} = b_{ij}/(2b_i) + b_{ij}/(2b_j)$ b_i is the contour length of segment i, b_{ij} is the length of the border between segments i and j, a_k is the area occupied by local interest region k, and a_{ik} is the area of overlap between segment i and interest region k. Fig. 10 describes the procedure for computing CRF potential tables. There are 3 free parameters, θ_ϕ, θ_ψ and θ_μ, which control the strength of the potentials. At this point, there is no learning; we tune the parameters by hand. The hidden labels are inferred by running loopy belief propagation [24].

Fig. 11 quantitatively compares the estimated segmentations with the ground truth segmentations (the ROC plot is obtained by thresholding the label probabilities). Some successful predictions are shown in Fig. 12, and some less successful car recognition results are displayed in Fig. 13. Localization failed when neighbouring superpixels and SIFT descriptors failed to complement each other with reliable label estimates. Notice we did not tailor the CRF to the car class, so there is reason to expect recognition performance generalizes to other visual object classes.

5 Conclusions

In this paper, we extended the discriminative power of local scale-invariant features using Bayesian learning. We showed that constrained semi-supervised learning using MCMC is remarkably well-behaved in the face of noisy high-dimensional features and wide variability in the unlabeled training data. Our method allows us to solve the important problem of selecting local features for classification. In addition, we proposed a generic, probabilistic method for robust object localization by integrating multiple visual cues learned through our model. One future extension to our learning framework is the incorporation of soft constraints through Bayesian priors, which adapt the number of positive labels according to the size of the object relative to the the background.

Figure 12: Good localization results on car test images. Darker patches are more likely to correspond to cars.

Figure 13: Some of the worst car localization results. Darker patches are more likely to belong to the car class.

References

[1] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[2] K. Mikolajczyk and C. Schmid. Selection of scale-invariant parts for object class recognition. In *ICCV*, 2001.

[3] T. Lindeberg. Feature detection with automatic scale selection. *IJCV*, 30(2), 1998.

[4] T. Kadir and M. Brady. Scale, saliency and image description. *IJCV*, 45(2):83–105, 2001.

[5] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, 2002.

[6] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV*, 2002.

[7] G. Dorkó and C. Schmid. Selection of scale invariant neighborhoods for object class recognition. In *ICCV*, 2003.

[8] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *ECCV*, 2004.

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

[10] K. Murphy, A. Torralba, and W. T. Freeman. Using the forest to see the trees: a graphical model relating features, objects, and scenes. *NIPS*, 2003.

[11] T. Miller, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. Learned-Miller, and D. A. Forsyth. Faces and names in the news. In *CVPR*, 2004.

[12] H. Kück, P. Carbonetto, and N. de Freitas. A Constrained semi-supervised learning approach to data association. In *ECCV*, 2004.

[13] Q. Tao, S. Scott, N. V. Vinodchandran, and T. T. Osugi. SVM-based generalized multiple-instance learning via approximate box counting. In *ICML*, 2004.

[14] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *JMLR*, 1:211–244, 2001.

[15] P. Duygulu, K. Barnard, N. de Freitas, and D. A. Forsyth. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.

[16] P. Carbonetto, N. de Freitas, P. Gustafson, and N. Thompson. Bayesian feature weighting for unsupervised learning, with application to object recognition. In *AI-Stats*, 2003.

[17] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Conf. on Research and Development in IR*, 2003.

[18] X. Ren and J. Malik. Learning a classification model for segmentation. In *ICCV*, 2003.

[19] K. Mikolajczyk and C. Schmid. A Performance evaluation of local descriptors. In *CVPR*, 2003.

[20] S. S. Tham, A. Doucet, and R. Kotagiri. Sparse Bayesian learning for regression and classification using Markov Chain Monte Carlo. In *ICML*, 2002.

[21] D. McFadden. A Method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57:995–1026, 1989.

[22] Anon. Removed as per guidelines. Technical report, 2004.

[23] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields. In *ICML*, 2001.

[24] K. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference. In *UAI*, 1999.