A Semi-Supervised Learning Approach to Object Recognition with Spatial Integration of Local Features and Segmentation Cues

Peter Carbonetto[1], Gyuri Dorkó[2], Cordelia Schmid[2], Hendrik Kück[1], and Nando de Freitas[1]

[1] University of British Columbia, Vancouver, Canada
{pcarbo,kueck,nando}@cs.ubc.ca
[2] INRIA Rhône-Alpes, Grenoble, France
{Gyuri.Dorko,Cordelia.Schmid@inrialpes.fr

Abstract. This chapter presents a principled way of formulating models for automatic local feature selection in object class recognition when there is little supervised data. Moreover, it discusses how one could formulate sensible spatial image context models using a conditional random field for integrating local features and segmentation cues (superpixels). By adopting sparse kernel methods and Bayesian model selection and data association, the proposed model identifies the most relevant sets of local features for recognizing object classes, achieves performance comparable to the fully supervised setting, and consistently outperforms existing methods for image classification.

1   Introduction

Over the past few years, researchers in high-level vision have shifted their focus from matching specific objects to the significantly more challenging problem of recognizing visual categories of objects. Since solutions exists to some image classification problems, there is a push to address more difficult problems such as object localization (segmenting an object from the background). There has also been success in learning robust representations of specific classes in constrained situations, notably frontal faces [1] and pedestrians in street scenes [2, 3], but models that can be trained to recognize generic object categories remain elusive.

A wealth of complementary developments in vision and machine learning have lead to improvements in general representations of object classes [4–7]. This paper furthers the state-of-the-art by adopting a principled probabilistic model for data association and model selection in object recognition. Our approach consists of the following three steps:

1. Extract a sparse set of *a priori* informative regions of the scene [5, 8], also called *keypoints* [9, 10]. Local interest regions bring tolerance to clutter, occlusion and deformable objects, and their sparsity reduces the complexity of subsequent learning and inference. Good detectors extract a sparse set

of interest regions without sacrificing information content, and select the same regions when observed at different viewpoints and scales. There exist many definitions as to what constitutes a good interest region, predicated on maximizing disparate criteria. Therefore, we expect that using multiple detectors will provide complementary information, and hence improve recognition. Sec. 6.1 describes how interest regions are extracted and represented as feature vectors.

2. Train the Bayesian classification model developed in [11] with an efficient Markov Chain Monte Carlo (MCMC) algorithm for Bayesian learning. The algorithm learns a sparse object class representation from the interest region descriptors, and does so with little supervision by explicitly modeling data association. See Sections 2-4 for more details.

3. For localization of objects, integrate two types of visual cues: interest regions and low-level segmentation using *superpixels* [12]. On their own, independent, local interest regions do not contain enough information to segment the object from the background, so we propose a simple conditional random field [13] that propagates information across neighbouring superpixels and weights the superpixel labels by the scores of overlapping interest regions. It is described in detail in Sec. 5.

The resulting representations accurately detect and locate objects in a wide variety of scenes at different poses and scales, even when training under very little supervision from the user.

We start with an example that illustrates the need for a model of data association in object recognition. After that, we motivate our proposed Bayesian hierarchical model for data association and object recognition.

1.1   A Case for data association in object recognition

Consider the toy training set in Fig. 1. It consists of three images, each with a caption indicating the presence or absence of cars in the scene. The circles depict some of the extracted features at their characteristic scale. The first image does not contain a car, so we can justifiably say that none of the circles are car features. In the second and third training images, however, we cannot conclude with certainty which features belong to a car. The conventional approach to this problem is to treat unlabeled features in the background as noise [4, 5, 7], an approach which degrades significantly when the object in question occupies only a small part of the unlabeled image, as in the second image. A more sensible strategy is to explicitly model the feature labels, allowing the learning algorithm to exploit the unlabeled background features instead of being hindered by them. This is precisely the solution we propose in this paper.

Each feature label is a binary variable indicating whether it belongs to a car (positive) or to the background (negative). In this setting, data association is closely related to the multiple instance learning problem [14, 15]. In the classical multiple instance formulation, a positive group label (the images are the groups) indicates that *at least one* of the individuals in the group has a positive label (this
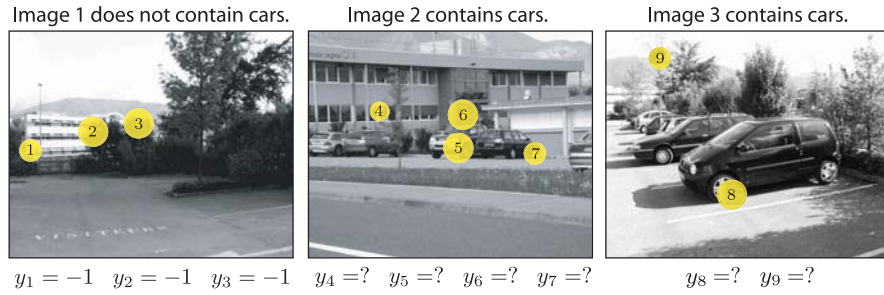
Image 1 does not contain cars.   Image 2 contains cars.   Image 3 contains cars.

$y_1 = -1 \quad y_2 = -1 \quad y_3 = -1 \quad y_4 =? \quad y_5 =? \quad y_6 =? \quad y_7 =? \qquad y_8 =? \quad y_9 =?$

Fig. 1. Three annotated images from the INRIA car training set. The circles represent some of the extracted features. The feature labels y_1 to y_3 in the first image are known. In the second and third images, we don't know the correspondence between the features and the labels, hence the question marks on the y_i's. Notice there is no image that contains only car features, and the size of the cars varies considerably. The correct correspondence is likely $y_4 = -1, y_5 = 1, y_6 = -1, y_7 = 1,$ $y_8 = 1, y_9 = -1$ (1 means "car" and -1 signifies "not car").

corresponds to a "contains cars" caption), while a negative group label implies that *all* individuals in the group have a negative label. For our purposes, this formulation is not sufficiently informative for learning the correct association, since an image may contain *hundreds* of unlabeled points and in the multiple instance setting only one of them is enforced to have a positive label.[3] We propose two alternatives. In the first, we introduce image-level constraints enforcing a certain number of the features to belong to the positive class.

The problem is that it may be hard to identify appropriate constraints. Referring back to Fig. 1, the cars in the third image occupy much more space than in the second, so the third image is likely to contain more features associated with the car class. The best we can do with hard constraints to set a conservative lower bound on the number of positives per image. We suggest a better route: specify a ratio that indicates the expected fraction of individuals with a positive label, along with a level of confidence in such an expectation. When objects vary widely in size, a low confidence on the expected fraction allows the model to adapt the number of positive labels to each image. We call this approach data association with group statistics. It was first proposed in [17].

One might be skeptical that it is possible to achieve recognition in this setting, given the wide variability exhibited in the training images, the high dimension of the features, and the fact that there are hundreds of unlabeled points per image. However, the alternative, complete supervision, is not only unappealing but also unrealistic for general object recognition problems. Complete supervision requires the user to annotate and segment objects from the background. This is not only a time-consuming task, but also poorly defined since people tend to

---

[3] Data association is also commonly studied as a case of *semi-supervised learning* [16]. This formulation is less compatible since it has no notion of groups.

Fig. 2. Two sample images from the MIT-CSAIL database [18]. Yellow lines indicate car annotations. The annotations are incomplete in both images, so learning from data association is still appropriate in the presence of annotated data.

segment scenes differently. It also inhibits exploitation of the vast quantities of captioned images available on the Internet (in the form of news photos, for example [19]). The experiments in Sec. 6.3 show that our data association scheme largely compensates for the lack of annotation data.

Even when annotations are provided, a recognition system might still benefit from multiple instance learning. Consider images from the MIT-CSAIL database [18], pain-stakingly annotated with more than 30 object classes, including cars, fire hydrants and coffee machines. Despite the effort in producing the scene labelings, the annotations shown in Fig. 2 are still far from complete. By learning the labels in the unannotated areas, the model can better exploit such training data.

There have been several previous attempts in tackling the problem of data association in object recognition, but they failed to extend to realistic domains. Duygulu *et al.* [20] studied the problem from the perspective of statistical machine translation. They formulated data association as a mixture model, using expectation maximization (EM) to learn the parameters and the unknown labels. Later, the translation model was extended to handle continuous image features [21] and spatial relations [22]. The problem with their approach is that the posterior over the parameters of the mixture model is highly multimodal, so EM tends to get stuck in local minima. The situation is no better when applying MCMC simulation techniques to mixture models, due to a factorial explosion in the number of modes [23]. More complex representations only exacerbate the issue, so mixture models are limited to simple, unimodal object classes. While [22, 21, 20] tackle multi-category classification, we can do likewise by combining responses from multiple binary classifiers [24]. Others have extended the multiple instance learning paradigm. We refer the reader to [17] for further references.

1.2   A Case for Bayesian learning in object recognition

We employ the augmented Bayesian classification model developed in [11] with an efficient Markov Chain Monte Carlo (MCMC) algorithm for Bayesian learning. The algorithm accomplishes two things simultaneously: 1.) it learns the unobserved labels, and 2.) it selects a sparse object class representation from the high-dimensional feature vectors of the interest regions. We introduce a generalized Gibbs sampler to explore the space of labels that satisfy the constraints or group statistics.

Bayesian learning comprehends approximation of the posterior distribution through integration of multiple hypotheses. This is a crucial ingredient for robust performance in noisy environments, and helps resolve sensitivity to initialization. In the presence of uncertainty about the labels, Bayesian learning allows us to be open about multiple possible interpretations, and is honest regarding its confidence in a hypothesis. The latter is of particular importance for integrating multiple visual cues for recognition (see Sec. 5), since it helps weigh the decisions of multiple models. The same cannot be said for learning through optimization of the model posterior, using EM for example.

Another advantage over other methods is that we do not need to reduce the dimension of the features through unsupervised techniques which may purge valuable information. Monte Carlo methods have received little attention in high-level vision, but our results show that they can be both effective and efficient in solving difficult problems.

In effect, what we describe is a *bag of keypoints* model [9] that chooses the features that best identify an object (e.g. the car model should select features that describe wheels or rear-view mirrors). It is widely appreciated that bag of keypoints methods — which treat individual features as being independent — are inadequate for identifying and locating objects in scenes (a person is not just an elbow!), and there has been much success in learning relations between parts [7] and global context [22, 18]. Despite these objections, independent parts models are not only efficient and simple to implement, but also remain the state-of-the-art in detection systems [9, 25] and, as we show, can function as a basis for more complex localization systems.

2   Bayesian kernel machine for classification

We start by assuming complete supervision. In other words, each data point x_i has a known label $y_i^k \in \{-1, 1\}$. The next section considers the case when some of the labels are unknown.

The training data consists of a set of D labeled images, and each image j, for $j = 1, 2, \ldots, D$, contains a set of exemplars or feature vectors $\{x_i \mid i \in d_j\}$. The set of exemplars for all the images used during training is $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$, where N is the total number of training exemplars. Sec. 6.1 describes how to obtain the feature vectors beginning with the raw pixel data.

We use a sparse kernel machine to classify the interest region descriptors. The classification output depends on the feature being classified, x_i, and its relation

to a subset of relevant exemplars. The outputs of the classifier are then mapped to the probability of the discrete labels using the probit link function. Following Tham, Doucet and Kotagiri [26], we have

$$p(y_i = 1 \mid x_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \Phi\left(f(x_i, \boldsymbol{\beta}, \boldsymbol{\gamma})\right), \tag{1}$$

where the unknown regression function f is given by

$$f(x_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{k=1}^{N} \gamma_k \beta_k \psi(x_i, x_k).$$

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.

The kernel function is denoted by ψ. We use the Gaussian kernel $\psi(x_i, x_k) = \exp(-(x_i - x_k)^2/\sigma)$ since it worked well in our experiments, but other choices are possible. We denote the vector of regression coefficients by $\boldsymbol{\beta} \triangleq [\beta_1 \ \beta_2 \ \cdots \ \beta_N]^T$. Our model is *discriminative* because it is specified as a conditional probability distribution of labels given observations, and not the other way around as in a *generative* mixture model.

We introduce sparsity through a set of feature selection parameters $\boldsymbol{\gamma} \triangleq [\gamma_1 \ \gamma_2 \ \cdots \ \gamma_N]$, where $\gamma_k \in \{0, 1\}$. 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 exemplars. By learning $\boldsymbol{\gamma}$, we learn the relevant set of feature vectors, or prototypes, for each class.

It is convenient to express (1) in matrix notation,

$$p(y_i = 1 \mid x_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \Phi\left(\boldsymbol{\Psi}_{i,\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right), \tag{2}$$

where $\boldsymbol{\Psi} \in \mathbb{R}^{N \times N}$ is the kernel matrix with entries $\boldsymbol{\Psi}_{i,k} = \psi(x_i, x_k)$, $\boldsymbol{\Psi}_{i,\boldsymbol{\gamma}}$ is the ith row of the kernel matrix with zeroed columns corresponding to inactive entries of $\boldsymbol{\gamma}$, and $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the reduced version of $\boldsymbol{\beta}$ containing only the coefficients of the active kernels. Thus, the vector product in (2) is shorthand for

$$\boldsymbol{\Psi}_{i,\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}} = \left[\, \psi(x_i, x_1)\beta_1 \ \psi(x_i, x_2)\beta_2 \ \cdots \ \psi(x_i, x_N)\beta_N \,\right].$$

We follow a hierarchical Bayesian strategy [27], where the unknown parameters $\{\boldsymbol{\gamma}, \boldsymbol{\beta}\}$ are drawn from appropriate prior distributions. The intuition behind this hierarchical approach is that by increasing the levels of inference, we can make the higher level priors increasingly more diffuse. That is, we avoid having to specify sensitive parameters and therefore are more likely to obtain results that are independent of parameter tuning.

We place a regularized maximum entropy g-prior on the regression coefficients $p(\boldsymbol{\beta} \mid \delta, \boldsymbol{\gamma}) = \mathcal{N}(0, \delta^2 \mathbf{S}_{\boldsymbol{\gamma}})$, where $\mathbf{S}_{\boldsymbol{\gamma}} = (\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T \boldsymbol{\Psi}_{\boldsymbol{\gamma}} + \epsilon I_N)^{-1})$ and ϵ is a small value that helps maintain a prior covariance with full rank. The regularization term δ^2 is in turn assigned an inverse Gamma prior with two hyperparameters $\frac{\mu}{2}, \frac{\nu}{2}$
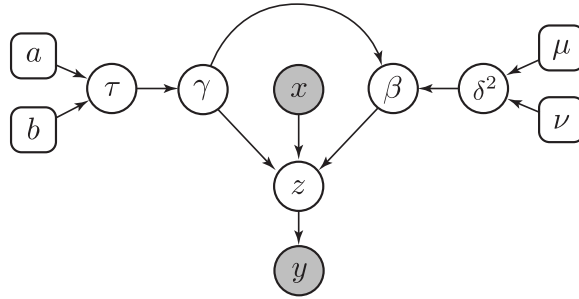
Fig. 3. The directed graphical representation of the fully-supervised classification model. Shaded nodes are observed during training, and square nodes are fixed hyperparameters.

specified by the user. One could argue that this is worse than the single parameter δ^2. However, the parameters of this hyperprior have much less direct influence than δ^2 itself, and therefore are less critical in determining the performance of the model [27]. Typically, we set μ and ν to near-uninformative values.

Following [11], each γ_k 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 data to automatically determine the complexity of the model according to the principle of Occam's razor, while allowing the user some control over the prior. Setting $b \gg a$ on large data sets initializes the learning algorithm to a reasonable number of active kernels.

The model is highly intractable. In particular, it is non-linear 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 introducing easy to sample low-dimensional variables $\boldsymbol{z}$. Then, by conditioning on the samples of these latent 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

$$p(z_i \,|\, \boldsymbol{\gamma}, \boldsymbol{\beta}, x_i) = \mathcal{N}(\boldsymbol{\Psi}_{i,\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1). \tag{3}$$

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 simple trick, first introduced by Nobel Laureate Daniel McFadden, is important to Bayesian data analysis since it reduces a difficult high-dimensional inference problem to a much simpler problem of sampling independent low-dimensional variables [28]. To recover the binary labels, we have

$$y_i = \begin{cases} 1 & \text{if } f(x_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) > 0, \\ -1 & \text{otherwise.} \end{cases}$$

The directed graphical model in Fig. 3 summarizes the Bayesian kernel machine for classification.

3   Two augmented models for data association

The model presented up to this point is nearly identical to the one proposed in [26]. It assumes all the labels in the training data are known. In this section, we augment the model with either constraints (Sec. 3.1) or group statistics (Sec. 3.2) in order to handle weak supervision.

3.1   Constrained multiple instance learning

When the image caption says that no object is present, all the labels are observed to be negative, and we can recover the latent regression variables z_i following (3), as in [28, 26]. We denote observed labels by y_i^k.

When the image contains an instance of the object, the unknown labels y_i^u must satisfy constraints on the minimum number of features of each class. We define $n_{(+)}$ to be the constraint on the minimum number of positive points in an image, and $n_{(-)}$ to be the minimum number of negatively classified points. The prior on the regression coefficients is

$$p(\{z_i^u\}|\boldsymbol{\gamma},\boldsymbol{\beta},\{x_i\}) \propto \prod_i \mathcal{N}(z_i^u|\boldsymbol{\Psi}_{\boldsymbol{\gamma},i}\boldsymbol{\beta},1)\,\mathbb{I}_{C_{(-)}}(\{z_i^u\})\mathbb{I}_{C_{(+)}}(\{z_i^u\}),$$

where i ranges over the set of exemplars in the image, $C_{(-)}$ is the set of assignments to y_i^u (and accordingly z_i^u) that obey the negative labels constraint $n_{(-)}$, $C_{(+)}$ is the set of assignments to y_i^u that satisfy the constraint $n_{(+)}$, 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 can be realistically handled by MCMC algorithms [11].

3.2   Learning with group statistics

An alternative to constrained data association is to augment the training data with two user-defined statistics: an estimate of the fraction of positive instances for each image j, $m_j \in [0,1]$, and a global parameter χ quantifying the confidence in these guesses. Higher values indicate higher confidence, while $\chi = 0$ is a complete lack of confidence, resulting in unsupervised learning.

The observed value m_j is an estimate of the true fraction of positives, λ_j, which in turn is deterministically computed from the labels in the image according to

$$\lambda_j = \frac{1}{N_j} \sum_{i \in d_j} \mathbb{I}_{(0,+\infty)}(z_i^u), \tag{4}$$

where N_j is the total number of extracted feature vectors in image j. Note that we implicitly integrate out y_i^u in (4). We use the Beta distribution to model this noisy measurement process, so the prior on m_j is

$$p(m_j \,|\, \lambda_j, \chi) = \text{Beta}\left(\chi\lambda_j + 1, \chi(1 - \lambda_j) + 1\right) \propto m_j^{\chi\lambda_j}(1 - m_j)^{\chi(1-\lambda_j)}.$$

The augmented classification model with group statistics is summarized in Fig. 4.
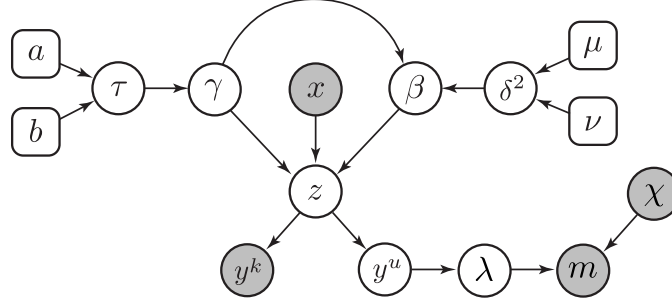
Fig. 4. The directed graphical representation of the classification model with group statistics. Shaded nodes are observed during training, and square nodes are fixed hyperparameters.

4   Model computation

The classification objective is to estimate the density

$$p(y_{N+1}=1|x_{N+1}, \boldsymbol{x}, \boldsymbol{y}^k) = \int p(y_{N+1}=1|x_{N+1}, \theta)p(\theta|\boldsymbol{x}, \boldsymbol{y}^k)d\theta$$

for an unseen point x_{N+1}, given the training data $\{\boldsymbol{x}, \boldsymbol{y}^k\}$, where $\theta = \{\boldsymbol{\gamma}, \boldsymbol{\beta}\}$ is the set of parameters that directly influence prediction. Obtaining this probability requires a solution to an intractable integral, so we approximate it with the Monte Carlo point-mass estimate

$$p(y_{N+1}=1 \,|\, x_{N+1}, \boldsymbol{x}, \boldsymbol{y}^k) \approx \frac{1}{n_s} \sum_{s=1}^{n_s} p(y_{N+1}=1 \,|\, x_{N+1}, \theta^{(s)})$$

$$\approx 1 - \frac{1}{n_s} \sum_{s=1}^{n_s} \Phi\big(-\boldsymbol{\Psi}_{N+1, \boldsymbol{\gamma}^{(s)}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(s)}\big),$$

where n_s is the number of samples, and each sample $\theta^{(s)} = \{\boldsymbol{\gamma}^{(s)}, \boldsymbol{\beta}^{(s)}\}$ is distributed according to the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{y}^k)$. Kück *et al.* [11] develop an MCMC algorithm for sampling from the posterior by augmenting the original blocked Gibbs sampler [26] 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 [11] uses rejection sampling to sample the unknown labels subject to the constraints or group statistics, while we adopt a more efficient MCMC scheme and sample from the full conditionals in each document.

5   Conditional random field for integration of multiple cues

Even though positively classified local features often lie on the object (see the experimental results of Sec. 6.3), they are inadequate for separating the object from the background. Interest regions have been used successfully as a basis for image classification, but there are few positive results extending to the localization of objects. Here, we add an additional layer to localize the objects in an image. The basic intuition behind our approach is that labels on nearby interest regions and neighbouring segments should be useful in predicting a segment label. We propose a simple conditional random field that incorporates segmentation cues and the interest region labels predicted by our Bayesian kernel machine. Spatial integration is achieved in a generic fashion, so we expect our localization scheme applies to a variety of object classes.

The first step is to learn a classifier using the Bayesian learning algorithm described in Sections 2-4. Next, the image is decomposed into *superpixels* — small segments which induce a low compression [12]. We use the Normalized Cuts algorithm [29] to segment images, but other (less expensive) methods could possibly be used with similar returns. The extracted features of small segments are hardly sufficient for locating object classes in cluttered scenes, so the novel step is the construction of a conditional random field [13] (CRF) that propagates information across an image's neighbouring superpixels and interest regions.

Interest region labels influence the segment labels through CRF potentials. The strength of a potential is determined according to the overlap between the interest region and the segment. Defining a_i to be the area occupied by interest region i, and a_{ik} to be the overlap between segment k and interest region i, the potential on the kth segment label y_k^s is defined to be

$$\phi_k(y_k^s) = \sum_i \frac{a_{ik}}{a_i} \delta(y_k^s = y_i),\tag{5}$$

where y_i is the interest region label predicted by the sparse kernel machine classifier (1), i ranges over the set of interest regions in the image, and $\delta(x=y)$ is the delta-Dirac indicator which returns 1 when x is equal to y, and 0 otherwise.

Next, we define the potential between two adjacent segments k and l to be

$$\mu_{kl}(y_k^s, y_l^s) = \theta_\mu + \left(\frac{b_{kl}}{2b_k} + \frac{b_{kl}}{2b_l}\right)\delta(y_k^s = y_l^s),\tag{6}$$

where b_k is the contour length of segment k and b_{kl} is the length of the border shared by segments k and l. The pairwise potential (6) is the prior compatibility of the labels of neighbouring segments.

Putting the potentials (5,6) together, the joint probability of the segment labels $\boldsymbol{y}^s$ is given by

$$p(\boldsymbol{y}^s \mid \boldsymbol{y}) = \frac{1}{Z(\boldsymbol{y})} \prod_k \phi_k(y_k^s) \prod_l \mu_{kl}(y_k^s, y_l^s),\tag{7}$$

where the partition function $Z(\boldsymbol{y}) = \sum_{\boldsymbol{y}^s} \prod_k \phi_k(y_k^s) \prod_l \mu_{kl}(y_k^s, y_l^s)$ ensures that the probabilities sum to unity. There is only a single free parameter, θ_μ, which controls the strength of the potential. At this point, there is no learning; we tune the parameter by hand. In our experiments, we set θ_μ to a relatively strong prior, 0.1, which encourages neighbouring segments to have the same labels.

Even though equation (7) contains a product over all pairs (k, l) of segments in the image, the adjacency graph is sparse since only a few superpixels will share a border, so it is reasonable to run an inference algorithm suitable for sparse graphs. We use the tree sampling algorithm of [30] to infer the hidden labels $\boldsymbol{y}^s$.

6   Experiments

We conduct three sets of experiments. 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 cues in a principled manner, we demonstrate reliable localization of objects. We start by describing the setup used in our experiments.

6.1   Experiment setup

We use interest region detectors which select informative or stable regions of the image. We use three different scale-invariant detectors: the Harris-Laplace detector [31] which finds corner-like features, the Kadir-Brady detector [32] which proposes circular regions with maximum grey-level entropy, and the Laplacian method [33] which detects blob-like structures. Based on earlier studies [34], we chose the Scale Invariant Feature Transform (SIFT) [10] 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.

For fair comparison, we adjust 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. Note Fergus *et al.* [7] extract only 20 features per image on average, owing in part to the expense of training, while Opelt *et al.* [6] learn from several hundred regions per image.

For all our experiments using the constrained data association model (Sec. 3.1), we fix the label constraint n_0 to 0 and set n_1 between 15 and 30, depending on the object 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. When employing the group statistics model (Sec. 3.2), we set the parameters to be approximately $m = 0.3$ and $\chi = 400$. We set $a = 1$ and b according to a feature selection prior of approximately 200 active kernel centres, and we bestow near uninformative priors on the rest of the model parameters. In all our experiments, we set σ 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 second per image on a 2 GHz Pentium machine. The code and data for our experiments are available at http://lear.inrialpes.fr/objrecls.

6.2   Image classification

The experiments in this section quantify our model's capacity for identifying the presence or absence of objects in images. We refer to this task as image classification. 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. 1 for some example images. For purposes of comparison with other methods, we also present results on some existing databases of airplanes, motorbikes, wildcats, bicycles and people. The experiment data is summarized in Table 1.

class	Training images		Test images	
	with object	without	with object	without
airplanes	400	450	400	450
motorbikes	400	450	400	450
wildcats	100	450	100	450
bicycles	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 motorbikes (side) and airplanes (side) categories (http://www.vision.caltech.edu/html-files/archive.html), the Corel Image database for the Wildcats, the Graz bicycles and people data sets (http://www.emt.tugraz.at/∼pinz/data/GRAZ_01), and the INRIA car database (http://lear.inrialpes.fr/data).

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 Table 2. We report performance using the

data set	H-L	K-B	LoG	Combo	Csurka	Fergus	Opelt
airplanes	0.985	0.993	0.938	0.998	0.962	0.902	0.889
motorbikes	0.988	0.998	0.983	1.000	0.980	0.925	0.922
wildcats	0.960	0.980	0.930	0.990	0.920	0.900	—
bicycles	0.920	0.880	0.840	0.900	0.880	—	0.865
people	0.800	0.740	0.840	0.820	0.780	—	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 two three columns refer to the performance reported by Fergus *et al.* [7] and Opelt *et al.* [6]. The third column from the right is a reimplementation of the bag-of-keypoints model of Csurka *et al.* [9] using affine-invariant Harris-Laplace interest regions. 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 [31], Kadir-Brady entropy detector [32], Laplacian of Gaussians [33], and combination of the three detectors.

Receiver Operating Characteristic (ROC) equal error rate, a standard evaluation criterion [6, 7]. 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. We used the constrained data association model for these experiments, since constraints were easier to specify for most of the existing data sets.

Observe that our model in combination with the three detectors always produces the best image classification (at least when comparisons with other methods are available). Moreover, our model does very well in classifying car images in spite of the aforementioned challenges posed by the training examples. We omitted error bars because independent MCMC trials with fixed priors exhibited little variance.

One of the more interesting results of Table 2 is that no single detector dominates over the rest. This highlights the importance of having a wide variety of feature types for object class recognition.

Training with the combination of the Harris-Laplace, Kadir-Brady and LoG detectors often — albeit inconsistently — improves the equal error rate. For instance, we see that the ROC equal error rate decreases in the combination scenario for car, people and bicycle classification. Upon closer inspection, however, the ROC equal error rate can be deceptive. If we examine the full ROC plots in Fig. 5, the combination of detectors now appears to be equally advantageous. Importantly, a precision-recall plot for the task of labeling individual features as belonging to cars in Fig. 6 shows that our classifier picks the best individual features first when given the choice between three detectors in the combination scenario (the ground truth was determined according to manual object-background segmentations of the scenes). Note that in Fig. 6 the Harris-
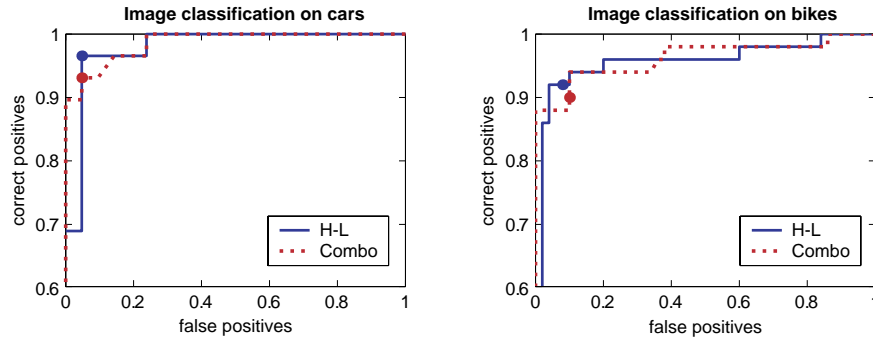
Image classification on cars

Image classification on bikes

Fig. 5. The graph on the left plots the ROC curve for classification performance of car test images using the Harris-Laplace detector (blue solid line) and the combination of three detectors (red dotted line). The graph on the right shows analogous results for the bicycles test set. In both cases, the equal error rate (indicated by a large dot) is inferior in the combination, but according to the full ROC curve it may perform slightly better.

Point classification on training cars

Point classification on test cars
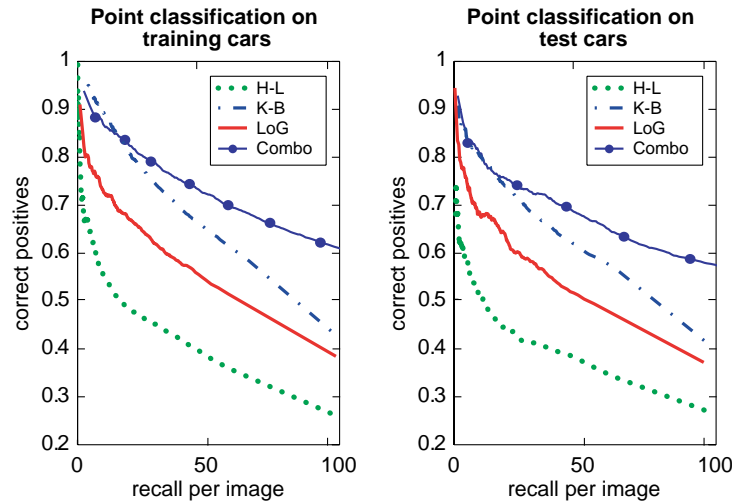
Fig. 6. Plots of precision (percentage of correct positives) versus average recall per image for the task of labeling individual features as belonging to cars. Our definition of recall here is not standard since we do not divide by the number of regions in the image. The combination scenario extends to 300 along the x-axis, but we cut it off at 100. 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).

Laplace detector is overly penalized because it often selects corner-like features that are near, but not on, cars. Fig. 7 shows a couple examples where learning a model with a combination of detectors results in an improved image classification.
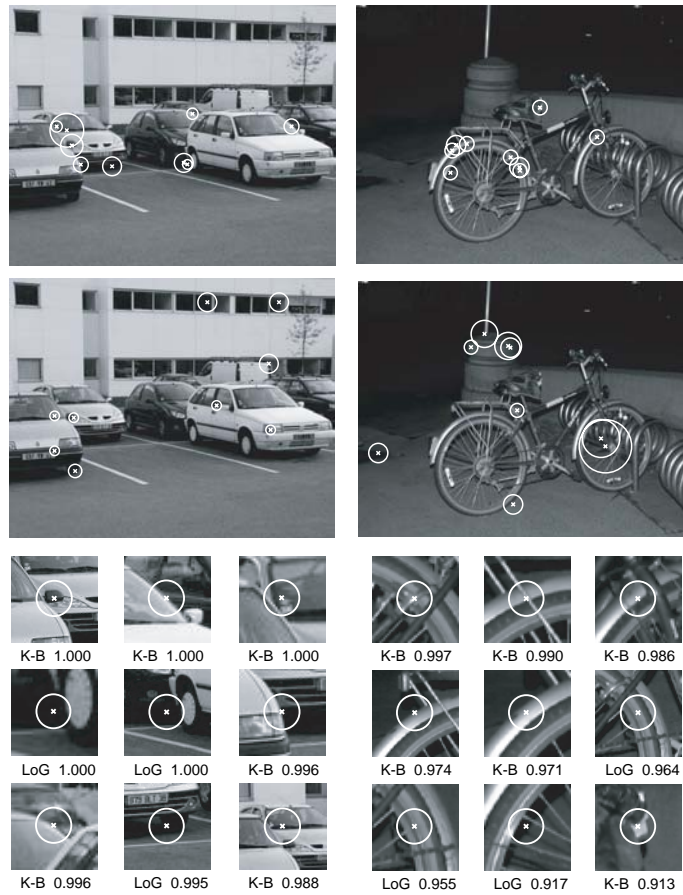

Fig. 7. Two examples in which the combination of detectors (top row) results in improved image classification over the Harris-Laplace detector (middle row). The circles represent the 9 interest regions that are most likely to belong to cars or bicycles. The bottom row shows the top features along with 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.

We show examples of correctly and incorrectly classified images, along with the interest regions extracted by the detectors, in Fig. 8. Incorrectly classified
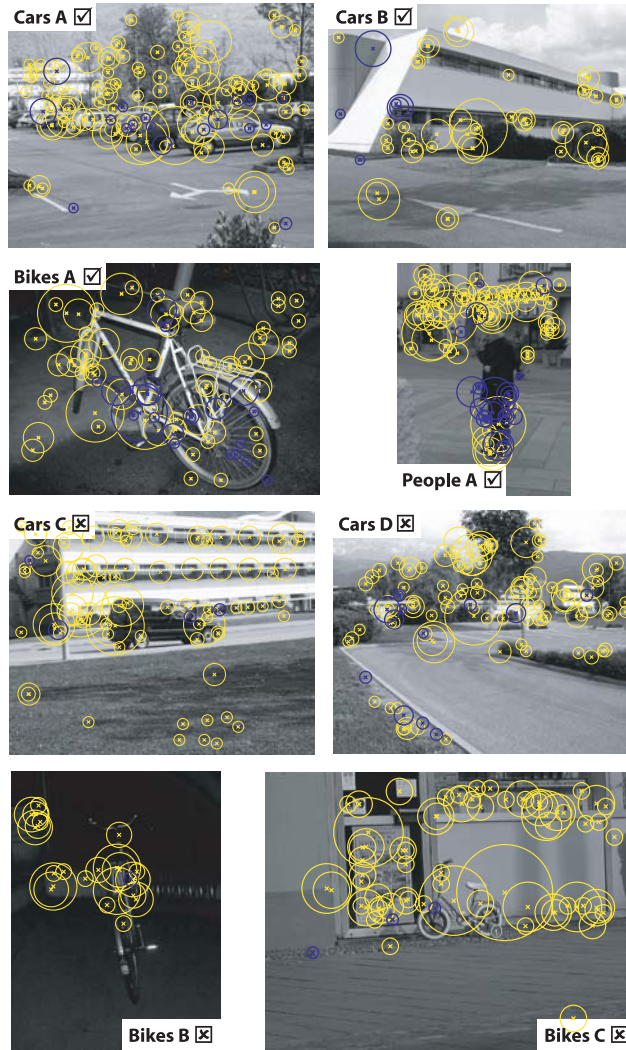
Fig. 8. Test images correctly (top four images) and incorrectly (bottom four) classified using interest regions extracted by the Harris-Laplace (for cars and bicycles) or LoG detector (for people). Dark blue circles represent local interest regions that are more likely to belong to the object, while yellow circles more probably belong to the background.

images tended to be unlike any of the images observed during training, such as the van and the child's bicycle. Problematic images also tended to exhibit unusual illumination conditions.
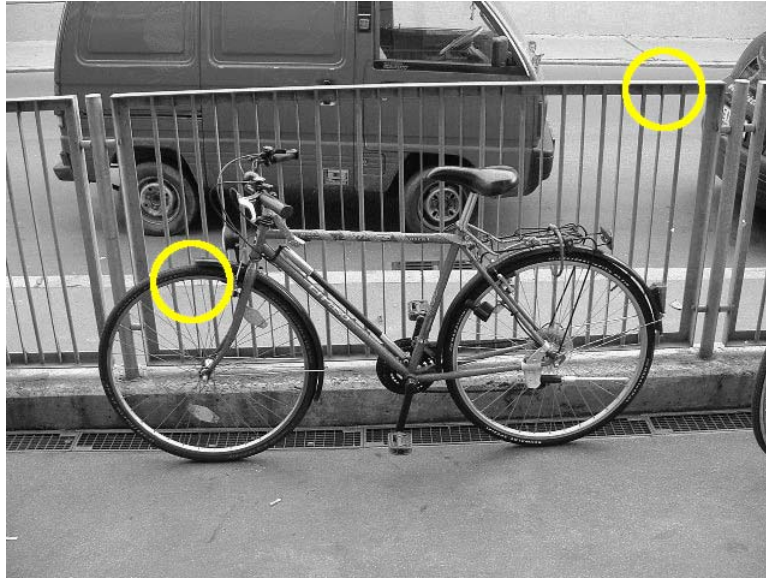
Fig. 9. The yellow circles are two interest regions extracted by the entropy detector. By looking only at the pixels inside the yellow circle, it is difficult to tell which one belongs to the bicycle and which one belongs to the background.

6.3   Investigation of data association

In this section, we ask to what extent our proposed scheme for data association correctly labels the individual features, given that it is provided very little information. In some sense, this task is unfair since many individual interest regions cannot discriminate the object class. Fig. 9 shows two Kadir-Brady interest regions that do not help discriminate bicycles. Even under the best of conditions, we should not expect the classifier to predict the feature labels perfectly.

We frame the data association question as follows: if manual segmentations were provided, how much would we gain over image caption data? The answer of course depends on the nature and quality of the data. At the very least, we should expect that our model predicts the correct labels of the discriminative features in the INRIA car database, since it appears to exhibit sufficient information to delineate positive and negative instances.

We conduct the experiment on the car database using the interest regions extracted from the Harris-Laplace detector. We use both hard constraints and group statistics. We increase supervision by setting some unknown labels y_i^u known to fall on cars to $y_i^k = 1$. Note that there is some noise in this process, since an interest region near a car may or may nor be associated with it. The results are presented in Fig. 10. The ROC curves show how the accuracy in labeling individual features changes with different levels of supervision. As expected, the addition of a few hand-labeled points improves recognition in training images.
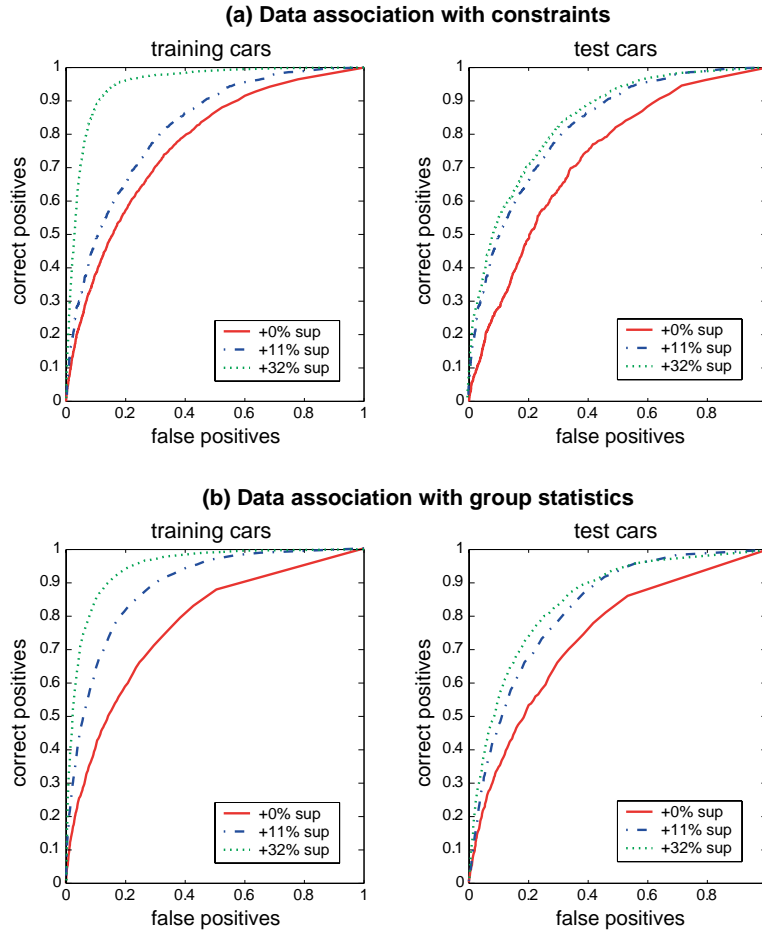
(a) Data association with constraints

(b) Data association with group statistics

Fig. 10. The ROC plots demonstrate how learning with different proportions of hand-labeled points affects performance on labeling individual car features. *(a)* Labeling accuracy using the constrained data association model (Sec. 3.1). *(b)* Labeling accuracy using the data association with group statistics model (Sec. 3.2). The Harris-Laplace detector is used for both these experiments. With a lot of supervision, the models predict near-perfect feature labels in the training images, but there is little improvement in the test images.

However, further upgrades in supervision result in almost no gains to recognition in test images. This shows that our data association schemes largely compensate for the lack of annotations in the data. Fig. 11 demonstrates this effect on a single image.
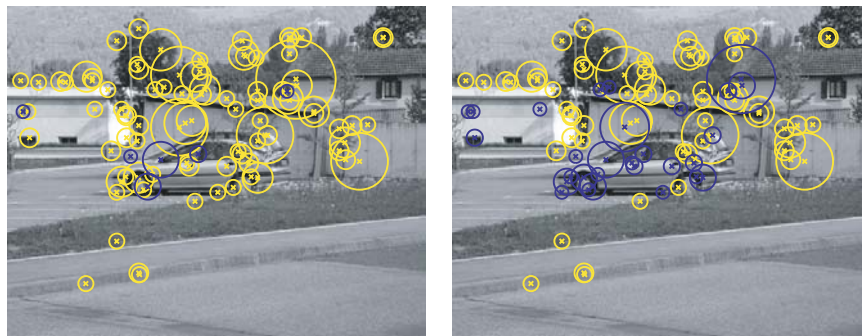
Fig. 11. Labeling of individual interest regions using the model augmented with data association constraints. The model was trained with various levels of supervision (see Fig. 10). *Left:* Car test image, no observed car labels during training. *Right:* The same image, except that the model was trained with an additional 11% observed feature labels. Dark blue circles are more likely than not to belong to the object, and light yellow circles are more likely to belong to the background.

6.4   Object localization

In this section, we evaluate the proposed object localization model. In order to quantify its effectiveness, we compare the object-background segmentation predicted by the model with those drawn by hand. Some examples of manual segmentations are shown in Fig. 12. Perfect localization requires: 1.) that
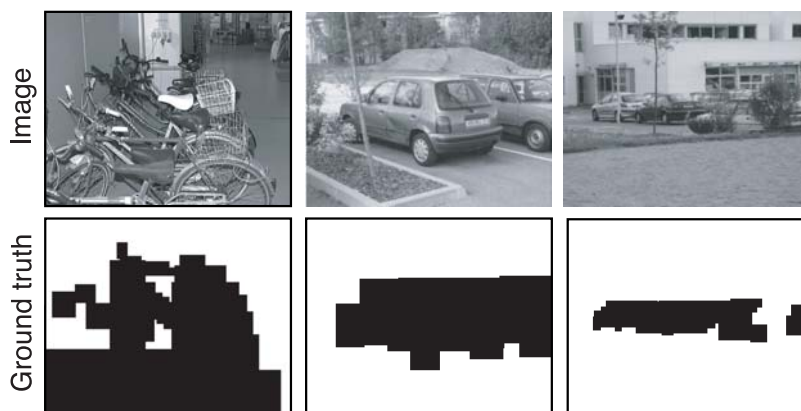

Fig. 12. Examples of ground truth segmentations from the bicycle and car databases.

the boundaries of the segments follow the object boundaries, and 2.) that the

conditional random field predicts the segment labels correctly. Even then, the evaluation may not be precise since the ground truth annotations contain some error, as evidenced by the examples in Fig. 12.
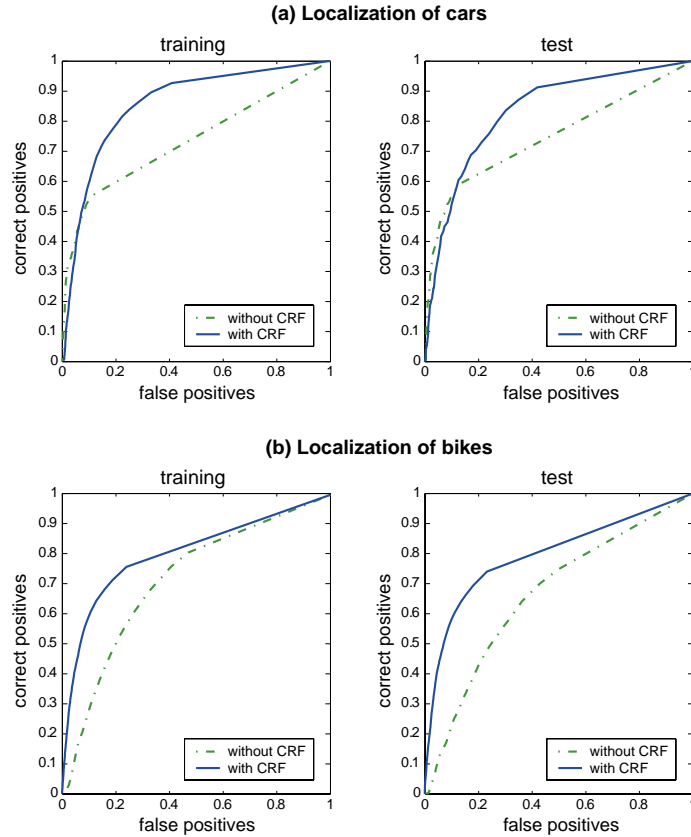

Fig. 13. ROC plots for localization of *(a)* bicycles and *(b)* cars, with (solid blue line) and without (dashed green line) the proposed CRF model. We use the Harris-Laplace detector for the cars, and the Kadir-Brady entropy detector for extracting interest regions in the bicycles database. Notice that the addition of the superpixels with the conditional random field dramatically improve the quality of the object-background separation.

The ROC curves in Fig. 13 report the quality of the estimated segmentations in the car and bicycle databases. The ROC plots are obtained by thresholding the label probabilities on the segments and then finding the intersection with the ground truth segmentations. We use the Harris-Laplace detector for the car images and the Kadir-Brady entropy detector for the bicycles. The "without
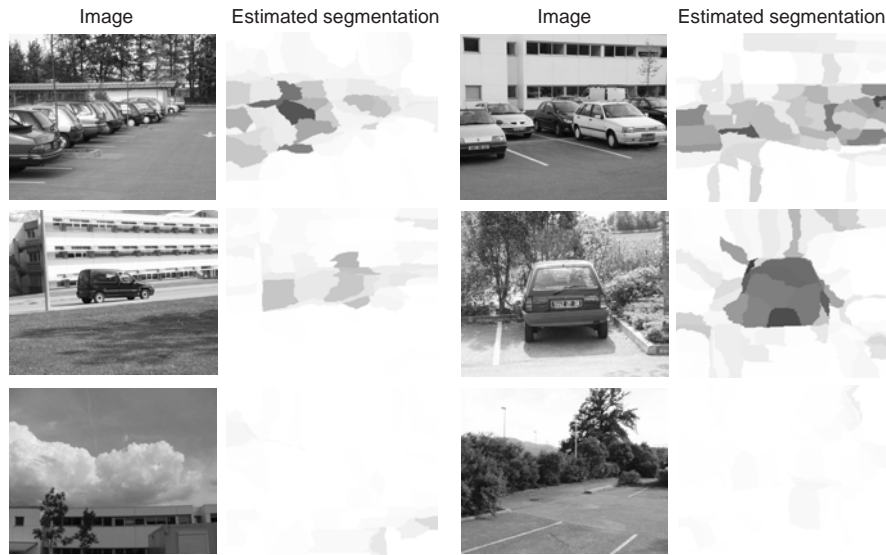
Fig. 14. Good localization results on car test images. Darker patches are more likely to correspond to cars.
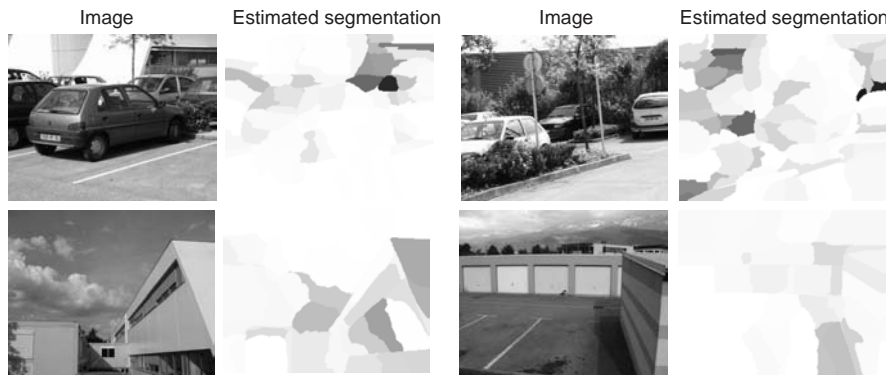

Fig. 15. Poor localization results on car test images. Darker patches are more likely to belong to the car class.

CRF" results in Fig. 13 do not use the superpixels; the spatial information is acquired from the location and scale of the interest regions. Our results show that we gain a lot in localization by using the segments to propagate interest region labels. The results in Fig. 13 show that our method is more reliable for locating cars in images. Without the CRF, Fig. 13a shows that the first selected labels selected are almost always within the boundary of cars, but the model

cannot make any predictions in areas where no interest regions are extracted by the detector.

Some successful predictions in car test images are shown in Fig. 14, and some less successful car recognition results are displayed in Fig. 15. Localization failed when the interest regions and superpixels failed to complement each other. Notice we did not tailor the CRF to an object class, so recognition performance might very well generalize to other visual object classes.

7   Conclusions and Discussion

In this paper, we extended the discriminative power of local scale-invariant features using Bayesian learning. We showed that both models for generalized multiple instance learning — constrained data association and learning with group statistics — are 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. The experiments show our method successfully segments the object from the background. The important implication is that our Bayesian model selects the features that really lie on or near the object.

The conditional random field we proposed does not adapt its parameters to the object class in question since there is no learning involved. An important question is whether our Bayesian methods for data association can be extended to more advanced models for learning to recognize objects, such as those that incorporate context, shape information, correlations between features and different types of features. We suspect that it is as much a challenge for machine learning as it is for vision.

Acknowledgments

References

1. Paul Viola and Michael J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.
2. Bastian Leibe, Edgar Seemann, and Bernt Schiele. Pedestrian detection in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
3. Krystian Mikolajczyk, Cordelia Schmid, and Andrew Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proceedings of the European Conference on Computer Vision*, 2004.

4. S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *Proceedings of the European Conference on Computer Vision*, 2002.

5. G. Dorkó and C. Schmid. Selection of scale invariant neighborhoods for object class recognition. In *Proceedings of the International Conference on Computer Vision*, 2003.

6. A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proceedings of the European Conference on Computer Vision*, 2004.

7. R.Fergus, P.Perona, and A.Zisserman. Object class recognition by unsupervised scale-invariant learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003.

8. Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *Proceedings of the International Conference on Computer Vision*, 2001.

9. Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Proceedings of the ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

10. David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

11. H. Kück, P. Carbonetto, and N. de Freitas. A Constrained semi-supervised learning approach to data association. In *Proceedings of the European Conference on Computer Vision*, 2004.

12. X. Ren and J. Malik. Learning a classification model for segmentation. In *Proceedings of the International Conference on Computer Vision*, 2003.

13. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields. In *Proceedings of the International Conference on Machine Learning*, 2001.

14. Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Multiple instance learning with generalized support vector machines. In *Proceedings of the AAAI National Conference on Artificial Intelligence*, 2002.

15. T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solving the multiple instance learning with axis-parallel rectangles. *Artificial Intelligence*, 89(1):31–71, 1997.

16. X. Zhu, Zoubin Ghahramani, and John Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning*, 2003.

17. Hendrik Kück and Nando de Freitas. Learning about individuals from group statistics. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2005.

18. A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin. Context-based vision system for place and object recognition. In *Proceedings of the International Conference on Computer Vision*, 2003.

19. 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

20. 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 *Proceedings of the European Conference on Computer Vision*, 2002.

21. P. Carbonetto, N. de Freitas, P. Gustafson, and N. Thompson. Bayesian feature weighting for unsupervised learning, with application to object recognition. In *Proceedings of the Workshop on Artificial Intelligence and Statistics*, 2003.

22. P. Carbonetto, N. de Freitas, and K. Barnard. A Statistical model for general contextual object recognition. In *Proceedings of the European Conference on Computer Vision*, 2004.
23. G. Celeux, M. Hurn, and C. P. Robert. Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95:957–970, 2000.
24. M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.
25. Thomas Serre, Lior Wolf, and Tomaso Poggio. Object recognition with features inspired by visual cortex. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005.
26. S. S. Tham, A. Doucet, and R. Kotagiri. Sparse Bayesian learning for regression and classification using Markov Chain Monte Carlo. In *Proceedings of the International Conference on Machine Learning*, 2002.
27. J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley and Sons, 2000.
28. D. McFadden. A Method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57:995–1026, 1989.
29. Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1997.
30. Firas Hamze and Nando de Freitas. From fields to trees. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 2004.
31. K. Mikolajczyk and C. Schmid. Selection of scale-invariant parts for object class recognition. In *Proceedings of the International Conference on Computer Vision*, 2001.
32. T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.
33. T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2), 1998.
34. K. Mikolajczyk and C. Schmid. A Performance evaluation of local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003.