International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Learning to recognize objects with little supervision

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

[1] University of British Columbia, Vancouver, Canada
[2] INRIA Rhône-Alpes, Grenoble, France

The date of receipt and acceptance will be inserted by the editor

Abstract *This paper shows (i) improvements over state-of-the-art local feature recognition systems, (ii) how to formulate principled models for automatic local feature selection in object class recognition when there is little supervised data, and (iii) how to formulate sensible spatial image context models using a conditional random field for integrating local features and segmentation cues (superpixels). By adopting sparse kernel methods, Bayesian learning techniques and data association with constraints, the proposed model identifies the most relevant sets of local features for recognizing object classes, achieves performance comparable to the fully supervised setting, and and obtains excellent results for image classification.*

1 Introduction

Over the past few years, researchers in high-level vision have shifted their focus from matching specific objects to a significantly more challenging problem: recognizing visual categories of objects. Effective solutions do exist for some image classification problems, so the big push is to address more difficult problems such as object localization (*i.e.* segmenting an object from the background). There has been much success in learning robust representations of specific classes in constrained situations, notably frontal faces [52] and pedestrians in street scenes [27,36], but general-purpose models that can be trained to recognize object categories remain elusive.

A wealth of complementary developments in vision and machine learning have lead to improvements in general representations of object classes [1,15,18,38]. 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* local, informative regions of the scene [15,34], often called *keypoints* [12, 31]. Such local *interest regions* bring tolerance to clutter, occlusion and deformable objects, and their sparsity reduces the complexity of subsequent learning and inference. In general, a good detector is one that extracts 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 different criteria. Therefore, we expect that using multiple detectors will provide complementary information, 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 [22] with an efficient Markov Chain Monte Carlo (MCMC) algorithm for approximate probabilistic inference. The inference algorithm identifies a sparse and effective object class representation from the interest region descriptors, and does so with little supervision by explicitly representing the correspondence between the extracted image keypoints and the set of objects. We refer to this as "data association"; see Sections 2-4 for more details.
3. For object localization, integrate two types of visual cues: interest regions and low-level segmentation using *superpixels* [41]. On their own, independent, local interest regions do not contain enough information to segment the object from the background. We propose a simple conditional random field [25] that overcomes this deficiency by propagating information across neighbouring superpixels and weighting superpixel labels by scores obtained from overlapping interest regions. These ideas are described in full 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 user supervision.

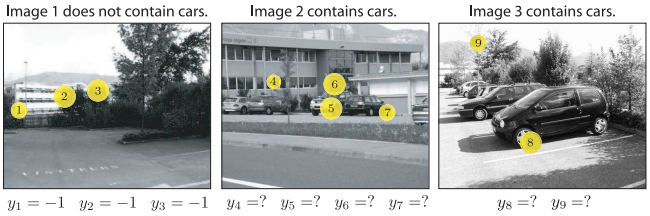

Fig. 1 Three annotated images from the INRIA car training set. The circles represent some of the extracted features. The feature labels y_1, y_2 and y_3 in the first image are known. In the second and third images, we don't know the correspondence between the features and the labels, hence the question 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").

We start with an example that illustrates the need for explicitly modeling data association in object recognition. After that, we motivate the 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 that indicates the presence or absence of one or more cars in the scene. The numbered circles represent a few of the extracted interest regions at their characteristic scale. The first image does not contain a car, so we can justifiably deduce that none of the circles are features of the car object class. In the second and third training images, however, we cannot conclude with certainty which of the regions belong to a car. The conventional approach to this problem is to treat unlabeled feature vectors in the background as noise [1,15,18], an approach which degrades significantly when the object in question occupies only a small part of the unlabeled image, as in the second image.[1] A more sensible strategy is to explicitly model the individual labels, allowing the learning algorithm to exploit the unlabeled background features instead of being hindered by them. This is precisely the approach proposed in this paper.

Each label is a binary variable that indicates whether the image keypoint belongs to a car (positive) or to the background (negative). Data association is the problem of determining the correspondence between the observations (image keypoints) and the set of objects. This problem has been well-studied in the context of citation matching [39]. In the setting we explore here, in which

[1] FIX. The cited method [18] has a latent variable that indexes the parts of an object, and an index of 0 corresponds an occluded part. Curiously, they do not use this latent variable to solve the correspondence problem; hence their "unsupervised" learning approach.

Fig. 2 Two sample images from the MIT-CSAIL database [51]. Yellow lines indicate car annotations. The annotations are incomplete in both images, so learning with data association is still appropriate in the presence of annotated data.

there are only two classes (positive and negative), data association is closely related to the multiple instance learning problem [2,14]. In the classical multiple instance formulation, a positive group label (here, groups are images) indicates that *at least one* of the individuals in the group has a positive label—corresponding to a "contains cars" caption—while a negative group label implies that *all* individuals in the group have a negative label.[2] 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. We propose two alternatives. In the first, we introduce image-level constraints that enforce a certain number of the image keypoints 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. There is 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 significantly 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 [23].

One might be skeptical that it is possible to achieve proper recognition in this setting, given the wide variability exhibited in the training images, the high dimensionality of the features, and the fact that there are hundreds of unlabeled points per image. One alternative—discovery of object categories without *any* labels—is extremely sensitive to the composition of the data set, and works best when the images contain isolated, unoccluded instances of the object [47]. Such *unsupervised*

[2] Data association is also commonly studied as a special case of *semi-supervised learning* [57]. This formulation is less compatible since it has no notion of groups.

methods are especially prone to learning artifacts of the data set. The other alternative—complete supervision—is not only unappealing but also unrealistic for general object recognition problems. Complete supervision requires the user to segment and annotate objects from the background. This is not only a time-consuming task, but also poorly defined since people tend to 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 [37]). 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 [51], painstakingly 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, our model can better exploit such training data.

Attempts have already been made in tackling the problem of data association in object recognition. Duygulu *et al.* [16] 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 [8] and spatial relations [7]. The problem with these approaches 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 [10]. More complex representations only exacerbate the issue, so mixture models are limited to simple, unimodal object classes. While [7,8,16] tackle multi-category classification, we can do likewise by combining responses from multiple binary classifiers [50].

1.2 A Case for Bayesian learning in object recognition

We employ the augmented Bayesian classification model developed in [22] 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, which results in a single point estimate.

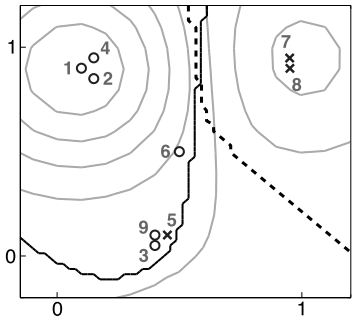

Fig. 3 The 9 feature vectors extracted for the small car data set (Fig. 1). Points marked with an × are labeled as "car", and circles represent negative instances. The lightly shaded lines are contours of the kernel response function (2) with $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and σ set following to the description in the text. The dark, solid line is a decision boundary obtained from simulation of the model with full supervision (Sec. 2). The dashed line is a decision boundary obtained from simulation of the model with incomplete supervision (Sec. 3.1).

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 [12] that chooses the features that best identify an object (e.g. the model for cars 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 [18] and global context [7,51]. Despite these objections, independent parts models are not only efficient and relatively simple to implement, but also remain the state-of-the-art in detection systems [12,45] and, as we show, can function as a basis for localization.

2 Bayesian kernel machine for classification

We start with a description of the model that assumes complete supervision. In other words, each feature vector x_i has a known label $y_i^k \in \{-1, 1\}$. The next section considers the case when some of the labels are not obesrved. We use a small running example throughout this section to illustrate key concepts.

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 feature vectors $\{x_i \,|\, i \in d_j\}$. The set of feature vectors for all the images used during training is $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$, where N is the total size of the training set. Our running example is the small car data set depicted in Fig. 1, in which a hypothetical method has extracted $N = 9$ feature vectors from the 3 images:

$$\boldsymbol{x} = \left\{ \begin{matrix} 0.10 & 0.15 & 0.40 & 0.15 & 0.45 & 0.50 & 0.95 & 0.95 & 0.40 \\ 0.90 & 0.85 & 0.05 & 0.95 & 0.10 & 0.50 & 0.95 & 0.90 & 0.10 \end{matrix} \right\}.$$

Sec. 6.1 describes how to automatically obtain the feature vectors beginning with the raw pixel data. Under the assumption of complete supervision, the question marks in Fig. 1 have been replaced with the correspondence described in the image caption. The feature vectors x_i are plotted in Fig. 3. Points marked with an × are positive instances, and circles represent negative instances.

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 discrete labels using the probit link function. Following Tham *et al.* [49], we have

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

where the unknown function f is a weighted sum of kernel responses given by

$$f(x_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{k=1}^{N} \gamma_k \beta_k \psi(x_i, x_k). \tag{2}$$

The probit link $\Phi(\cdot)$ is the cumulative density function of the standard Normal distribution; it provides a continuous and monotonic mapping from the reals to the range $[0, 1]$, thus producing a valid probability. By convention, researchers tend to adopt a logistic (sigmoidal) link function, but the probit link is equally valid and will lead to an efficient sampler.

The kernel or similarity 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 $\boldsymbol{\beta} = [\beta_1\ \beta_2\ \cdots\ \beta_N]^T$. In this semi-parameteric modeling approach, feature vectors are mapped to a nonlinear, high-dimensional kernel manifold. There are as many parameters β_k as there are data instances, but as we will see the variables γ_k reduce the model (1) to a much lower manifold. Our model is *discriminative*, because it is specified as a conditional probability distribution of labels $\boldsymbol{y}$ given observations $\boldsymbol{x}$, and not as a joint distribution on $\{\boldsymbol{x}, \boldsymbol{y}\}$, as in a *generative* model. This means we do not expend extra computational effort in modeling quantities other than the variables of interest, the labels $\boldsymbol{y}$.

We introduce sparsity through a set of parameters $\boldsymbol{\gamma} = [\gamma_1 \cdots \gamma_N]$, 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 feature vectors. By learning $\boldsymbol{\gamma}$, we learn the relevant set of feature vectors, or prototypes, for each class.

Let's illustrate these ideas by returning to our small example. Suppose we were to choose the interest regions 1, 3 and 8 as prototypes, so $\gamma_1 = \gamma_3 = \gamma_8 = 1$. A reasonable choice for the regression coefficients might then be $\beta_1 = -1$, $\beta_3 = 0.01$ and $\beta_8 = 0.5$. To see why this is a good choice, observe that the contours of the kernel response function (2)—depicted by the lightly shaded lines in Fig. 3—closely follow the decision boundary obtained by simulation, depicted by the dark, solid line in Fig. 3. (See Sec. 4 for more details on the simulation.) Here we set $\sigma = 1/4$. Notice that x_7 and x_8 are isolated from negative instances, and therefore do a good job discriminating cars. The fifth feature vector, on the other hand, responds in a similar way to background features, hence lies on the decision boundary.

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

$$p(y_i = 1 \,|\, x_i, \boldsymbol{\beta}, \boldsymbol{\gamma}) = \Phi\left(\boldsymbol{\Psi}_{i,\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}\right), \tag{3}$$

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 inner product in (3) is shorthand for

$$\boldsymbol{\Psi}_{i,\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}} = \psi(x_i, x_1)\beta_1 + \cdots + \psi(x_i, x_N)\beta_N.$$

The reduced kernel matrix $\boldsymbol{\Psi}_{\boldsymbol{\gamma}}$ has height N and width equal to the number of active kernels.

We follow a hierarchical Bayesian strategy [4], 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.

A basic result in statistics states that when the variance of the noise is known up to a multiplicative factor (denoted by δ^2) and $\boldsymbol{\gamma}$ is fixed, the maximum likelihood estimator of the regression coefficients $\boldsymbol{\beta}$ follows a Normal distribution with covariance $\delta^2(\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T \boldsymbol{\Psi}_{\boldsymbol{\gamma}})^{-1}$ [42]. This motivates a conjugate prior of the form

$$p(\boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \delta^2) = \mathcal{N}(0, \delta^2 \mathbf{S}_{\boldsymbol{\gamma}}), \tag{4}$$

where $\mathbf{S}_{\boldsymbol{\gamma}} = (\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T \boldsymbol{\Psi}_{\boldsymbol{\gamma}} + \epsilon I)^{-1}$, I is the identity matrix, and ϵ is a small value that helps maintain a prior covariance with full rank.[3] This is a stable version of the maximum

[3] For the Gaussian prior to be well-defined, the covariance matrix must be positive definite. It is easy to come up with

entropy *g-prior* originally recommended by Zellner [55]. While some consider this choice of conjugate prior unorthodox due to its dependence on the data, Zellner's g-prior is still widely adopted in econometrics [21]. Also, from a practical point of view, we avoid having to estimate a full covariance matrix.

The multiplicative factor δ^2 is in turn assigned an inverse Gamma prior with two hyperparameters $\frac{\mu}{2}$, $\frac{\nu}{2}$ 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 [4,42]. Typically, we set μ and ν to near-uninformative values.

Following [22], 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 Gibbs sampler to a reasonable number of active kernels.

The model is highly intractable. In particular, it is non-linear and the posterior of the coefficients $\boldsymbol{\beta}$ is a correlated, hard to sample, high-dimensional distribution. However, we can simplify the problem enormously by introducing easy to sample low-dimensional variables $\boldsymbol{z} = \{z_1, z_2, \ldots, z_N\}$. These can be seen as a continuous version of the binary labels satisfying

$$y_i = \begin{cases} +1 & \text{if } z_i > 0, \\ -1 & \text{otherwise.} \end{cases} \tag{5}$$

From the probit model (1), the z_i's are independently distributed according to

$$p(z_i \,|\, \boldsymbol{\gamma}, \boldsymbol{\beta}, x_i) = \mathcal{N}(\boldsymbol{\Psi}_{i,\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1). \tag{6}$$

The choice of distributions (5) and (6) ensures that the original model (1) is recovered if we marginalize out $\boldsymbol{z}$.[4] 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 inference problem to a much simpler problem of sampling independent low-dimensional variables [33].

Putting all the pieces together, the discriminative model is given by a joint density over the possible configurations of labels $\boldsymbol{y}$ and unknowns $\boldsymbol{\theta} = \{\boldsymbol{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \delta^2, \tau\}$

small examples in which the Gram matrix $\boldsymbol{\Psi}^T \boldsymbol{\Psi}$ has one or more eigenvalues that are close to zero, leading to numerical instability.

[4] The unit variance of z_i given x_i, $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ follows directly from the probit model (1), and in no way limits the flexibility of the classifier.

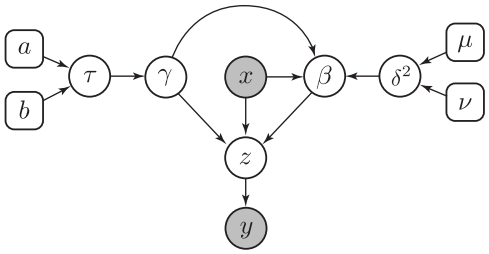

Fig. 4 The directed graphical model of the fully-supervised classification model. Shaded nodes are observed during training, and square nodes are fixed hyperparameters.

conditioned on the observed quantities $\boldsymbol{x}$ and the hyperparameters:

$$p(\boldsymbol{y}, \boldsymbol{\theta} \,|\, \boldsymbol{x}) = p(\tau \,|\, a, b)\, p(\delta^2 \,|\, \mu, \nu)\, p(\boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \delta^2) \times \prod_k p(\gamma_k \,|\, \tau) \prod_i p(y_i \,|\, z_i)\, p(z_i \,|\, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{\gamma}). \tag{7}$$

The joint density (7) is *complete* due to the appearance of the latent variables $\boldsymbol{z}$ [30]. The directed graphical model in Fig. 4 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 [49]. 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. We assume the practitioner has the privilege of setting the constraints or group statistics priors in an informed manner. For instance, the practitioner might know in advance whether the object occupies a little or a lot of space in the images. One might contend that training is no longer fully automatic in this setting. Still, we maintain that this is better than either no supervision or full supervision and, for that matter, experiments (Sec. 6) show that our approach bears considerable improvement other existing methods even when little thought is put into the choice of constraints.

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 recover the latent regression variables z_i following (6), as in [33,49]. This situation occurs in the first image of the cars data set (Fig. 1). The observed labels are denoted by y_i^k, and their corresponding real-valued responses are denoted by z_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 hidden variables $\{z_i^u\}$ in each image is

$$p(\{z_i^u\} \,|\, \{x_i\}, \boldsymbol{\gamma}, \boldsymbol{\beta}) \propto \prod_i \mathcal{N}(z_i^u \,|\, \boldsymbol{\Psi}_{\boldsymbol{\gamma},i}\boldsymbol{\beta}, 1) \times \mathbb{I}_{C_{(-)}}(\{z_i^u\})\, \mathbb{I}_{C_{(+)}}(\{z_i^u\}), \quad (8)$$

where i ranges over the set of extracted feature vectors 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: it evaluates to 1 if $\omega \in \Omega$, and 0 otherwise. The prior (8) specifies a region of support only for those label assignments that satisfy the constraints. We assume that the region of support is non-empty. Discrete constraints in non-convex continuous optimization problems can be highly problematic. However, they can be realistically handled by MCMC algorithms [22].

To illustrate the effect of learning under weak supervision, we test a conservative constraint of $n_{(+)} = 1$, $n_{(-)} = 0$ on our token car data set. Small simulations of the model with 100 samples (see Sec. 4) regularly get the correct labels for most of the image keypoints. Note that the shape of the posterior is sensitive to the choice of hyperparameters because the data set is small. Regardless, it is evident that the image labels do not contain enough information to identify the 5th and 6th interest regions correctly or reliably. For the purposes of illustration we show a sample decision boundary obtained under incomplete supervision, depicted by the dashed line in Fig. 3.

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 $m_j \in [0, 1]$ of the fraction of positive instances for each image j, and a global parameter χ quantifying the confidence in these guesses. Higher values indicate greater confidence, and thus greater sensitivity to the choice of m_j, while $\chi = 0$ is a complete lack of confidence, resulting in unsupervised learning. It is up to the practitioner to regulate the level of sensitivity. In our small example, a moderate setting of $m_j = \frac{1}{2}$ and $\chi = 1$ produces classification results and a decision boundary similar to that obtained in the constrained setting (see the dashed line in Fig. 3).

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

$$\lambda_j = \frac{1}{N_j} \sum_{i \in d_j} \mathbb{I}_{(0,+\infty)}(z_i^u), \quad (9)$$

where N_j is the total number of extracted feature vectors in image j. Note that we implicitly integrate out y_i^u in (9). As in [23], we use the Beta distribution to model this noisy measurement process, so the prior on m_j is

$$p(m_j \,|\, \lambda_j, \chi) = \text{Beta}\,(\chi\lambda_j + 1, \chi(1-\lambda_j) + 1) \propto m_j^{\chi\lambda_j}(1-m_j)^{\chi(1-\lambda_j)}. \quad (10)$$

The augmented classification model with group statistics is summarized in Fig. 5.

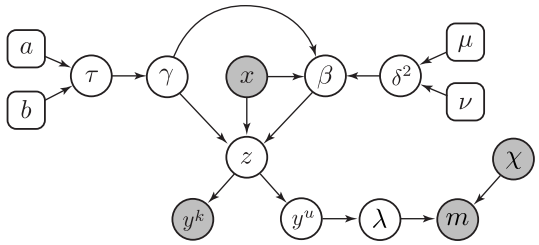

Fig. 5 The directed graphical representation of the classification model with group statistics. Shaded nodes are observed during training, and square nodes are fixed hyperparameters.

4 Model computation

The classification objective is to estimate the predictive density

$$p(y_{N+1} = 1 \,|\, x_{N+1}, \boldsymbol{x}, \boldsymbol{y}^k) = \int p(y_{N+1} = 1 \,|\, x_{N+1}, \boldsymbol{\gamma}, \boldsymbol{\beta})\, p(\boldsymbol{\gamma}, \boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{y}^k)\, d\boldsymbol{\theta} \quad (11)$$

for an unseen feature vector x_{N+1} given the training data $\{\boldsymbol{x}, \boldsymbol{y}^k\}$, where $p(\boldsymbol{\gamma}, \boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{y}^k)$ is the posterior density. This unseen feature vector might respresent the descriptor for an interest region extracted in a test image. Obtaining the probability (11) requires a solution to several intractable integrals, including the normalizing constant of the posterior which arises from Bayes' rule, $\int p(\boldsymbol{y}^k, \boldsymbol{\gamma}, \boldsymbol{\beta} \,|\, \boldsymbol{x})\, d\boldsymbol{\gamma}\, d\boldsymbol{\beta}$, usually called the *marginal likelihood*. Probabilistic inference is further exacerbated by the presence of uncertainty in the labels $\boldsymbol{y}^u$. The approach we take here is to approximate the posterior with a Monte Carlo point-mass estimate. From the definition of the model (1), we have

$$p(y_{N+1} = 1 \,|\, x_{N+1}, \boldsymbol{x}, \boldsymbol{y}^k) \approx \frac{1}{n_s} \sum_{s=1}^{n_s} p(y_{N+1} = 1 \,|\, x_{N+1}, \boldsymbol{\gamma}^{(s)}, \boldsymbol{\beta}^{(s)}) \approx \frac{1}{n_s} \sum_{s=1}^{n_s} \Phi\big(\boldsymbol{\Psi}_{N+1,\boldsymbol{\gamma}^{(s)}} \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(s)}\big), \quad (12)$$

where n_s is the number of samples $\{\boldsymbol{\gamma}^{(s)}, \boldsymbol{\beta}^{(s)}\}$. Ideally, we would like draw independent samples from the posterior $p(\boldsymbol{\gamma}, \boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{y}^k)$, but doing so using rejection sampling or importance sampling would not be advisable in this setting. MCMC is a particular strategy for generating samples as states of a Markov chain. The states do not constitute *i.i.d.* samples from the posterior, but an important central limit theorem tells us that over time the

1 Choose initial states $\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{z}$ and δ^2.
2 For $s = 1, \ldots, n_s$,
3 For all known labels i,
4 Sample $z_i^k \sim p(\cdot \,|\, x_i, y_i^k, \boldsymbol{\gamma}, \boldsymbol{\beta})$; see (15).
5 For all unknown labels i,
6 Sample $z_i^u \sim p(\cdot \,|\, x_i, \{z_{-i}^u\}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \ldots)$; see (15,16,17).
7 Sample $\hat{\alpha}^2 \sim \pi(\cdot \,|\, \mu_\alpha, \nu_\alpha)$.
8 Set $\boldsymbol{w} \leftarrow \hat{\alpha} \times \boldsymbol{z}$.
9 Sample $\alpha^2 \sim \pi(\cdot \,|\, \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\gamma}, \delta^2, \mu_\alpha, \nu_\alpha)$; see (22).
10 Set $\boldsymbol{z}' \leftarrow \sqrt{\hat{\alpha}^2/\alpha^2} \times \boldsymbol{z}$.
11 Sample $\boldsymbol{\beta} \sim p(\cdot \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \delta^2, \boldsymbol{z}')$; see (14).
12 Sample $\delta^2 \sim p(\cdot \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mu, \nu)$; see (18).
13 Sample $\boldsymbol{\gamma}$; see Fig. 7.
14 Set $\boldsymbol{\gamma}^{(s)} \leftarrow \boldsymbol{\gamma}$ and $\boldsymbol{\beta}^{(s)} \leftarrow \boldsymbol{\beta}$.

Fig. 6 Algorithm summarizing the parameter expanded, blocked Gibbs sampler for the constrained data association model (Sec. 3.1). The output is the collection of samples $\{\boldsymbol{\gamma}^{(s)}, \boldsymbol{\beta}^{(s)}\}$, for $s = 1, \ldots, n_s$.

distribution of the samples closely approximates the true distribution, provided that the Markov chain is *ergodic* and satisfies the *detailed balance equation* [44].

Kück *et al.* [22] develop an MCMC algorithm for sampling from the posterior by augmenting the original blocked Gibbs sampler [49] to the data association scenario. Gibbs samplers are generally easy to implement and prove their correctness, but they can be extremely slow to converge to the true distribution if the random variables exhibit strong correlation—and strong correlations certainly abound here. For instance, the selection of the prototypes via $\boldsymbol{\gamma}$ is strongly influenced by the choice of regression coefficients $\boldsymbol{\beta}$, and vice versa. Thus, we implement all the known strategies for acceleration [22,49], including blocked moves and reparameterization (Sec. 4.2). In addition, we use the Schur complement [6] to derive fast matrix updates (Sec. 4.3). One key difference is that [22] 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.

We spend the rest of this section developing the MCMC algorithm that produces correlated samples from the posterior. An overall recipe is provided in Fig. 6, which may serve as guidance throughout the rest of the section. Since the algorithm is quite involved, we split up the derivation into sections. In Sec. 4.1, we derive the blocked Gibbs sampler with the exception of the sampler for $\boldsymbol{\gamma}$, which we leave until Sec. 4.3 since it is an elaborate step. Sec. 4.2 improves upon the expected convergence rate of the Gibbs sampler through introduction of an expansion parameter.

4.1 Generalized blocked Gibbs sampler

If two variables are highly correlated, an ordinary two-stage Gibbs sampler [44] will slowly navigate through the joint space because it samples from the full conditional distributions. We will see shortly that we can analytically derive the posterior of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ jointly, conditioned on $\boldsymbol{z}$ and δ^2, resulting in so-called "blocked" moves which converge faster than a Markov chain derived from the full conditionals [29]. We factorize the joint conditional of the regression coefficients and variable selection parameters as

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma} \,|\, \boldsymbol{x}, \boldsymbol{z}, \delta^2, a, b) = p(\boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\gamma}, \delta^2)\, p(\boldsymbol{\gamma} \,|\, \boldsymbol{x}, \boldsymbol{z}, \delta^2, a, b). \tag{13}$$

Following Bayes' rule, a straightforward derivation shows that the conditional posterior for sampling the regression coefficients is

$$\begin{aligned} p(\boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\gamma}, \delta^2) &\propto p(\boldsymbol{z} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\beta})\, p(\boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \delta^2) \\ &= \mathcal{N}\left(\boldsymbol{\beta} \;\middle|\; \mathbf{Q}_{\boldsymbol{\gamma}} \boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T \mathbf{Z}, \mathbf{Q}_{\boldsymbol{\gamma}}\right). \end{aligned} \tag{14}$$

where $\mathbf{Z} = [z_1, \ldots, z_N]^T$, $\mathbf{Q}_{\boldsymbol{\gamma}} = (\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T \boldsymbol{\Psi}_{\boldsymbol{\gamma}} + (\delta^2 \mathbf{S}_{\boldsymbol{\gamma}})^{-1})^{-1}$ and $\mathbf{S}_{\boldsymbol{\gamma}}$ is defined in Sec. 2.

We can sample the z_i^k's easily since they are independent of one other. Again, we apply Bayes' rule to obtain the posterior distribution for z_i^k conditioned on assignments to the rest of the unknowns. When the label is positive, we have

$$\begin{aligned} p(z_i^k \,|\, x_i, y_i^k = 1, \boldsymbol{\gamma}, \boldsymbol{\beta}) &\propto p(y_i^k = 1 \,|\, z_i^k)\, p(z_i^k \,|\, x_i, \boldsymbol{\gamma}, \boldsymbol{\beta}) \\ &= \mathcal{N}(\boldsymbol{\Psi}_{\boldsymbol{\gamma},i} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1)\, \mathbb{I}_{(0,+\infty)}(z_i^k). \end{aligned} \tag{15a}$$

Otherwise, when $y_i^k = -1$ the conditional posterior is

$$p(z_i^k \,|\, x_i, y_i^k = -1, \boldsymbol{\gamma}, \boldsymbol{\beta}) = \mathcal{N}(\boldsymbol{\Psi}_{\boldsymbol{\gamma},i} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1)\, \mathbb{I}_{(-\infty,0]}(z_i^k). \tag{15b}$$

The conditional posterior of z_i^k is a Normal density truncated either to the right or to the left of the origin. Efficient methods have been developed for drawing samples from this distribution [43].

Sampling the z_i^u's when the labels are unknown is not quite as simple because the joint posterior does not factorize nicely. While [22,23] use rejection sampling to sample the unknown labels subject to the constraints or group statistics, we adopt a more efficient MCMC scheme and sample from the full conditionals in each document. For the augmented model with constraints (Sec. 3.1), we need to consider three cases. When the number of positive labels—not counting the ith label—is less than $n_{(+)}$, the new sample z_i^u is required to be positive, thus it is drawn according to the left-truncated Normal (15a). When the number of negative labels other than y_i^u is less than $n_{(-)}$, the sampling distribution is the right-truncated Normal (15b). Finally, we need to consider a third case when the remaining labels satisfy all the constraints, so that either $y_i^u = 1$ or $y_i^u = -1$ is allowed; the full conditional is then simply

$$p(z_i^u \,|\, \{z_{-i}^u\}, \boldsymbol{\gamma}, \boldsymbol{\beta}, x_i) = \mathcal{N}(\boldsymbol{\Psi}_{\boldsymbol{\gamma},i} \boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1), \tag{16}$$

where $i \in d_j$, and $\{z^u_{-i}\}$ refers to the collection of latent variables z_i in a particular image with the exception of z^u_i. Note that at least one of $C_{(-)}$ and $C_{(+)}$ is always satisfied provided $\boldsymbol{z}$ is initialized within the region of support.

Following Bayes' rule and the conditional identity, the Gibbs sampling step for data association with group statistics (Sec. 3.2) is given by

$$\begin{aligned} p(z^u_i \,|\, x_i, \{z^u_{-i}\}, \boldsymbol{\gamma}, \boldsymbol{\beta}, m_j, \chi) \\ &\propto p(z_i \,|\, x_i, \boldsymbol{\beta}, \boldsymbol{\gamma})\, p(m_j \,|\, z^u_i, \{z^u_{-i}\}, \chi) \\ &\propto \mathcal{N}(z_i \,|\, \boldsymbol{\Psi}_{\boldsymbol{\gamma},i}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, 1)\, \mathrm{Beta}(m_j \,|\, \chi\lambda_j{+}1, \chi(1-\lambda_j){+}1). \end{aligned} \tag{17}$$

This density is a weighted mixture of a left-truncated and a right-truncated Normal density.

We sample the variance parameter δ^2 according to its conditional posterior,

$$\begin{aligned} p(\delta^2 \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \mu, \nu) &\propto p(\boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \delta^2)\, p(\delta^2 \,|\, \mu, \nu) \\ &= \mathcal{IG}\big(\tfrac{1}{2}(\mu + \Sigma\boldsymbol{\gamma}), \tfrac{1}{2}\big(\nu + \boldsymbol{\beta}_{\boldsymbol{\gamma}}^T \mathbf{S}_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\beta}_{\boldsymbol{\gamma}}\big)\big), \end{aligned} \tag{18}$$

where we define $\Sigma\boldsymbol{\gamma} = \sum_{k=1}^N \gamma_k$ to be the number of active kernels.

4.2 Parameter expansion

The convergence rate of the Gibbs sampler suffers from high correlation between parameter $\boldsymbol{\beta}$ and the latent variables $\boldsymbol{z}$ [29]. We start by introducing a scaled version of the latent variables,

$$w_i = \alpha z_i \tag{19}$$

with auxiliary parameter α. The idea is then to come up with a new, overparameterized model $\pi(\boldsymbol{y}, \boldsymbol{w} \,|\, \boldsymbol{x}, \boldsymbol{\theta}, \alpha)$ that agrees with the original model (7) but has more desirable variance properties. By agreeing, we mean the new model conditioned on the labels $\boldsymbol{y}$ should satisfy

$$\int \pi(\boldsymbol{y}, \boldsymbol{w} \,|\, \boldsymbol{x}, \boldsymbol{\theta}, \alpha)\, d\boldsymbol{w} = \int p(\boldsymbol{y}, \boldsymbol{z} \,|\, \boldsymbol{x}, \boldsymbol{\theta})\, d\boldsymbol{z}. \tag{20}$$

Here we denote the variables transformed under the scaling (19) by $\boldsymbol{w} = \{w_1, \ldots, w_N\}$. If we specify the new, parameter expanded prior as $\pi(\boldsymbol{\theta}, \alpha) = p(\alpha \,|\, \boldsymbol{\theta})\, p(\boldsymbol{\theta})$, the posterior distribution will be the same for both models provided (20) is satisfied. We can ensure that (20) is satisified by appealing to the change of variables theorem from multivariate calculus [32], and setting

$$\pi(\boldsymbol{y}, \boldsymbol{w} \,|\, \boldsymbol{x}, \boldsymbol{\theta}, \alpha) = p(\boldsymbol{y}, \boldsymbol{z} \,|\, \boldsymbol{x}, \boldsymbol{\theta})\, |J_\alpha(\boldsymbol{w})|, \tag{21}$$

where $J_\alpha(\boldsymbol{w})$ is the Jacobian of the vector-valued function that transforms the quantities $\boldsymbol{w}$ back to their original values $\boldsymbol{z}$. Under the change of scale (19), the determinant of the Jacobian is $J_\alpha(\boldsymbol{w}) = \alpha^{-N}$.

Liu *et al.* [30] suggest placing an improper Haar prior on the scaling parameter α since it leads to optimal convergence.[5] This is also the prior used by Tham *et al.* [49]. In practice, however, the Haar prior tends to be unstable, leading to very small or very large values of α. An alternative is the inverse Gamma prior, $\alpha^2 \sim \mathcal{IG}(\frac{\mu_\alpha}{2}, \frac{\nu_\alpha}{2})$. It achieves an improved convergence rate while allowing the user to tune μ_α and ν_α for stability.

One can still sample from the expanded model (21) without knowing its exact form. A general strategy for sampling is presented in [30], and specifics behind the derivations are given in our technical report [9]. The parameter expanded Gibbs sampler for $\boldsymbol{\beta}$ and $\boldsymbol{z}$ consists of the following steps:

1. Generate a new sample $\boldsymbol{z}$ as explained in Sec. 4.1.
2. Draw a sample $\hat{\alpha}^2$ from the prior $\pi(\alpha^2 \,|\, \mu_\alpha, \nu_\alpha)$.
3. Draw a sample α^2 according to the conditional posterior $\pi(\alpha^2 \,|\, \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\gamma}, \delta^2, \mu_\alpha, \nu_\alpha)$ where $\boldsymbol{w} = \hat{\alpha} \times \boldsymbol{z}$.
4. Obtain a new sample $\boldsymbol{\beta}$ from the conditional posterior density $\pi(\boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \delta^2, \boldsymbol{z}')$ where $\boldsymbol{z}' = \hat{\alpha}/\alpha \times \boldsymbol{z}$.

The remaining steps of the Gibbs sampler are not affected by the inclusion of the expansion parameter. We obtain the following "blocked" step by employing Bayes' rule and integrating out the regression coefficients, as we do for the $\boldsymbol{\gamma}$ variables:

$$\begin{aligned} \pi(\alpha^2 \,|\, \boldsymbol{x}, \boldsymbol{w}, \boldsymbol{\gamma}, \delta^2, \mu_\alpha, \nu_\alpha) \\ &\propto \pi(\boldsymbol{w} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \delta^2, \alpha^2)\, \pi(\alpha^2 \,|\, \mu, \nu) \\ &\propto p(\boldsymbol{z} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \delta^2)\, |J_\alpha(\boldsymbol{w})|\, \pi(\alpha^2 \,|\, \mu, \nu) \\ &= \mathcal{IG}\big(\alpha^2 \,|\, \tfrac{\mu_\alpha + N}{2}, \tfrac{1}{2}\big[\nu_\alpha + \mathbf{W}^T\big(I_N - \boldsymbol{\Psi}_{\boldsymbol{\gamma}}\mathbf{Q}_{\boldsymbol{\gamma}}\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T\big)\mathbf{W}\big]\big), \end{aligned} \tag{22}$$

where $\boldsymbol{z} = \boldsymbol{w}/\alpha$, $\mathbf{W} = [w_1 \cdots w_N]^T$, $\mathbf{Q}_{\boldsymbol{\gamma}}$ is defined as before, and the marginalized likelihood term in (22) is

$$\begin{aligned} p(\boldsymbol{z} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \delta^2) &= \int p(\boldsymbol{z} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\beta})\, p(\boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \delta^2)\, d\boldsymbol{\beta} \\ &\propto \exp\big\{-\tfrac{1}{2}\big[\mathbf{Z}^T\big(I_N - \boldsymbol{\Psi}_{\boldsymbol{\gamma}}\mathbf{Q}_{\boldsymbol{\gamma}}\boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T\big)\mathbf{Z}\big]\big\}. \end{aligned}$$

The steps for sampling the expanded model are summarized in lines 3-7 of Fig. 6.

4.3 Sampling the variable selection parameter

Exact sampling from the $\boldsymbol{\gamma}$ posterior (13) conditioned on δ^2 and $\boldsymbol{z}$ is impractical because it requires evaluation of 2^N possible configurations of $\boldsymbol{\gamma}$. We outline an alternative first proposed in [48]. It is essentially a Gibbs sampler implemented using Metropolis-Hastings (M-H) proposals for more efficient computation. While sampling from the posterior of $\boldsymbol{\gamma}$ is impractical, it is possible to

[5] By improper, we mean that the prior is not a probability measure as it does not integrate to one [42]. This is often acceptable in practice because an improper prior can still lead to a proper posterior. The Haar prior in particular has strong justifications from a frequentist standpoint.

1 For $k = 1, \ldots, N$,
2 If $\gamma_k = 1$,
3 Sample $u \sim \mathcal{U}_{[0,1]}$.
4 If $u < \frac{\Sigma\boldsymbol{\gamma}_{-k}+a}{N+a+b-1}$,
5 Sample $r \sim \mathcal{U}_{[0,1]}$.
6 If $r < \mathcal{A}(\gamma_k = 1, \gamma_k^\star = 0)$, then $\gamma_k \leftarrow 0$.
7 Else if $\gamma_k = 0$,
8 Sample $u \sim \mathcal{U}_{[0,1]}$.
9 If $u < \frac{N-\sum\boldsymbol{\gamma}_{-k}+b-1}{N+a+b-1}$,
10 Sample $r \sim \mathcal{U}_{[0,1]}$.
11 If $r < \mathcal{A}(\gamma_k = 0, \gamma_k^\star = 1)$, then $\gamma_k \leftarrow 1$.

Fig. 7 Metropolized Gibbs sampler for $\boldsymbol{\gamma}$. $\mathcal{U}_{[0,1]}$ is the uniform distribution on the unit interval.

sample each γ_k from its full conditional. One could employ the Gibbs sampler to accomplish this, since the full conditional probabilities can be computed as

$$p(\gamma_k \,|\, \boldsymbol{x}, \boldsymbol{z}, \boldsymbol{\gamma}_{-k}, \delta^2, a, b) = \frac{p(\gamma_k, \boldsymbol{\gamma}_{-k} \,|\, \boldsymbol{x}, \boldsymbol{z}, \delta^2, a, b)}{p(\gamma_k = 0, \boldsymbol{\gamma}_{-k} \,|\, \ldots) + p(\gamma_k = 1, \boldsymbol{\gamma}_{-k} \,|\, \ldots)},$$

where $\boldsymbol{\gamma}_{-k}$ is the collection of variable selection parameters with the exception of the kth one.[6] Even though this Gibbs sampler is a vast improvement over sampling directly from the conditional posterior of $\boldsymbol{\gamma}$, it is still costly since each of the N steps requires the evaluation of two terms in the denominator, each which involves the inversion of a large matrix. Thus, it is worth our while to reduce the expense of these computations as much as possible.

The main idea is to propose a change to $\boldsymbol{\gamma}$ infrequently, and to use Metropolis-Hastings to correct for any discrepancy between our proposal mechanism and the posterior distribution. A single step of the Metropolis-Hastings algorithm [3,11] consists of sampling a candidate value $\boldsymbol{\gamma}^\star$ from a proposal distribution $q(\boldsymbol{\gamma}^\star \,|\, \boldsymbol{\gamma})$, and then accepting the candidate $\boldsymbol{\gamma}^{\text{new}} \leftarrow \boldsymbol{\gamma}^\star$ with probability $\mathcal{A}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^\star)$. Otherwise, the new sample remains unchanged, and $\boldsymbol{\gamma}^{(\text{new})} \leftarrow \boldsymbol{\gamma}$. The M-H acceptance probability is given by

$$\mathcal{A}(\boldsymbol{\gamma}, \boldsymbol{\gamma}^\star) = \min\left\{1, \frac{p(\boldsymbol{\gamma}^\star \,|\, \boldsymbol{x}, \boldsymbol{z}, \delta^2, a, b)\, q(\boldsymbol{\gamma} \,|\, \boldsymbol{\gamma}^\star)}{p(\boldsymbol{\gamma} \,|\, \boldsymbol{x}, \boldsymbol{z}, \delta^2, a, b)\, q(\boldsymbol{\gamma}^\star \,|\, \boldsymbol{\gamma})}\right\}.$$

We need to consider two cases for the Metropolized Gibbs sampler: when $\gamma_k = 0$ and when $\gamma_k = 1$. When kernel k is inactive in the current sample $\boldsymbol{\gamma}$, our proposal consists of flipping γ_k to 1 with probability proportional to the prior:

$$\begin{aligned} p(\gamma_k^\star = 1 \,|\, \boldsymbol{\gamma}_{-k}, a, b) &\propto p(\gamma_k^\star = 1, \boldsymbol{\gamma}_{-k} \,|\, a, b) \\ &= \frac{p(\gamma_k = 1, \boldsymbol{\gamma}_{-k} | a, b)}{p(\gamma_k = 1, \boldsymbol{\gamma}_{-k} | a, b) + p(\gamma_k = 0, \boldsymbol{\gamma}_{-k} | a, b)} \\ &= \frac{\Sigma\boldsymbol{\gamma}_{-k} + a}{N + a + b - 1}, \end{aligned} \tag{23}$$

[6] Note the terms in the denominator do not sum to one.

where $\sum\boldsymbol{\gamma}_{-k} = \sum_{k' \neq k} \gamma_{k'}$ is the number of active kernels, not counting the kth one. When a is much smaller than b and the number of active kernels is small compared to N, so the proposal is unlikely to activate an inactive kernel, and hence the expensive part—computing the M-H acceptance probability (step 11 of the algorithm in Fig. 7)—is avoided. Depending on the strength of the prior, the Metropolized Gibbs sampler can filter out a lot of poor candidates while maintaining a desirable M-H acceptance rate. When a flip is proposed, the change is accepted with probability

$$\mathcal{A}\big(\gamma_k = 0, \gamma_k^\star = 1\big) = \min\left\{1, \frac{p(\boldsymbol{z} \,|\, \boldsymbol{x}, \gamma_k^\star = 1, \boldsymbol{\gamma}_{-k}, \delta^2)}{p(\boldsymbol{z} \,|\, \boldsymbol{x}, \gamma_k = 0, \boldsymbol{\gamma}_{-k}, \delta^2)}\right\}. \tag{24}$$

The acceptance ratio reduces to a ratio of likelihoods because the proposal terms cancel with the prior terms. The likelihood terms that appear in (24) are given by the following expression, discarding factors that do not implicate the variable selection parameter $\boldsymbol{\gamma}$:

$$\begin{aligned} &p(\boldsymbol{z} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \delta^2) \\ &\quad = \textstyle\int p(\boldsymbol{z} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \boldsymbol{\beta})\, p(\boldsymbol{\beta} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}, \delta^2)\, d\boldsymbol{\beta} \\ &\quad = \left((2\pi)^N \left|\delta^2 \mathbf{S}_{\boldsymbol{\gamma}} \mathbf{Q}_{\boldsymbol{\gamma}}^{-1}\right| \exp\left\{\left[\mathbf{Z}^T \left(I_N - \boldsymbol{\Psi}_{\boldsymbol{\gamma}} \mathbf{Q}_{\boldsymbol{\gamma}} \boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T\right) \mathbf{Z}\right]\right\}\right)^{-1/2} \\ &\qquad \times \textstyle\int \mathcal{N}(\boldsymbol{\beta} \,|\, \mathbf{Q}_{\boldsymbol{\gamma}} \boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T \mathbf{Z}, \mathbf{Q}_{\boldsymbol{\gamma}})\, d\boldsymbol{\beta} \\ &\quad \propto \left|\tfrac{1}{\delta^2} \mathbf{Q}_{\boldsymbol{\gamma}} \mathbf{S}_{\boldsymbol{\gamma}}^{-1}\right|^{-1/2} \exp\left\{\tfrac{1}{2} \mathbf{Z}^T \boldsymbol{\Psi}_{\boldsymbol{\gamma}} \mathbf{Q}_{\boldsymbol{\gamma}} \boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T \mathbf{Z}\right\}. \end{aligned} \tag{25}$$

Naively, evaluation of the acceptance ratio (24) involves computing the inverse of two large matrices in $p(\boldsymbol{z} \,|\, \boldsymbol{x}, \boldsymbol{\gamma}^\star, \delta^2)$, $\mathbf{S}_{\boldsymbol{\gamma}^\star}$ and $\mathbf{Q}_{\boldsymbol{\gamma}^\star}$. We can do better by decomposing the inverse of $\mathbf{Q}_{\boldsymbol{\gamma}^\star}$ into a 2×2 block matrix,

$$\begin{aligned} \mathbf{Q}_{\boldsymbol{\gamma}^\star}^{-1} &= \boldsymbol{\Psi}_{\boldsymbol{\gamma}^\star}^T \boldsymbol{\Psi}_{\boldsymbol{\gamma}^\star} + (\delta^2 \mathbf{S}_{\boldsymbol{\gamma}^\star})^{-1} \\ &= \tfrac{1+\delta^2}{\delta^2} \boldsymbol{\Psi}_{\boldsymbol{\gamma}^\star}^T \boldsymbol{\Psi}_{\boldsymbol{\gamma}^\star} + \tfrac{\epsilon}{\delta^2} I \\ &= \begin{bmatrix} \mathbf{Q}_{\boldsymbol{\gamma}}^{-1} & \frac{1+\delta^2}{\delta} \boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T \boldsymbol{\Psi}_{\gamma_k^\star} \\ \frac{1+\delta^2}{\delta} \boldsymbol{\Psi}_{\gamma_k^\star}^T \boldsymbol{\Psi}_{\boldsymbol{\gamma}} & \frac{1+\delta^2}{\delta} \boldsymbol{\Psi}_{\gamma_k^\star}^T \boldsymbol{\Psi}_{\gamma_k^\star} + \frac{\epsilon}{\delta^2} \end{bmatrix}, \end{aligned} \tag{26}$$

where $\boldsymbol{\Psi}_{\gamma_k^\star}$ is newly activated column of the kernel matrix. Notice that the bottom-right entry of (26) is a scalar. Invoking the formula for the inverse of a block matrix [6], $\mathbf{Q}_{\boldsymbol{\gamma}^\star}$ resolves to

$$\mathbf{Q}_{\boldsymbol{\gamma}^\star} = \begin{bmatrix} \mathbf{Q}_{\boldsymbol{\gamma}} + v^2 d_k \mathbf{Q}_{\boldsymbol{\gamma}} \boldsymbol{A}_k \boldsymbol{A}_k^T \mathbf{Q}_{\boldsymbol{\gamma}} & -v d_k \mathbf{Q}_{\boldsymbol{\gamma}} \boldsymbol{A}_k \\ -v d_k \boldsymbol{A}_k^T \mathbf{Q}_{\boldsymbol{\gamma}} & d_k \end{bmatrix}, \tag{27}$$

where $v = \frac{1+\delta^2}{\delta^2}$, $\boldsymbol{A}_k = \boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T \boldsymbol{\Psi}_{\gamma_k^\star}$ and

$$d_k^{-1} = v \boldsymbol{\Psi}_{\gamma_k^\star}^T \boldsymbol{\Psi}_{\gamma_k^\star} + \tfrac{\epsilon}{\delta^2} - v^2 \boldsymbol{A}_k^T \mathbf{Q}_{\boldsymbol{\gamma}} \boldsymbol{A}_k$$

is the *Schur complement* of $\mathbf{Q}_{\boldsymbol{\gamma}}^{-1}$. We obtain an iterative expression for $\mathbf{S}_{\boldsymbol{\gamma}^\star}$ in an analogous manner. Plugging (27) into the ratio (24), we obtain the expression

$$\begin{aligned} \mathcal{A}\big(\gamma_k = 0, \gamma_k^\star = 1\big) = \min\Big\{&1, \sqrt{d_k / (c_k \delta^2)} \\ &\times \exp\left(\tfrac{1}{2} d_k \left(\mathbf{Z}^T \left(I_N - v \boldsymbol{\Psi}_{\boldsymbol{\gamma}} \mathbf{Q}_{\boldsymbol{\gamma}} \boldsymbol{\Psi}_{\boldsymbol{\gamma}}^T\right) \boldsymbol{\Psi}_{\gamma_k^\star}\right)^2\right)\Big\}, \end{aligned} \tag{28a}$$

where $c_k^{-1} = \boldsymbol{\Psi}_{\gamma_k^\star}^T \boldsymbol{\Psi}_{\gamma_k^\star} + \epsilon - \boldsymbol{A}_k^T \mathbf{S}_{\boldsymbol{\gamma}} \boldsymbol{A}_k$ is the Schur complement of $\mathbf{S}_{\boldsymbol{\gamma}}^{-1}$.

When kernel k is active, it is deactivated with probability

$$p(\gamma_k^\star = 0 \,|\, \boldsymbol{\gamma}_{-k}, a, b) \propto p(\gamma_k^\star = 0, \boldsymbol{\gamma}_{-k} \,|\, a, b) = \frac{N - \sum \boldsymbol{\gamma}_{-k} + b - 1}{N + a + b - 1},$$

and then accept the change with probability equal to

$$\mathcal{A}(\gamma_k = 1, \gamma_k^\star = 0) = \min\left\{1, \frac{p(\boldsymbol{z} \,|\, \boldsymbol{x}, \gamma_k^\star = 0, \boldsymbol{\gamma}_{-k}, \delta^2)}{p(\boldsymbol{z} \,|\, \boldsymbol{x}, \gamma_k = 1, \boldsymbol{\gamma}_{-k}, \delta^2)}\right\} = \min\Big\{1, \sqrt{\delta^2 c_k^\star / d_k^\star} \times \exp\left(-\tfrac{1}{2} d_k^\star \left(\mathbf{Z}^T \left(I_N - v \boldsymbol{\Psi}_{\boldsymbol{\gamma}^\star} \mathbf{Q}_{\boldsymbol{\gamma}^\star} \boldsymbol{\Psi}_{\boldsymbol{\gamma}^\star}^T\right) \boldsymbol{\Psi}_{\gamma_k}\right)^2\right)\Big\}, \tag{28b}$$

where the scalar $c_k^\star$ is the inverse of the Schur complement of $\mathbf{S}_{\boldsymbol{\gamma}^\star}^{-1}$ and $d_k^\star$ is the inverse of the Schur complement of $\mathbf{Q}_{\boldsymbol{\gamma}^\star}^{-1}$. Both these values are easily recovered as the kth row, kth colum entry of their respective matrices, basically by reversing the block operations detailed above. The Metropolized Gibbs sampler is summarized in Fig. 7. Note that a proper implementation should treat boundary cases.

The computational bottleneck of our MCMC algorithm is the sampling of the variable selection parameter $\boldsymbol{\gamma}$. By manipulating the order of the matrix computation, we have shown that it is possible to compute the necessary acceptance probabilities (28) with $O(N)$ elementary operations, provided the model has selected only a small number of prototypes. In the worst case—when the Metropolization trick offers little advantage and when a lot of the kernels are active—the computational complexity of the learning algorithm approaches $O(N^3)$. In practice, we never come close to the worst case scenario because both the prior and data favour sparsity.

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. 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 that our localization scheme applies to a wide variety of object classes.

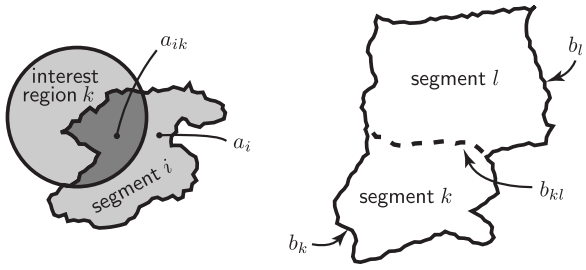

Fig. 8 Diagrams illustrating how the quantities a_i, a_{ik}, b_k, b_l and b_{kl} are computed for the conditional random field. The contour length b_k includes both the solid and dashed lines surrounding segment k, and likewise for the length b_l.

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 [41]. We use the Normalized Cuts algorithm [46] to segment images, but other, less expensive methods could possibly be used with similar returns. An example low-compression segmentation produced by Normalized Cuts is shown in Fig. 9. The extracted features of small segments are hardly sufficient for locating object classes in cluttered scenes, so the next step is to construct a conditional random field [25] (CRF) that propagates information across an image's neighbouring superpixels and interest regions. CRFs have been shown to be effective models for image classification and region labeling [24, 40].

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

$$\phi_k(y_k^s) = \sum_i \frac{a_{ik}}{a_i} \delta(y_k^s = y_i), \tag{29}$$

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. Since there is no overlap between most segments and interest regions, most segment labels are only influenced by small number of labels y_i. See the left of Fig. 8 for an example that demonstrates how the quantities a_i and a_{ik} are computed.

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

$$\mu_{kl}(y_k^s, y_l^s) = \eta + \frac{b_{kl}}{2}\left(\frac{1}{b_k} + \frac{1}{b_l}\right)\delta(y_k^s = y_l^s), \tag{30}$$

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. See

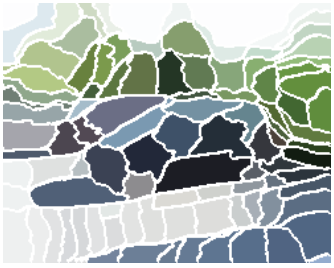

Fig. 9 Superpixels for the third image in Fig. 1.

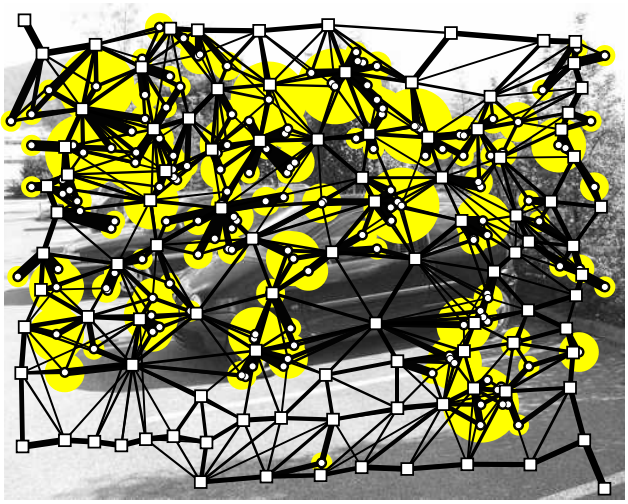

Fig. 10 Spatial layout of the conditional random field for the third image in the token data set (see Fig. 1). White squares correspond to superpixels (see Fig. 9). Black lines are drawn between neighbouring superpixels, such that thicker lines represent stronger interactions, hence stronger enforcement of label compatibility. Yellow circles depict the extracted image keypoints at their characteristic scale. A thick black line between a white circle i and a white square k means that the label y_k^s is strongly encouraged to be equal to the keypoint label y_i.

the right-hand side of Fig. 8 for an example. The pairwise potential (30) is the prior compatibility of the labels of neighbouring segments. For instance, if two identical, neighbouring segments share the same label, and the shared border is one half of the segment contour lengths, the response (30) resolves to $\eta + \frac{1}{2}$; when the labels differ, the potential is η. A large value of η means the prior only infirmly enforces compatibility.

An unpublished theorem by Hammersley and Clifford tells us that

$$p(\boldsymbol{y}^s \mid \boldsymbol{y}) = \frac{1}{Z(\boldsymbol{y})} \prod_k \phi_k(y_k^s) \prod_l \mu_{kl}(y_k^s, y_l^s), \qquad (31)$$

defines a joint probability over the segment labels $\boldsymbol{y}^s$ provided the potentials are always positive and the partition function $Z(\boldsymbol{y})$ is chosen so that (31) sums to unity [5]. The partition function is given by

$$Z(\boldsymbol{y}) = \sum_{\boldsymbol{y}^s} \prod_k \phi_k(y_k^s) \prod_l \mu_{kl}(y_k^s, y_l^s).$$

What we have defined in (31) is a probability density over the segmented labels $\boldsymbol{y}^s$ *conditioned* on known values—obtained by simulations—of the interest region labels $\boldsymbol{y}$. This is our conditional random field. Fig. 10 gives an example of the spatial layout of the conditional random field for an image containing a car. Notice that some of the segment labels (squares) are not directly influenced by any inferred labels y_i (circles), but the influence is still propagated through the potentials of the CRF. In other words, all the segment labels are dependent on each other.

There is only a single parameter η controlling the strength of the potentials. At this point, there is no learning; we tune the parameter by hand. In our experiments, we set η to a relatively strong prior, $\frac{1}{10}$, encouraging neighbouring segments to have the same labels.

Even though equation (31) involves a product over all pairs (k, l) of segments in the image, the adjacency graph is sparse since only a few superpixels share a common border, and as a result it is reasonable to use an inference algorithm designed for sparse graphs. We use the tree sampling algorithm of [19] 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 for 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 [34] which finds corner-like features, the Kadir-Brady detector [20] which proposes circular regions with maximum grey-level entropy, and the Laplacian method [28] which detects blob-like structures. Based on earlier studies [35], we chose the Scale Invariant Feature Transform (SIFT) [31] 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. A library for detecting Harris-Laplace and Laplacian interest regions is available on the Web at lear.inrialpes.fr/software. This library also includes SIFT routines. A MATLAB implementation of the Kadir-Brady detector was obtained from www.robots.ox.ac.uk/~timork/salscale.html. 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 that Fergus *et al.* [18] extract only 20 features per image on average, owing in part to the expense of training, while Opelt *et al.* [38] 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. For instance, we have a rough idea that cars in the INRIA data set often occupy a small portion of the scene, so it is reasonable to set a small constraint ($n_1 = 15$) for the INRIA car experiments in Sec. 6.2. 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. Clearly, the constraints do have an impact on prediction, but experience dictates that small constraints are sufficient to recover good results. See Sec. 6.3 for further experiments and discussion pertaining to this issue. When employing the group statistics model (Sec. 3.2), we set the parameters to be approximately $m = 0.3$ and $\chi = 400$. They were modified slightly according to the variability exhibited in the data set. As we discussed in Sec. 3.1, it is assumed that the user is able to assign the prior in an informed manner.

We set $a = 1$ and b according to a feature selection prior of approximately 200 active kernel centres (thus the prior depends on N), and we bestow near uninformative priors on the rest of the model parameters: $\mu = \nu = 0.01$ and $\mu_\alpha = \nu_\alpha = 0.01$. In all our experiments, we set σ to 0.01 with $\epsilon = 0.01$ 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 ($n_s = 1900$) 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 learning stage, which involved running the blocked Gibbs sampler for 2000 iterations, took a couple hours on the smallest data sets, and over a day on data sets composed of almost a thousand images.

We have made the code and data for all our experiments available at 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, "INRIA cars", consisting of images arising from the same environment: parking lots with and without cars.

	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
cars (rear)	400	400	400	400
cars (side)	360	450	360	450
faces	218	450	217	450
bicycles	100	100	50	50
people	100	100	50	50
INRIA cars	50	50	29	21

Table 1 Summary of experiment data. The sources for the nine image databases are the following: the Caltech motorbikes (side) and airplanes (side) were obtained from www.vision.caltech.edu/html-files/archive.html, the wildcats come from the Corel Image database, the cars (side) data set was composed of the Caltech background images and other images available at l2r.cs.uiuc.edu/∼cogcomp/Data/Car, the Graz bicycles and people data sets are archived at www.emt.tugraz.at/∼pinz/data/GRAZ_01, and the INRIA car database can be downloaded from lear.inrialpes.fr/data.

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, cars, motorbikes, wildcats, bicycle, faces and people. The experiment data is summarized in Table 1. Note that the "cars (side)" data set is the only data set in which positive labels are observed during training, because there are isolated instances of the object.

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 Receiver Operating Characteristic (ROC) equal error rate, a standard evaluation criterion [18,38]. 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 we found the constraints to be generally easier to specify. We omitted error bars in our results because independent MCMC simulations with the same choice of priors exhibited little variance. Table 2 compares the results of our model with those obtained from several previously proposed methods: the boosting method proposed in [38], the constellation model described in [18], the "bag-of-keypoints" method developed at XEROX [53], the discriminative representations based on histograms of image patches [13], and the support vector machine classifier using SIFT and SPIN descrip-

data set	H-L	K-B	LoG	Combo	MT	Fergus	Opelt	Xerox	Deselaers	Zhang
airplanes	0.985	0.993	0.938	0.998	0.955	0.902	0.889	0.971	0.986	0.988
motorbikes	0.988	0.998	0.983	1.000	0.990	0.925	0.922	0.980	0.989	0.985
wildcats	0.960	0.980	0.930	0.990	0.940	0.900	—	—	—	0.970
cars (rear)	0.995	0.990	0.975	0.998	0.980	0.903	—	0.986	—	0.983
cars (side)	0.958	*0.875*	0.964	*0.969*	0.928	0.885	0.830	0.873	—	0.950
faces	0.972	0.935	0.963	0.963	0.871	0.964	0.935	0.993	0.963	1.000
bicycles	0.920	0.880	0.840	0.900	0.860	—	0.865	—	—	0.920
people	0.800	0.740	0.840	0.820	0.780	—	0.808	—	—	0.880
INRIA cars	0.966	0.897	0.897	0.931	0.793	—	—	—	—	—

Table 2 Image classification performance on test sets measured using the ROC equal error rate. The last five columns refer to performance reported by Fergus *et al.* [18], Opelt *et al.* [38], Willamowski *et al.* [53] ("Xerox"), Deselaers *et al.* [13], and Zhang *et al.* [56]. The fifth column "MT" is our implementation of the statistical machine translation model of [16] with a vocabulary obtained by quantizing the Harris-Laplace interest regions into 1000 clusters. 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 (H-L) [34], Kadir-Brady (K-B) entropy detector [20], Laplacian of Gaussians (LoG) [28], and combination of the three detectors (Combo).

tors [56]. All methods except [18] use SIFT descriptors. The last approach combines SIFT with rotationally invariant descriptors based on "spin images" [26].

In addition, as a baseline comparison we implemented the statistical machine translation model for object recognition described in [16]. The machine translation model ("MT") is designed to handle multi-category classification, so our image classification task poses as a special case. It is a mixture model which handles the correspondence problem in unlabeled images through latent variables. However, there are two principal differences between the machine translation model and our Bayesian kernel machine: 1) the feature vectors must be quantized with k-means in order to obtain a vocabulary of discrete tokens, or "blobs" [16], 2) simulation is a notoriously difficult problem for mixture models, so EM is used to approximate the posterior by a single sample. As suggested by [47], we generated a vocabulary of size 1000 for each of the data sets. Despite the apparent simplicity of the machine translation model, obtaining a single sample with EM sometimes took nearly as long as obtaining 2000 samples of our model posterior using MCMC. The reason is that the number of active kernels in the MCMC simulations tended to be considerably smaller than 1000.

Our model in combination with the three detectors often produces the best image classification, at least when comparisons with other methods are available. One exception is the Xerox classifier which performed better on the faces data set [53]. The other exception is the classifier recently proposed by Zhang *et al.* [56]; their method is noticeably better at predicting the presence of people and faces in images. We suspect that the rotationally invariant SPIN descriptors enable their object representations to achieve better generalization in certain cases. All the results of image classification should be taken with some reservation; in most of these experiments, it is not so clear whether classification properly validates recognition.

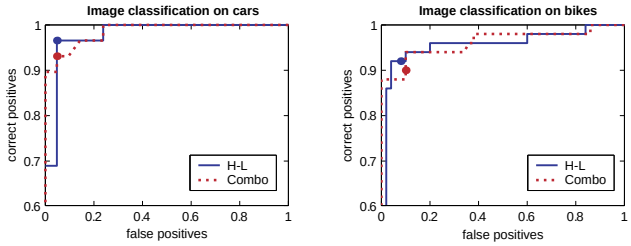

Fig. 11 The graph on the left plots the ROC curve for classification performance of car test images using the Harris-Laplace detector (blue solid line) and the combination of three detectors (red dotted line). The graph on the right shows analogous results for the bicycles test set. In both cases, the 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.

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.

Another surprising result is that the "baseline" method—the statistical machine translation model of [16]—occasionally outperforms existing methods. This suggests that the image classification task on some of the data sets is less challenging than on others. The figures reported in the last row of Table 2 confirms our hypothesis that the INRIA car data set offers a greater object recognition challenge. In spite of the difficulties posed by this data set, our model does very well in classifying the images.

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 (hence the results in

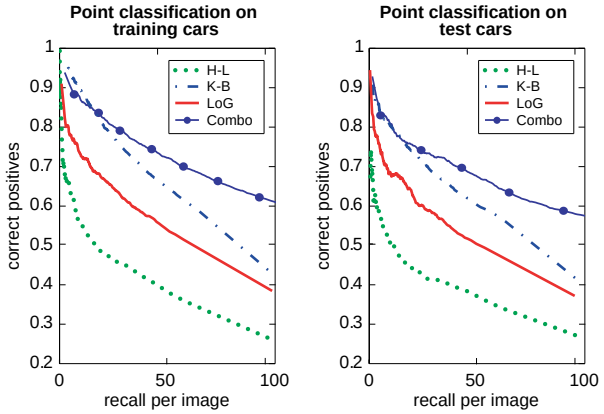

Fig. 12 Plots of precision (percentage of correct positives) versus average recall per image for the task of labeling individual features as belonging to cars. Our definition of recall here is not standard since we do not divide by the number of regions in the image. The combination scenario extends 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).

Table 2 should be viewed with a dose of skepticism). If we examine the full ROC plots in Fig. 11, 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. 12 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. 12 the Harris-Laplace detector is overly penalized because it often selects corner-like features that are near, but not on, cars. Fig. 13 shows a couple examples where learning a model with a combination of detectors results 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. 14. Incorrectly classified images tended to be unlike any of the images observed during training, such as the van and the child's bicycle. Also, problematic images tended to exhibit unusual illumination conditions.

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. Fig. 15 gives two examples of 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.

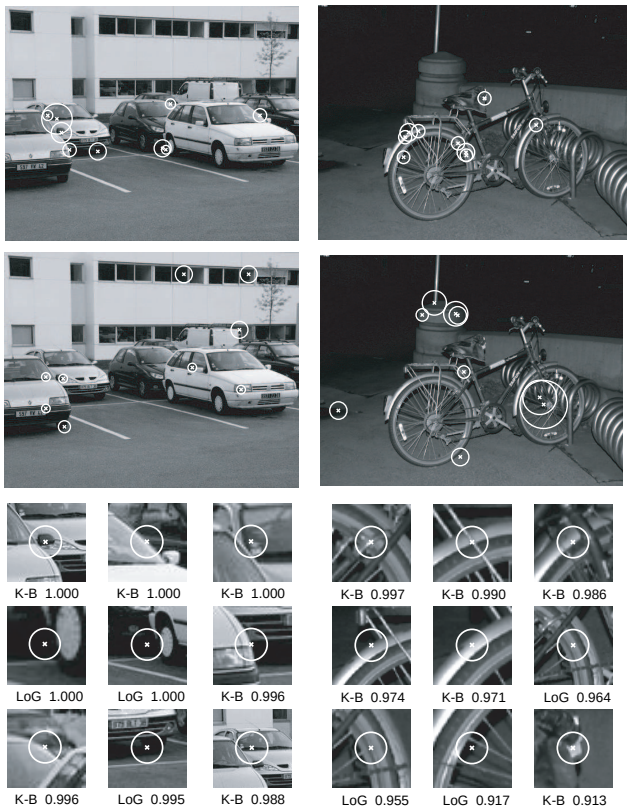

Fig. 13 Two examples in which the combination of detectors (top row) results in improved image classification over the Harris-Laplace detector (middle row). The circles represent the 9 interest regions that are most likely to belong to cars or bicycles. The bottom row shows the top features along with 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 frame the the investigation as follows: if manual segmentations were provided, how much would it be an improvement over image caption data? The answer certainly 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 test 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 degree of noise associated with this process, since an interest region near a car may or may nor correspond to the car category. The results are presented in Fig. 16. 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. However, further upgrades in supervision result in al-

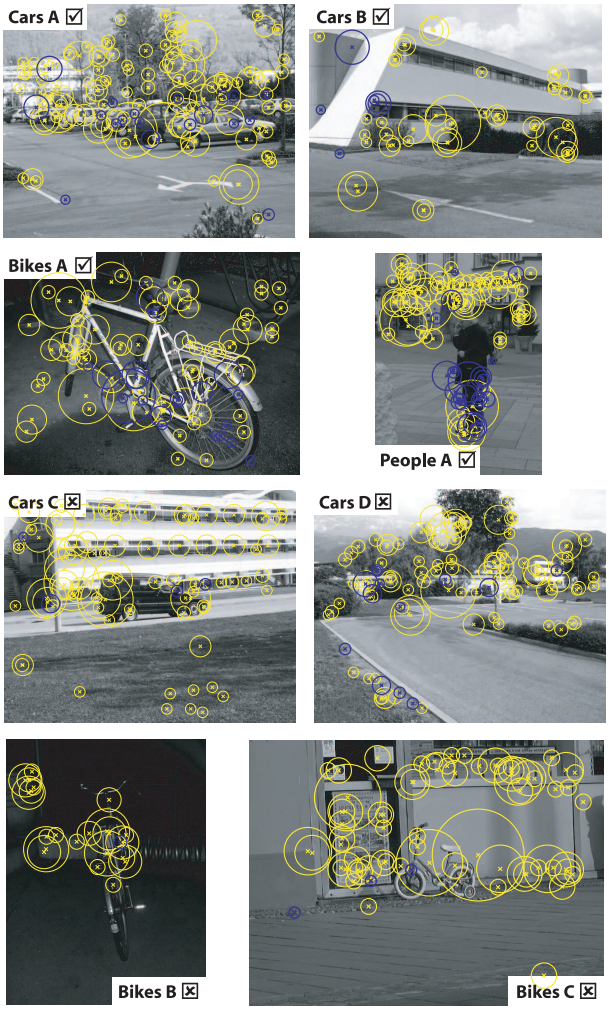

Fig. 14 Test images correctly (top four images) and incorrectly (bottom four) classified using interest regions extracted by the Harris-Laplace (for cars and bicycles) or LoG detector (for people). Dark blue circles represent local interest regions that are more likely to belong to the object, while yellow circles more probably belong to the background.

most 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. 17 demonstrates this effect on a single image.

6.4 Object localization

In this section, we evaluate the proposed conditional random field for object localization. Precisely, object localization is the task of segmenting the object from the background; it is often referred to as "region labeling" [24]. This contrasts with a stronger notion of object localization, in which individual instances of the object must be isolated from the background; as such, a successful localization implies correctly counting the number of objects in the scene. This problem is more considerably difficult to solve and its evaluation is more arbitrary [1, 17, 54].

Fig. 15 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.

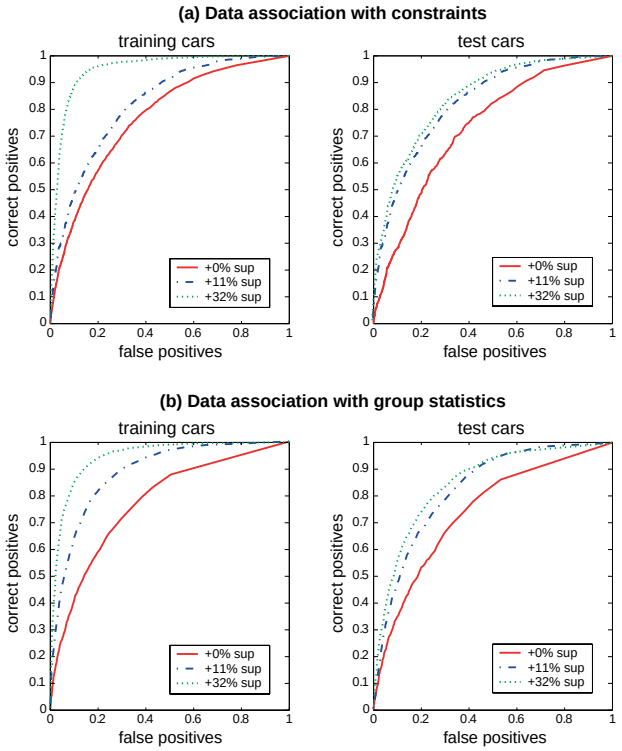

Fig. 16 The ROC plots demonstrate how learning with different proportions of hand-labeled points affects performance on labeling individual car features. *(a)* Labeling accuracy using the constrained data association model (Sec. 3.1). *(b)* Labeling accuracy using the data association with group statistics 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.

In order to quantify the accuracy of the conditional random field model, we compare the object-background segmentation predicted by the model with those drawn by hand. Some examples of manual segmentations are shown in Fig. 18. Perfect localization requires: 1.) that 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

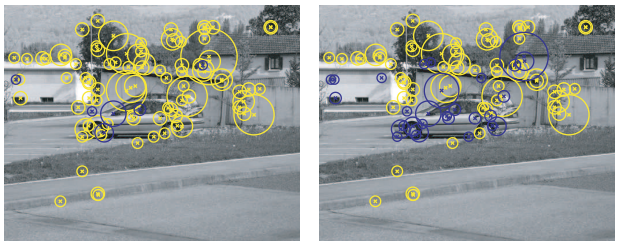

Fig. 17 Labeling of individual interest regions using the model augmented with data association constraints. The model was trained with various levels of supervision (see Fig. 16). *Left:* Car test image, no observed car labels during training. *Right:* The same image, except that the model was trained 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.

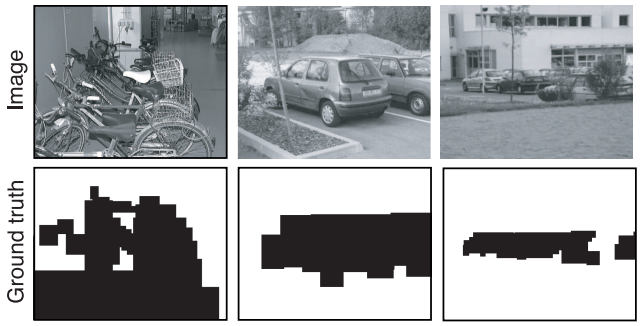

Fig. 18 Examples of ground truth segmentations from the bicycle and car databases.

contain some error, as evidenced by the examples in Fig. 18.

The ROC curves in Fig. 19 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 CRF" results in Fig. 19 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. 19 show that our method is more reliable for locating cars in images. Without the CRF, Fig. 19a 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. 20, and some less successful car recognition results are shown in Fig. 21. We observe poor localization when the interest regions and superpixels fail to complement each other. We did not tailor the CRF to

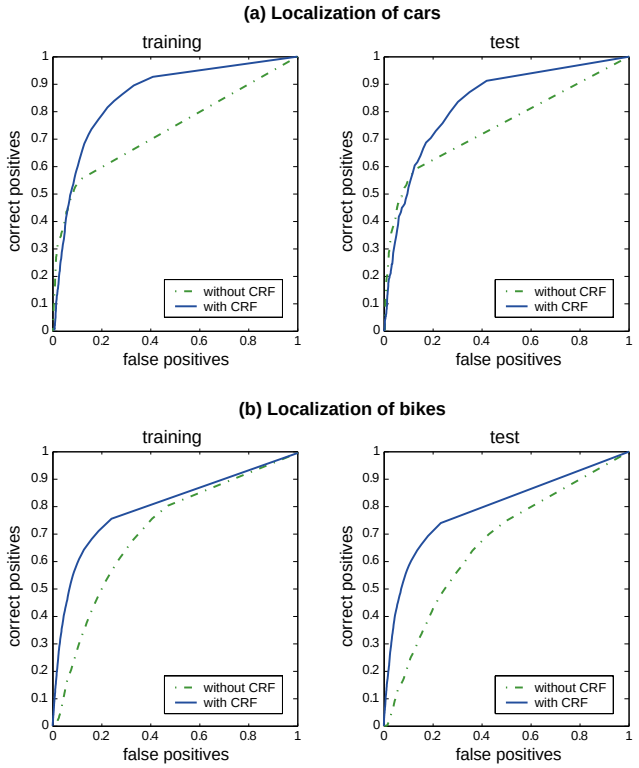

Fig. 19 ROC plots for localization of *(a)* bicycles and *(b)* cars, with (solid blue line) and without (dashed green line) the proposed CRF model. We use the Harris-Laplace detector for the cars, and the Kadir-Brady entropy detector for extracting interest regions in the bicycles database. Notice that the addition of the superpixels with the conditional random field dramatically improve the quality of the object-background separation.

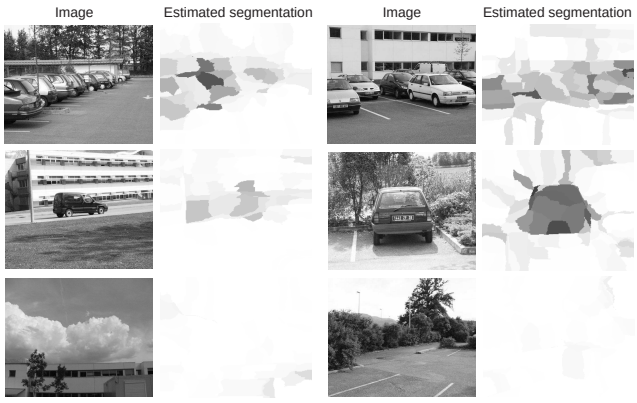

Fig. 20 Good localization results on car test images. Darker patches are more likely to correspond to cars.

a particular object class, so such results might very well extend 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 learn-

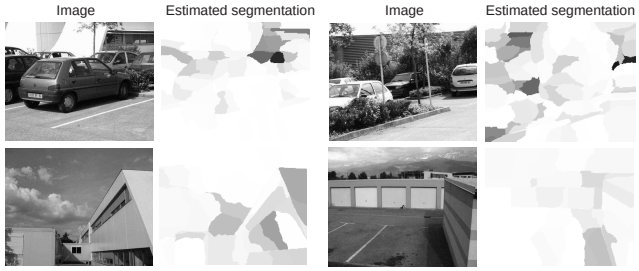

Fig. 21 Poor localization results on car test images. Darker patches are more likely to belong to the car class.

ing 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 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

We thank Guillaume Bouchard, Navneet Dalal, Daniel Eaton and Michael Marszałek and Kevin Murphy for their help, and the reviewers for their valuable critiques. We also acknowledge the financial support of the European project LAVA, the PASCAL Network of Excellence, and NSERC in Canada.

References

1. S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26:1475–1490, 2004.
2. S. Andrews, I. Tsochantaridis, and T. Hofmann. Multiple instance learning with generalized support vector machines. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pages 943–944, 2002.
3. C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan. An Introduction to MCMC for machine learning. *Machine Learning*, 50(1-2):5–43, 2003.
4. J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. John Wiley and Sons, 2000.
5. J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.
6. S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
7. P. Carbonetto, N. de Freitas, and K. Barnard. A Statistical model for general contextual object recognition. In *Proceedings of the 8th European Conference on Computer Vision*, volume I, pages 350–362, 2004.
8. 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.
9. P. Carbonetto, G. Dorko, C. Schmid, and N. de Freitas. Bayesian learning for weakly supervised object classification. Technical report, INRIA Rhône-Alpes, 2004.
10. 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.
11. S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, November 1995.
12. G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Proceedings of the ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.
13. T. Deselaers, D. Keysers, and H. Ney. Discriminative training for object recognition using images patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 157–162, 2005.
14. 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.
15. G. Dorkó and C. Schmid. Selection of scale invariant neighborhoods for object class recognition. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, volume I, pages 634–640, 2003.
16. 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 7th European Conference on Computer Vision*, volume IV, pages 97–112, 2002.
17. M. Everingham, A. Zisserman, C. Williams, and L. V. Gool. The PASCAL visual object classes challenge 2006 (VOC2006) results. Technical report, 2006.
18. 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*, II:264–271, 2003.
19. F. Hamze and N. de Freitas. From fields to trees. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 243–250, 2004.
20. T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.
21. R. Kohn, M. Smith, and D. Chan. Nonparametric regresion using linear combinations of basis functions. *Statistics and Computing*, 11:313–322, 2001.
22. H. Kück, P. Carbonetto, and N. de Freitas. A Constrained semi-supervised learning approach to data asso-

ciation. In *Proceedings of the 8th European Conference on Computer Vision*, volume III, pages 1–12, 2004.
23. H. Kück and N. de Freitas. Learning about individuals from group statistics. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, pages 332–339, 2005.
24. S. Kumar and M. Hebert. Discriminative random fields. *International Journal of Computer Vision*, 26:179–201, 2006.
25. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*, 2001.
26. S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions, 2005. IEEE Transactions on Pattern Analysis and Machine Intelligence.
27. B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume I, pages 878–885, 2005.
28. T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision*, 30(2):79–116, 1998.
29. J. S. Liu, W. H. Wong, and A. Kong. Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, 81(1):27–40, 1994.
30. J. S. Liu and Y. N. Wu. Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94(448):1264–1274, December 1999.
31. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
32. J. E. Marsden and A. J. Tromba. *Vector Calculus*. W.H. Freeman & Company, 4th edition, 1999.
33. D. McFadden. A Method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57:995–1026, 1989.
34. K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proceedings of the 8th International Conference on Computer Vision*, volume I, pages 525–531, 2001.
35. K. Mikolajczyk and C. Schmid. A Performance evaluation of local descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 257–263, 2003.
36. K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *Proceedings of the 8th European Conference on Computer Vision*, volume I, pages 69–82, 2004.
37. T. Miller, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 848–854, 2004.
38. A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proceedings of the 8th European Conference on Computer Vision*, volume II, pages 71–84, 2004.
39. H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. Identity uncertainty and citation matching. In *Advances in Neural Information Processing Systems 15*, 2003.
40. A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *Advances in Neural Information Processing Systems 17*, pages 1097–1104, 2005.
41. X. Ren and J. Malik. Learning a classification model for segmentation. In *Proceedings of the 9th IEEE International Conference on Computer Vision*, volume I, pages 10–17, 2003.
42. C. P. Robert. *The Bayesian Choice*. Springer-Verlag, 1994.
43. C. P. Robert. Simulation of truncated normal variables. *Statistics and Computing*, 5:121–125, 1995.
44. C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, 2nd edition, 2004.
45. T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume II, pages 994–1000, 2005.
46. J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 731–737, 1997.
47. J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their locations in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume I, pages 370–377, 2005.
48. S. Tham. *Markov Chain Monte Carlo for sparse Bayesian regression and classification*. PhD thesis, University of Melbourne, August 2002.
49. S. S. Tham, A. Doucet, and R. Kotagiri. Sparse Bayesian learning for regression and classification using Markov Chain Monte Carlo. In *Proceedings of the 19th International Conference on Machine Learning*, 2002.
50. M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.
51. 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 9th IEEE International Conference on Computer Vision*, volume I, pages 273–280, 2003.
52. P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.
53. J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *Proceedings of the CVPR Workshop on Learning for Adaptable Visual Systems*, 2004.
54. J. Winn and J. Shotton. The layout consistent random field for recognizing and segmenting partially occluded objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 37–44, 2006.
55. A. Zellner. *An Introduction to Bayesian Inference in Econometrics*. J. Wiley, 1971.
56. J. Zhang, M. Marsałek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: a comprehensive study. *To appear in the International Journal of Computer Vision*, 2006.

57. X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the 20th International Conference on Machine Learning*, pages 912–919, 2003.