A Constrained Semi-Supervised Learning Approach to Data Association

Hendrik Kück, Peter Carbonetto, and Nando de Freitas

Dept. of Computer Science
University of British Columbia
Vancouver, Canada
{kueck,pcarbo,nando}@cs.ubc.ca

Abstract. Data association (obtaining correspondences) is a ubiquitous problem in computer vision. It appears when matching image features across multiple images, matching image features to object recognition models and matching image features to semantic concepts. In this paper, we show how a wide class of data association tasks arising in computer vision can be interpreted as a constrained semi-supervised learning problem. This interpretation opens up room for the development of new, more efficient data association methods. In particular, it leads to the formulation of a new principled probabilistic model for constrained semi-supervised learning that accounts for uncertainty in the parameters and missing data. By adopting an ingenious data augmentation strategy, it becomes possible to develop an efficient MCMC algorithm where the high-dimensional variables in the model can be sampled efficiently and directly from their posterior distributions. We demonstrate the new model and algorithm on synthetic data and the complex problem of matching image features to words in the image captions.

1 Introduction

Data association is an ubiquitous problem in computer vision. It manifests itself when matching images (*eg* stereo and motion data [1]), matching image features to object recognition models [2] and matching image features to language descriptions [3]. The data association task is commonly mapped to an unsupervised probabilistic mixture model [4, 1, 5]. The parameters of this model are typically learned with the EM algorithm or approximate variants. This approach is fraught with difficulties. EM often gets stuck in local minima and is highly dependent on the initial values of the parameters. Markov chain Monte Carlo (MCMC) methods also perform poorly in this mixture model scenario [6]. The reason for this failure is that the number of modes in the posterior distribution of the parameters is factorial in the number of mixture components [7]. Maximisation in such a highly peaked space is a formidable task and likely to fail in high dimensions. This is unfortunate as it is becoming clear that effective learning techniques for computer vision have to manage many mixture components and high dimensions.

Here, we take a new route to solve this vision problem. We cast the data association problem as one of constrained semi-supervised learning. We argue that it is possible to construct efficient MCMC algorithms in this new setting. Efficiency here is a result of using a data augmentation method, first introduced in econometrics by economics Nobel laureate Daniel McFadden [8], which enables us to compute the distribution of the high-dimensional variables analytically. That is, instead of sampling in high-dimensions with a Markov chain, we sample directly from the posterior distribution of the high-dimensional variables. This, so called *Rao-Blackwellised*, sampler achieves an important decrease in variance as predicted by well known theorems from Markov chain theory [9].

Our approach is similar in spirit to the multiple instance learning paradigm of Dietterich *et al* [10]. This approach is expanded in [11] where the authors adopt support vector machines to deal with the supervised part of the model and integer programming constraints to handle the missing labels. This optimisation approach suffers from two problems. First, it is NP-hard so one has to introduce heuristics. Second, it is an optimisation technique and as such it only gives us a point estimate of the decision boundary. That is, it lacks a probabilistic interpretation. The approach we propose here allows us to compute all probabilities of interest and consequently we are able to obtain not only point estimates, but also confidence measures. These measures are essential when the data association mechanism is embedded in a meta decision problem, as is often the case.

The problem of semi-supervised learning has received great attention in the recent machine learning literature. In particular, very efficient kernel methods have been proposed to attack this problem [12, 13]. Our approach, still based on kernel expansions, favours sparse solutions. Moreover, it does not require supervised samples from each category and, in addition, it is probabilistic. The most important point is that our approach allows for the introduction of constraints. Adding constraints to existing algorithms for semi-supervised learning leads to NP-hard problems, typically of the integer programming type as in [11].

We introduce a coherent, fully probabilistic Bayesian model for constrained semi-supervised learning. This enables us to account for uncertainty in both the parameters and unknown labels in a principled manner. The model applies to both regression and classification, but we focus on the problem of binary classification so as to demonstrate the method in the difficult task of matching image regions to words in the image caption [3].

Our contribution is therefore threefold: a new approach to a known complex data association (correspondence) problem, a general principled probabilistic model for constrained semi-supervised learning and a sophisticated blocked MCMC algorithm to carry out the necessary computations.

2 Data association as constrained semi-supervised learning

There are many large collections of annotated images on the web, galleries and news agencies. Figure 1 shows a few annotated images from the Corel image database. By, for example, segmenting the images, we can view object recognition as the process of finding the correct associations between the labels in the caption and the image segments. Knowing the associations allows us to build a translation model that takes as input image features and outputs the appropriate words; see [3] for a detailed description. A properly trained translation model takes images (without any captions) as input and outputs images with labelled regions.

What makes this approach feasible is that the training set of images like the leftmost three images in Figure 1 is vast and ever increasing. On the other hand, a supervised approach using training data like the right-most image, where segments have been annotated, is very problematic in practice, as labelling individual segments (or other local image features) is hard and time-consuming.

This data association problem can be formulated as a mixture model similar to the ones used in statistical machine translation. This is the approach originally proposed in [3] and extended in [14] to handle continuous image features. The parameters in both cases were learned with EM. The problem with this approach is that the posterior over parameters of the mixture model has a factorial number of modes and so EM tends to get stuck in local minima. The situation is no better for MCMC algorithms for mixture models because of this factorial explosion of modes [6]. This calls for a new approach.

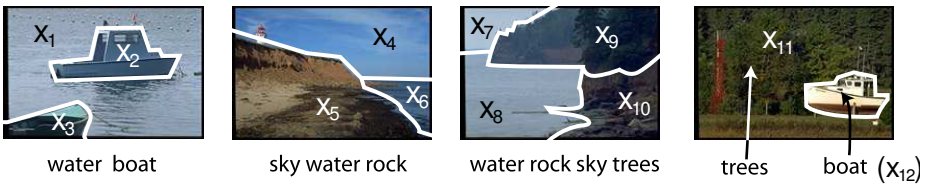


Fig. 1. Annotated images from the Corel database. We would like to automatically match image regions to words in the caption. That is we don't know the right associations (correspondences) between image features and text features.

We can convert the data association problem to a constrained semi-supervised learning problem. We demonstrate this with the toy example of Figure 1. Suppose we are interested in being able to detect boats in images. We could assume that if the word *boat* does not appear in the caption, then there are no boats in the image[1]. In this case, we assign the label 0 to each segment in the image. If however the word *boat* appears in the caption, then we know that at least one of the segments corresponds to a boat. The problem is that we do not know which. So we assign question marks to the labels in this image. Sometimes, we might be fortunate and have a few segment labels as in the rightmost image of Figure 1.

By letting x_i denote the feature vector corresponding to the i-th segment and y_i denote the existence of a boat, our data association problem is mapped to the following semi-supervised binary classification task

	image 1	image 2	image 3	image 4
Input x	x_1 x_2 x_3	x_4 x_5 x_6	x_7 x_8 x_9 x_{10}	x_{11} x_{12}
Labels y	? ? ?	0 0 0	0 0 0 0	0 1

Note that for the question marks, we have the constraint that at least one of them has to be a 1 (this is what leads to the integer programming problem in optimisation approaches). To be able to annotate all the image segments, we need to build one classifier for each word of interest. This is sound from an information retrieval point of view [11]. From an object recognition perspective, we would like to adopt multicategorical classifiers. Here, we opt for a simple solution by combining the responses of the various binary classifiers [15].

In more precise terms, given the training data $\mathcal{D}$ (a collection of images with captions) the goal is then to learn the predictive distribution $p(y = 1|\, x)$, where y is a binary indicator variable that is 1 iff the new test-set image segment represented by x is part of the concept. If we use a model with parameters θ, the Bayesian solution is given by

$$p(y = 1|\, x) = \int p(y = 1|\, x, \theta)\, p(\theta|\, \mathcal{D})\, d\theta.$$

That is, we integrate out the uncertainty of the parameters. The problem with this theoretical solution is that the integral is intractable. To overcome this problem, we sample θ according to $p(\theta|\, \mathcal{D})$ to obtain the following approximation

$$p(y = 1|\, x) \approx \frac{1}{N} \sum_i p(y = 1|\, x, \theta_i)$$

where θ_i is one of the samples. This approximation converges to the true solution by the Strong Law of Large Numbers. This approach not only allows us to compute point estimates, but also confidence intervals. In the next section, we outline the probabilistic model.

[1] Of course, this depends on how good the labels are, but as mentioned earlier, there are many databases with very good captions; see for example `www.corbis.com`. So for now we work under this assumption.

3 Parametrization and probabilistic model

Our training data $\mathcal{D}$ consists of two parts, the set of blob description vectors $\{x_{1:N}\}$ with $x_i \in \mathbb{R}^d$ for $i = 1, \ldots, N$ and a set of binary labels y^k. The full set of labels includes the known and unknown labels, $y \triangleq \{y^k, y^u\}$. Our classification model is as follows

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

where $\Phi(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-a^2/2\right) da$ is the cumulative function of the standard Normal distribution. This is the so-called probit link. By convention, researchers tend to adopt the logistic link function $\varphi(u) = (1 + \exp(-u))^{-1}$. However, from a Bayesian computational point of view, the probit link has many advantages and is equally valid. Following Tam, Doucet and Kotagiri[16], the unknown function is represented with a sparse kernel machine with kernels centered at the data points $x_{1:N}$:

$$f\left(x, \beta, \gamma\right) = \beta_0 + \sum_{i=1}^{N} \gamma_i \beta_i K\left(x, x_i\right). \tag{2}$$

Here β is a N-dimensional parameter vector and K is a kernel function. Typical choices for the kernel function K are:

- Linear: $K(x_i, x) = \|x_i - x\|$
- Cubic: $K(x_i, x) = \|x_i - x\|^3$
- Gaussian: $K(x_i, x) = \exp\left(-\lambda \|x_i - x\|^2\right)$
- Sigmoidal: $K(x_i, x) = \tanh\left(\lambda \|x_i - x\|^2\right)$

The last two kernels require a scale parameter λ to be chosen. The vector of unknown binary indicator variables $\gamma \in \{0, 1\}^N$ is used to control the complexity of the model. It leads to sparser solutions and updates, where the subset of active kernels adapts to the data. This is a well studied statistical model [16].

When all the kernels are active, we can express equation (2) in matrix notation

$$f\left(x_i, \beta\right) = \Psi_i^T \beta,$$

where Ψ_i denotes the i-th row of the kernel matrix

$$\Psi = \begin{bmatrix} 1 & K(x_1, x_1) & K(x_1, x_2) & \cdots & K(x_1, x_N) \\ 1 & K(x_2, x_1) & K(x_2, x_2) & \cdots & K(x_2, x_N) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & K(x_2, x_1) & K(x_N, x_2) & \cdots & K(x_N, x_N) \end{bmatrix} \tag{3}$$

When only a subset of kernels is active, we obtain a sparse model:

$$f\left(x_i, \beta_\gamma\right) = \Psi_{\gamma i}^T \beta_\gamma,$$

where Ψ_γ is the matrix consisting of the columns j of Ψ where $\gamma_j = 1$. $\Psi_{\gamma i}$ then is the i-th row of this matrix. β_γ is the reduced version of β, only containing

the coefficients for the activated kernels. In [16], this model is applied to supervised learning and shown to produce more accurate results than support vector machines and other kernel machines. Here, we need to extend the model to the more general scenario of semi-supervised learning with constraints in the labels.

We adopt a hierarchical Bayesian model [17]. We assume that each kernel is active with probability τ, i.e. $p(\gamma|\,\tau)$ is a Bernoulli distribution. Instead of having the user choose a fixed τ a priori, we deal with this parameter in the Bayesian way and assign a prior $p(\tau)$ to it. This way, the value of τ is allowed to adapt to the data. At the same time we can bias it by specifying the prior $p(\tau)$ according to our prior belief as to what the value of τ should be. While Tam, Doucet and Kotagiri [16] use the completely uninformative uniform prior, we instead choose to put a conjugate Beta-prior on τ which allows the user to exert as much control as desired over the percentage of active kernels

$$p(\tau) = \frac{\Gamma(a+b)}{\Gamma(a)\,\Gamma(b)}\tau^{a-1}(1-\tau)^{b-1}. \tag{4}$$

For the choice $a = b = 1.0$, we get the uninformative uniform distribution. We obtain the prior on the binary vector γ by integrating over τ

$$p(\gamma) = \int p(\gamma|\,\tau)\,p(\tau)\,d\tau = \frac{\Gamma(\Sigma\gamma + a)\,\Gamma(N - \Sigma\gamma + b)}{\Gamma(N + a + b)}, \tag{5}$$

where $\Sigma\gamma$ is the number of active kernels, i.e. the number of non zero elements in γ.

A (maximum entropy) g-prior is placed on the coefficients β:

$$p(\beta) = \mathcal{N}(0, \delta^2(\Psi_\gamma^T\Psi_\gamma)^{-1}) \tag{6}$$

where the regularisation parameter is assigned an inverse gamma prior:

$$p\left(\delta^2\right) = \mathcal{IG}(\frac{\mu}{2}, \frac{\nu}{2}). \tag{7}$$

This prior has two parameters μ and ν that have to be specified by the user. One could argue that this is worse than the single parameter δ^2. However, the parameters of this hyper-prior have a much less direct influence than δ^2 itself and are therefore less critical for the performance of the algorithms [17]. Assigning small values to these parameters results in an uninformative prior and allows δ^2 to adapt to the data.

3.1 Augmented Model

We augment the probabilistic model artificially in order to obtain an analytical expression for the posterior of the high-dimensional variables β. In particular, we introduce the set of independent variables $z_i \in \mathbb{R}$, such that

$$z_i = f(x_i, \beta, \gamma) + n_t, \tag{8}$$

where $n_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. The set of augmentation variables consists of two subsets $z \triangleq \{z^k, z^u\}$, one corresponding to the known labels y^k and the other to the unknown labels y^u. For the labelled data, we have

$$p\left(z_i^k \middle| \beta, \gamma, x_i\right) = \mathcal{N}\left(f\left(x_i, \beta, \gamma\right), 1\right) = \mathcal{N}\left(\Psi_{\gamma i}^T \beta_\gamma, 1\right). \tag{9}$$

We furthermore define

$$y_i^k = \begin{cases} 1 \text{ if } z_i^k > 0, \\ 0 \text{ otherwise.} \end{cases}$$

It is then easy to check that one has the required result:

$$\Pr\left(y_i^k = 1 \middle| x_i, \beta_\gamma, \gamma\right) = \Pr\left(z_i^k \geq 0 \middle| x_i, \beta_\gamma\right) = \Pr\left(n_i \geq -\Psi_{\gamma i}^T \beta_\gamma\right) = \Phi\left(\Psi_{\gamma i}^T \beta_\gamma\right).$$

Now, let $y_{k:k+l}^u$ denote the set of missing labels for a particular image (a set of question marks as described in Section 2). The prior distribution for the corresponding augmentation variables $z_{k:k+l}^u$ is then:

$$p(z_{k:k+l}^u | \beta, \gamma, x_i) \propto \left[\prod_{j=k}^{j=k+l} \mathcal{N}\left(\Psi_{\gamma j}^T \beta_\gamma, 1\right)\right] \mathbb{I}_{\mathcal{C}}(z_{k:k+l}^u) \tag{10}$$

where $\mathbb{I}_\Omega(\omega)$ is the set indicator function: 1 if $\omega \in \Omega$ and 0 otherwise. Our particular set of constraints is $\mathcal{C} \triangleq \{\text{one or more } z_j^u > 0\}$. That is, one or more of the z_j^u must be positive so that at least one of the y^u are positive. This prior is a truncated Normal distribution with the negative octant missing. The hierarchical Bayesian model is summarised in Figure 2.

3.2 Posterior distribution

The posterior distribution follows from Bayes rule

$$p(\beta, \gamma, \delta^2, z | y^k, x_{1:N}) \propto p(y^k|z^k) p(\gamma) p(\beta|\delta^2) p(\delta^2) p(z^u|\beta, \gamma, x) p(z^k|\beta, \gamma, x)$$

The key thing to note, by looking at our graphical model, is that by conditioning on the 1-dimensional variables z, the model reduces to a standard linear-Gaussian model [17]. We can as a result obtain analytical expressions for the

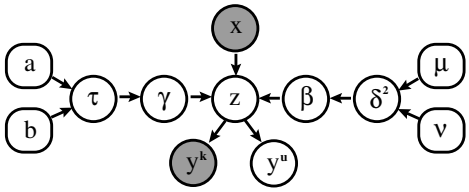


Fig. 2. Our directed acyclic graphical model. Note that by conditioning on z, y is independent of the model parameters.

conditional posteriors of the high-dimensional variables β and the regularisation parameter δ

$$p(\beta|z,x,\gamma,\delta^2) = \mathcal{N}\left(\frac{\delta^2}{1+\delta^2}\left(\Psi_\gamma^T\Psi_\gamma\right)^{-1}\Psi_\gamma^T z, \frac{\delta^2}{1+\delta^2}\left(\Psi_\gamma^T\Psi_\gamma\right)^{-1}\right) \tag{11}$$

$$p(\delta^2 \mid z,\beta,\gamma) = \mathcal{IG}\left(\frac{\mu+\Sigma_\gamma+1}{2}, \frac{\nu+\beta^T\Psi_\gamma^T\Psi_\gamma\beta}{2}\right) \tag{12}$$

where z is the vector $(z_1, z_2, \ldots, z_N)^T$. The posterior distribution of the augmentation variables z^k is given by the following truncated Normal distributions:

$$p(z_i^k|\beta,\gamma,x_i,y_i^k) \propto p(y_i^k|z_i^k)p(z_i^k|x_i,\beta,\gamma) = \begin{cases} \mathcal{N}\left(\Psi_{\gamma i}^{\mathtt{T}}\beta,1\right)\mathbb{I}_{(0,+\infty)}\left(z_i^k\right) \text{ if } y_i^k = 1 \\ \mathcal{N}\left(\Psi_{\gamma i}^{\mathtt{T}}\beta,1\right)\mathbb{I}_{(-\infty,0]}\left(z_i^k\right) \text{ if } y_i^k = 0 \end{cases} \tag{13}$$

4 MCMC Computation

We need to sample from the posterior distribution $p\left(\theta|\,\mathcal{D}\right)$, where θ represents the full set of parameters. To accomplish this, we introduce a Metropolised blocked Gibbs sampler. In short, we sample the high-dimensional parameters β and the regularisation parameters directly from their posterior distributions (equations (11) and (12)). It is important to note that only the components of β associated with the active kernels need to be updated. This computation is therefore very efficient. The γ are sampled with the efficient MCMC algorithm described in detail in [16]. The z^u are sampled from the truncated multivariate Gaussian in equation (10), while the z^k are sampled from the truncated distributions given by equation (13).

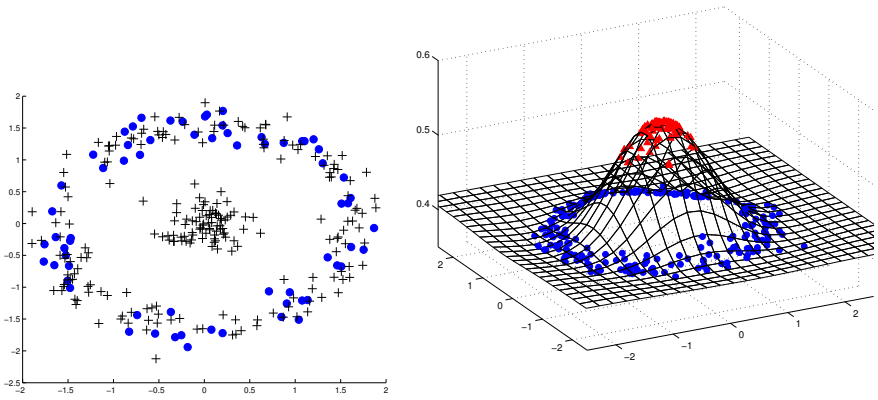
To sample from the truncated Gaussian distributions, we use the specialised routines described in [18]. These routines based on results from large deviation theory are essential in order to achieve good acceptance rates. We found in our experiments that the acceptance rate was satisfactory (70% to 80%).

5 Experiments

5.1 Synthetic data

In this first experiment we tested the performance of our algorithm on synthetic data. We sampled 300 data points from a mixture model consisting of a Gaussian and a surrounding ring with Gaussian cross section (see Figure 3(a)). Data points generated by the inner Gaussian were taken to be the positive instances, while those on the ring were assumed to be negative. The data points were then randomly grouped into groups (representing documents) of 6 data points each. In the given example, this resulted in 12 groups with exclusively negative data points, and 38 groups with both positives and negative instances. This corresponds to 72 data points with known negative labels and 228 data points with unknown but constrained labels.

We ran our algorithm on this data for 2000 samples (after a burn-in period of 1000 samples) using uninformative priors and a sigmoidal kernel with kernel parameter $\lambda = 1.0$. Although no data points were explicitly known to be positive in this case, the information of the constraints was sufficient to learn a nice distribution $p(y = 1|x)$ as shown in Figure 3(b). Using an appropriate threshold produces a perfect classification in this example.



(a) The given data points (b) Surface plot of $p(y = 1|x)$

Fig. 3. Experiment with synthetic data. (a) shows the generated data points. Instances with known negative labels are shown as filled circles whereas data points with unknown label are represented by the + symbol. The plot in (b) visualizes the probability distribution computed by our approach. The distribution obviously nicely separates positive and negative examples and thus provides an excellent classifier.

5.2 Object recognition data set

For this experiment, we used a set of 300 annotated images from the Corel database. The images in this set were annotated with in total 38 different words and each image was segmented into regions using normalised cuts [19]. Each of the regions is described by a 6-dimensional feature vector (CIE-Lab colour, y position in the image, boundary to area ratio and standard deviation of brightness). The data set was split into one training set containing 200 images with 2070 image regions and a test set of 100 images with 998 regions.

We compared two learning methods in this experiment. The first consisted of a mixture of Gaussians translation model trained with EM [3, 14]. The second is the method proposed in this paper. We adopted a vague hyper-prior for δ^2 ($\mu = \nu = 0.01$). Experiments with different types of kernels showed the sigmoidal kernel to work best for this data set. Not only did it produce better classifiers

than linear, multi-quadratic, cubic and Gaussian kernels, it also led to numerically more stable and sparser samplers. The average number of activated kernels per sample was between 5 and 20, depending on the learned concept.

We used both EM with the mixture model and our new constrained semi-supervised approach to learn binary classifiers for several of the words in this dataset. The Markov chains were run for 10,000 samples after a burn-in phase of 10,000 samples. On a 2.6 Ghz Pentium 4, run times for this were in the range of 5 to 10 minutes, which is perfectly acceptable in our setting.

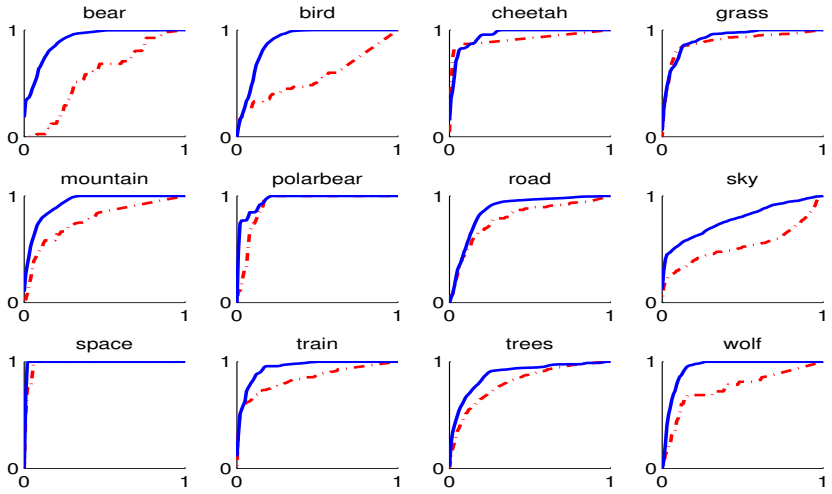
The performance of the learned classifiers was then evaluated by comparing their classification results for varying thresholds against a manual annotation of the individual image regions. The ROC plots in Figure 4 show the results averaged over 20 runs, plotting true positives against false positives. The plots show that the approach proposed in this paper yields significantly better classification performance than the EM mixture method. Given the relative simple features used and the small size of the data set, the performance is remarkably good. Figure 4(b) shows that the classifiers learned using the proposed approach generalize fairly well even where the EM mixture approach fails due to overfitting (look at the results for the concept 'space' for an example).

Figure 5 illustrates the dramatically higher consistency across runs of the algorithm proposed in this paper as compared to the EM algorithm for the mixture model. The error bars indicate the standard deviation of the ROC plots across the 20 runs. The large amount of variation indicates that the EM got stuck in local minima on several runs. While with the Corel data set this problem arose only for some of the categories, in larger and higher dimensional data sets, local minima are known to become a huge problem for EM.

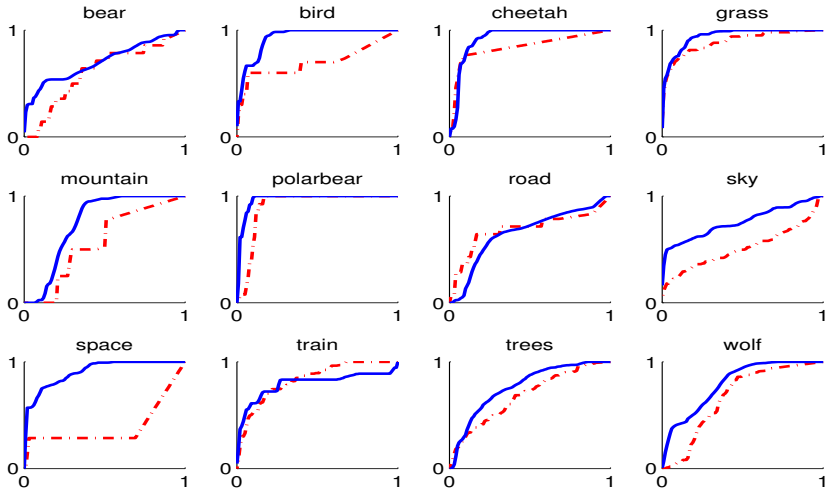
Finally, in Figure 6 we show some examples of the classifications generated by the two algorithms for 3 different images containing polar bears. In order to get a fair comparison of the 'polarbear' classifiers learned by the two approaches, we chose the thresholds so that both classifiers would produce 30 positive instances on the full test-set (which contains 27 patches manually labelled as 'polarbear'). The MCMC based approach presented in this paper manages to correctly classify the polar bears in the first two images while the mixture model trained with EM fails to do so. In the third example, both classifiers mistake the ice for the bear. This demonstrates a general problem in such data association tasks. If two concepts (like in this case polar bears and ice) appear together in all example documents, it can be very hard or even impossible to disambiguate them. A manually labelled example (such as in the rightmost image in Figure 1) could be of great value in such situations. The approach proposed in this paper can naturally handle such explicitly given associations (although none were used in the experiments presented here).

Acknowledgements

We would like to thank Arnaud Doucet, Kerry Mengersen and Herbie Lee. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Institute for Robotics and Intelligent Systems (IRIS) under the project title 'Robot Partners: Visually Guided Multi-agent Systems'.

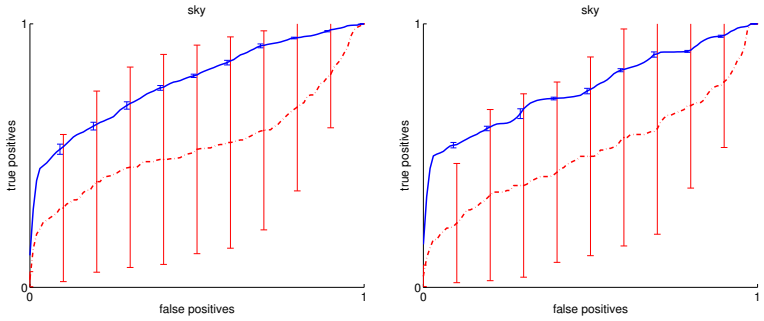


(a) training data



(b) test data

Fig. 4. ROC plots measuring the classification performance on image regions from the Corel image dataset of both the proposed algorithm (solid line) and the EM mixture algorithm (dashed line), averaged over 20 runs. The x axis measures $\frac{\text{negatives falsely classified as positives}}{\text{actual negatives}}$ while the y axis corresponds to $\frac{\text{correctly classified positives}}{\text{actual positives}}$. The plots are generated by using the learned probabilistic classifiers with varying thresholds and allow to compare the classifiers independent of a chosen fixed threshold value. The performance on the test set is remarkable considering that the algorithm only has access to simple image features (and no text in any form).



(a) performance on training data

(b) performance on test data

Fig. 5. ROC plots (as in Figure 4) for the annotation 'sky' . The average performance of our proposed approach is visualized by the solid line, that of the EM mixture algorithm by the dashed line. The error bars represent the standard deviation across 20 runs. It is clear from the plots that our proposed algorithm is more reliable and stable.

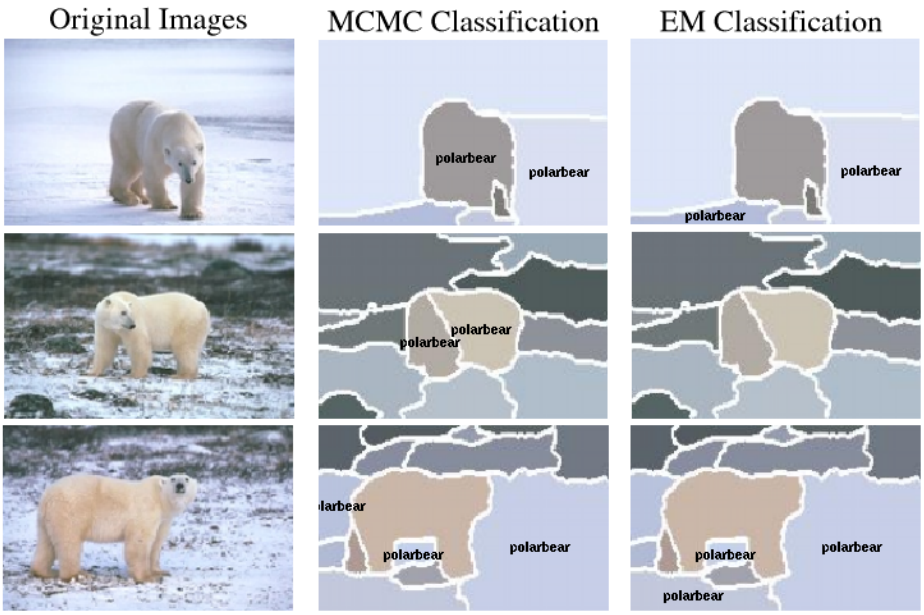


Fig. 6. Polar bear recognition results on three example images obtained with our constrained semi-supervised approach trained with MCMC and the mixture model trained with EM. The approach presented in this paper yields better results than the EM mixture model. As polar bears and ice appear together in all training images, it can be very hard to differentiate between them, as the last example illustrates.

References

1. Dellaert, F., Seitz, S., Thorpe, C., Thrun, S.: EM, MCMC, and chain flipping for structure from motion with unknown correspondence. Machine Learning **50** (2003) 45–71
2. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: IEEE Conference on Computer Vision and Pattern Recognition. (2003)
3. Duygulu, P., Barnard, K., de Freitas, N., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: ECCV. (2002) 97–112
4. Avitzour, D.: A maximum likelihood approach to data association. IEEE Transactions on Aerospace and Electronic Systems **28** (1992) 560–566
5. Blei, D., Jordan, M.: Modeling annotated data. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press (2003) 127–134
6. Celeux, G., Hurn, M., Robert, C.P.: Computational and inferential difficulties with mixture posterior distributions. Journal of the American Statistical Association **95** (2000) 957–970
7. Stephens, M.: Bayesian Methods for Mixtures of Normal Distributions. PhD thesis, Department of Statistics, Oxford University, England (1997)
8. McFadden, D.: A method of simulated moments for estimation of discrete response models without numerical integration. Econometrica **57** (1989) 995–1026
9. Liu, J., Wong, W.H., Kong, A.: Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. Biometrika **81** (1994) 27–40
10. Dietterich, T.G., Lathrop, R.H., Lozano-Perez, T.: Solving the multiple instance learning with axis-parallel rectangles. Artificial Intelligence **89** (1997) 31–71
11. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems 16, Cambridge, MA, MIT Press (2003)
12. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using Gaussian fields and harmonic functions. In: International Conference on Machine Learning. (2003)
13. Belkin, M., Niyogi, P.: Semi-supervised learning on manifolds. Technical Report TR-2002-12, Computer Science Department, The University of Chicago, MA (1994)
14. Carbonetto, P., de Freitas, N., Gustafson, P., Thompson, N.: Bayesian feature weighting for unsupervised learning, with application to object recognition. In: AI-STATS, Florida, USA (2003)
15. Tipping, M.E.: Sparse Bayesian learning and the relevance vector machine. Journal of Machine Learning Research **1** (2001) 211–244
16. Tham, S.S., Doucet, A., Ramamohanarao, K.: Sparse Bayesian learning for regression and classification using Markov chain Monte Carlo. In: International Conference on Machine Learning. (2002) 634–641
17. Bernardo, J.M., Smith, A.F.M.: Bayesian Theory. Wiley Series in Applied Probability and Statistics (1994)
18. Geweke, J.: Efficient simulation from the multivariate normal and Student t-distributions subject to linear constraints. In: Proceedings of 23rd Symp. Interface. (1991) 571–577
19. Shi, J., Malik, J.: Normalized cuts and image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. (1997) 731–737