Region Classification with Markov Field Aspect Models

Jakob Verbeek, Bill Triggs
LJK, INRIA, 655 avenue de l'Europe, Montbonnot 38330, France
Jakob.Verbeek@inria.fr, Bill.Triggs@imag.fr

Abstract

Considerable advances have been made in learning to recognize and localize visual object classes. Simple bag-of-feature approaches label each pixel or patch independently. More advanced models attempt to improve the coherence of the labellings by introducing some form of inter-patch coupling: traditional spatial models such as MRF's provide crisper local labellings by exploiting neighbourhood-level couplings, while aspect models such as PLSA and LDA use global relevance estimates (global mixing proportions for the classes appearing in the image) to shape the local choices. We point out that the two approaches are complementary, combining them to produce aspect-based spatial field models that outperform both approaches. We study two spatial models: one based on averaging over forests of minimal spanning trees linking neighboring image regions, the other on an efficient chain-based Expectation Propagation method for regular 8-neighbor Markov Random Fields. The models can be trained using either patch-level labels or image-level keywords. As input features they use factored observation models combining texture, color and position cues. Experimental results on the MSR Cambridge data sets show that combining spatial and aspect models significantly improves the region-level classification accuracy. In fact our models trained with image-level labels outperform PLSA trained with pixel-level ones.

1. Introduction

An ongoing theme in computer vision is the relationship between low-level image features and scene or object classes such as sky, water, cars, bicycles. Models that relate features to classes can be used to solve various tasks including *classification* (determining whether the image contains at least one instance of the class), *detection* (determining the positions of the class instances in the image), and *segmentation* (partitioning the image into regions covered by the classes present). When learning such models from annotated images, labeling can be done at various levels of specificity: image-level keywords only indicate that certain classes are present in the image; bounding boxes provide rectangular image regions in which the classes occur; and full segmentation masks assign individual pixels to classes.

Recently there has been considerable interest in methods for segmenting images into semantic classes [9, 11, 21]. Impressive results have been achieved with conditional random field [12] based models trained using detailed pixel-level labellings. Creating such labellings requires substantially more effort than providing image-level keywords and it would be useful to develop models that allow learning from keywords alone along the lines of [2, 7].

At the other extreme, some authors [20, 22] argue that aspect models like Probabilistic Latent Semantic Analysis (PLSA) [10] and Latent Dirichlet Allocation (LDA) [5] can recover semantic-level visual models even under completely unsupervised training (without any labeled training data). However others [13] find that unsupervised aspect models perform poorly in classification tasks, even in relatively simple cases where each image contains just one easily visible target class and inter-class variation is limited (*e.g.* the only views of cars are side views). It does seem clear that when only a few labelled training images are available, performance can be improved by representing them in terms of their aspect model topic mixtures rather than in terms of their original features [1, 13, 19].

In this paper we combine the advantages of spatial fields and aspect models, focusing particularly on the setting where the model is learned from image-level keywords without detailed pixel-level labelling. The key challenge is to associate image regions with the correct labels so as to estimate crisply defined class models. Our work is related to [18, 14] where segmentation models for textured animals are learned from images grouped by the class of animal. It differs in that: (i) there may be many classes in each image; (ii) we use dense image features to handle

This work was supported by an INRIA postdoctoral fellowship, the European Union funded research project CLASS, and the PASCAL network. We would like to thank J. van de Weijer for his colour descriptor code.

1-4244-1180-7/07/$25.00 ©2007 IEEE

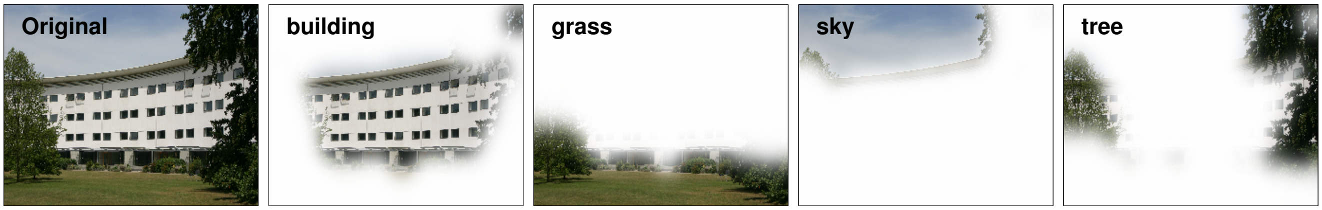

Figure 1. A training image labeled $\{$*building*, *grass*, *sky*, *tree*$\}$, and the corresponding soft region labelings inferred during learning.

classes with little texture; and most importantly (*iii*) we use well defined probabilistic models that capture both spatial coherence (local correlations between labels) and thematic coherence (image-wide correlations).

From each image we extract overlapping patches on a grid, representing them by color and texture descriptors and rough indications of their image positions. Label induction takes place at the patch level, but the results are propagated to pixel level for visualization and performance quantification. Figure 1 shows an example of a training image labeled with keywords together with the inferred association of image regions to labels. We assume that each image patch belongs either to one of the label classes or to a vague background class '*void*' that is available in every image.

Aspect models such as PLSA and LDA are probabilistic models that are well suited to this situation. They model (the patches of) each image as a mixture of latent aspects or 'topics'. Each image has its own mixing proportions but the topics and their attributes are shared by all images. Here we allocate a single latent topic for each semantic class. Extending this to multiple topics per class is trivial (but useful). To incorporate labelled training images we use only the topics corresponding to the the image's labels and *void* to model the image.

Existing aspect models do not enforce spatial coherence: they effectively assume that the labels of adjacent patches are independent, thus ignoring the strong local correlations that are found in real images. We develop two extensions of PLSA that are designed to capture some of this coherence: forests of random spanning trees and regular Markov random fields. Our experimental results show that they significantly outperform PLSA. In fact, even when trained using only image-level labels, they provide region-level classification accuracies similar to or better than those of PLSA trained using detailed pixel-level labellings.

The rest of the paper is organized as follows. Section 2 briefly introduces aspect models and describes how we use them to learn from labeled images. Section 3 sketches our image features and how we incorporate them. Section 4 is devoted to spatial extensions of aspect models. Section 5 gives comparative experimental results, and Section 6 concludes and sketches ongoing work.

2. Aspect models

Aspect models such as PLSA [10] and LDA [5] have been extensively studied as models of collections of text documents. Each document d is modeled generatively as a bag of words sampled from a document-specific mixture of T latent 'aspect' or 'topic' distributions. Each topic t is characterized by its distribution $p(w|t)$ over the W words of the dictionary, and each document d is characterized by its vector $\boldsymbol{\theta}_d$ of mixing weights over topics. Given $\boldsymbol{\theta}_d$, the N_d words of d are modelled as independent samples from the mixture. Letting z_n denote the unknown topic (mixture component) of word w_n and $\theta_{dt} = p(z_n{=}t|\boldsymbol{\theta}_d)$ denote the mixing weight of topic t in d, the probability of the document becomes:

$$p(w_1, \ldots, w_{N_d} \,|\, \boldsymbol{\theta}_d) \;=\; \prod_{n=1}^{N_d} \sum_{t=1}^{T} \theta_{dt}\, p(w_n \,|\, t). \qquad (1)$$

When aspect models are applied to image collections, the images take on the role of documents and the image patches that of words. Images are thus modeled as mixtures of latent aspects that generate appearance descriptors independently for each patch.

PLSA can be viewed as a probabilistic generalization of PCA – a low-rank nonnegative approximation of the matrix of (empirical word probabilities for each document) × (documents) using factor matrices $\{p(w|t)\} \times \{p(t|d)\}$ obtained by Expectation-Maximization (EM) [3]. There are various extensions. Notably, LDA adds a sparse (Dirichlet) prior for the topic weights $\boldsymbol{\theta}_d$ and treats these as hidden variables to be integrated out rather than as parameters to be estimated using Maximum Likelihood for each document. Here we use only PLSA. It is computationally more efficient than LDA and it has comparable accuracy in practice. The main benefit of aspect models is the low-rank projection of the document onto the aspect space, which regularizes the per-document word-probabilities $p(w|d)$ and helps to capture some of the underlying semantics via the per-image mixing coefficients $p(t|d)$. LDA provides additional regularization by encouraging the topic mixtures to be sparse and by averaging over their weights, but this only makes a significant difference for small documents and many topics, $N_d \not\gg T$.

We learn topic vectors $p(w|t)$ from image-level labels simply by setting θ_{dt} to zero for all classes (except *void*) that are not listed among the image labels. So only the images that are labeled with a topic contribute to learning its topic vector. The remaining θ_{dt} can have any (non-negative sum-to-one) values[1]. Section 5 shows that even such weak supervision allows good topic models to be learned.

PLSA learning is subject to local minima. To avoid some of the problems associated with these we initially hold the per-image topic proportions $\boldsymbol{\theta}_d$ fixed to uniform distributions over the given image-level labels (and zero for those not present). Once the topic vectors $p(w|t)$ have stabilized we allow both $\boldsymbol{\theta}_d$ and $p(w|t)$ to vary to obtain the final model. This makes the learned topic vectors significantly cleaner. Secondly, we force the topic vector for the *void* class to remain generic by fixing it to the uniform distribution rather than learning it.

3. A multi-modal aspect model to combine cues

We use cues from $M = 3$ distinct modalities to characterize each image patch: a SIFT descriptor [15] captures the local texture; the robust hue descriptor of [23] describes the local color distribution; and approximate image location is coded by the index of the cell into which the patch falls, where the image is covered by a $c \times c$ grid of regular cells. Our experiments used $c = 5$ and $c = 10$ – larger values reduce the performance. The SIFT and color descriptors are vector quantized into respectively 1000 and 100 bins using centers learned using k-means from all of the training set descriptors. The compound patch descriptor thus has $c^2 \cdot 10^3 \cdot 10^2$ possible values. It is infeasible to learn an aspect model over a vocabulary of this size. Instead we assume that the modalities are independent given the aspect and learn factored models for them, thus giving a generative model of the form:

$$p(\{w_n\}_{n=1}^{N_d} \,|\, \boldsymbol{\theta}_d) \;=\; \prod_{n=1}^{N_d} \sum_{t=1}^{T} \theta_{dt} \prod_{m=1}^{M} p(w_n^m \,|\, t)\,. \quad (2)$$

Note that we only assume independence *given* the aspect. For example, knowing that a region is green or blue can only alter the conditional distribution of its texture (*e.g.* favoring vertical grass-like textures or smooth sky-like ones) via the latent topics involved (*grass*, *sky*, . . .).

In contrast, [6] combines color and texture by calculating 128-D SIFT independently on each component of HSV color space and vector quantizing the resulting 384-D feature vectors. This captures within-topic interactions but the dimension of the resulting feature spaces rapidly becomes prohibitive. Unlike multi-modal LDA [4], in our model the three modalities of each patch share a single common topic.

4. Spatial extensions to aspect models

Aspect models such as PLSA and LDA ignore the spatial structure of the image, modeling its patches as independent draws from the topic mixture $\boldsymbol{\theta}_d$. We now discuss two models that capture more of the spatial coherence of natural region labels. Random fields are an attractive way of doing this but exact inference tends to be intractable for ones that are densely connected enough to capture spatial interactions well. Our first model uses tree-structured approximations of the conditional dependencies between region labels, thus allowing efficient exact inference with standard Belief Propagation. Our second model uses a conventional 8-neighbor Markov Random Field with efficient approximate inference based on Expectation Propagation (EP) on intersecting Markov chains.

4.1. Forests of spanning trees

A simple way to capture some of the local dependencies between aspect labels in a computationally tractable form is to connect the nodes of the image $Z = \{z_1, \ldots, z_{N_d}\}$ in a tree (*e.g.* a spanning subtree of the usual MRF grid) and impose a tree-structured prior

$$p(Z) \;\propto\; \exp\left(\sum_i \psi(z_i, z_{\pi(i)}) + \log\theta_{dz_i}\right), \quad (3)$$

where $\pi(i)$ denotes the unique parent of node i in the tree. There are a great many possible trees, even if we restrict attention to ones whose edges link near neighbors. Neighbors with similar appearance often belong to the same class so they are preferred candidates for coupling. To sample the trees, we initially connect all neighboring patches, assigning edges the value of 1 if they have the same SIFT or color index, and 0 otherwise. We then randomly sample maximal spanning trees from this 0-1 valued neighborhood graph. There are usually many such trees, so to reduce the dependence on arbitrary choices we randomly select 10 and average (marginal label probabilities) over them. Experimentally we find that such averaging helps but that the effect saturates after about 10 trees. There are also various ways to set the edge weights more progressively based on appearance, but we will not discuss these here.

Perhaps surprisingly, Potts-like pair-wise potentials

$$\psi(z_i, z_{\pi(i)}) \;=\; \rho \cdot [z_i = z_{\pi(i)}], \quad (4)$$

(where $[\cdot]$ denotes the indicator function of its argument) give more accurate classification results than ones estimated using fully supervised pixel-level labellings (again with averaging over 10 trees). We therefore use the Potts form below, determining ρ by cross-validation.

[1] Here we use only simple constraints that set the mixing proportions of topics not present to zero, but other convex constraints such as upper or lower bounds on the proportions or their relative sizes could easily be included. The negative data log-likelihood is a convex function of the proportions, so for any convex constraints on them, the optimal mixture can be found efficiently using convex optimization.

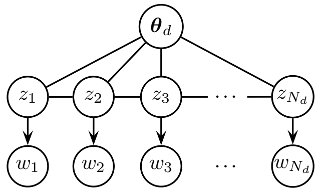

Figure 2. A graphical representation of our Markov field aspect model. (In reality, the z_n are coupled in a 2D lattice).

4.2. Markov field aspect model

The above tree-structured models can include only $N_d{-}1$ direct connections among N_d nodes, so spatial neighbors often end up being well separated in the trees. This typically limits the extent to which their correlations can be captured. A more realistic model is a Markov Random Field (MRF) with connections between all spatially adjacent nodes. The MRF has the following prior over node labels [3]:

$$p(Z) \;\propto\; \exp\left(\sum_i \log\theta_{dz_i} + \sum_{i\sim j}\phi(z_i,z_j)\right), \qquad (5)$$

where $i{\sim}j$ enumerates spatial neighbor nodes i, j. The pairwise potentials $\phi(z_i, z_j)$ encode the compatibility between neighboring labels. As before we use a Potts model

$$\phi(z_i,z_j) \;=\; \sigma\cdot[z_i = z_j]. \qquad (6)$$

See Figure 2 for a graphical representation.

Exact inference in MRFs – in particular, determining the posterior marginals $p(z_i|\{w_n\}_{n=1}^{N_d})$ that are needed to learn the topic models – requires time exponential in the field width (typically $O(\sqrt{N_d})$). However there are various methods for approximating marginals, including (structured) Gibbs sampling [8], Variational Mean-Field (VMF), Loopy Belief Propagation (LBP), and Expectation Propagation (EP). For an introduction to the latter three methods see [3], and for more details on EP and its relation to loopy BP, variational and other approximations see [16]. Gibbs sampling can in principle provide samples from the true posterior marginals, but it typically requires exponentially many iterations to do so. Variational methods and loopy BP converge more quickly, but tend to provide overconfident approximations: the variability and entropy of the marginals are under-estimated. EP is a recent approximation technique for exponential family models based on matching statistical moments. When the node marginals are multi-modal, EP tends to smooth over the peaks and hence over-estimate the marginal entropies, but in practice its marginals are often found to be more accurate than those of VMF and LBP [3]. In particular, MRF priors that can generate segmentation-like behaviour typically have many well-separated modes associated with collective local relabellings. For these, (sequential) Gibbs sampling is exponentially slow and VMF and LBP tend to lock onto just one mode, while EP tends to average over multiple modes.

EP approximates the MRF posterior $p(Z|\{w_n\}_{n=1}^{N_d})$ with a simpler model distribution $Q(Z)$ that is estimated from the observations and some tractable factorization of the MRF prior. Here we have a regular grid of patches with pairwise 8-neighbor connections. We use a completely factored approximation $Q(Z) = \prod_i q_i(z_i)$ and write the MRF prior as a product of four factors, respectively covering the contributions of horizontal, vertical, and the two types of diagonal edges:

$$p(Z) \;\propto\; t_1(Z)\,t_2(Z)\,t_3(Z)\,t_4(Z). \qquad (7)$$

Each factor is a set of 1-D MRFs whose exact marginals are calculated using the Baum-Welch algorithm [3]. This allows the EP projection step to account for the averaged influence of the 1-D couplings[2].

As with standard PLSA, we treat $(\boldsymbol{\theta}_d, \sigma_d)$ as parameters to be estimated for each image. Our full algorithm interleaves applications of EP to find marginals for the patch labels with M-steps that minimize a convex cost to update $(\boldsymbol{\theta}_d, \sigma_d)$ – and $p(w|t)$ too, if we are learning topic vectors. However we find that we can achieve essentially identical performance by using simple PLSA to estimate $\boldsymbol{\theta}_d$ and fixing $\sigma_d = \sigma$ empirically for all images. This is much faster than the full algorithm so we have used it in most of the below experiments.

5. Experimental results

Our experiments use the Microsoft Research Cambridge (MSRC) data sets[3]. The first set contains 240, 213×320 pixel images, each with a ground truth segmentation that labels each pixel with one of 13 semantic classes or *void*. The *void* pixels either do not belong to one of the 13 classes or lie near boundaries between classes and were labeled as *void* to simplify the task of manual segmentation. Here we also treat the classes *horse*, *mountain*, *sheep* and *water* as *void* as they occur rarely in the data set.

The algorithms below are based on 20×20 pixel patches extracted at 10 pixel intervals at a single scale. Each patch thus overlaps its eight nearest neighbors. Its ground truth label is taken to be the most frequent pixel label within it. On output, pixel-level posterior label probabilities are obtained by either: (i) using the posterior of the nearest patch centre – thus producing probabilities that are constant over 10×10 pixel blocks as in Figs. 3 and 4; or (ii) linearly interpolating the 4 adjacent patch-level posteriors to produce smooth

[2] Our formulation can also be interpreted as a particular message passing scheme in Loopy Belief Propagation – see [24].

[3] http://research.microsoft.com/vision/cambridge/recognition

	Building	Grass	Tree	Cow	Sky	Aeroplane	Face	Car	Bicycle	Average
S	51.1 (12.1)	74.0 (10.8)	68.1 (15.8)	59.0 (15.3)	59.2 (6.4)	52.1 (16.5)	52.5 (12.9)	59.4 (14.9)	76.3 (5.8)	61.3 (3.1)
C	50.4 (13.8)	77.6 (12.8)	46.8 (24.6)	51.0 (17.4)	81.2 (10.5)	20.8 (13.9)	77.2 (13.1)	58.5 (15.3)	38.0 (17.3)	55.7 (5.3)
P	0.0 (0.0)	86.6 (5.5)	0.0 (0.0)	0.0 (0.0)	68.9 (6.0)	3.5 (6.3)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	17.7 (1.0)
SC	66.6 (10.3)	84.0 (8.8)	59.5 (18.9)	74.8 (16.6)	89.4 (3.5)	74.8 (9.0)	80.7 (8.3)	73.9 (9.3)	73.0 (8.1)	75.2 (3.4)
SP	58.0 (8.9)	76.1 (7.3)	**62.6** (19.8)	74.0 (11.1)	81.0 (4.3)	69.4 (14.4)	55.9 (12.6)	69.3 (11.9)	**76.7** (5.1)	69.2 (3.6)
CP	60.5 (13.5)	80.2 (12.3)	38.5 (22.4)	57.2 (20.7)	89.5 (6.1)	48.4 (13.3)	76.6 (10.9)	63.6 (14.1)	34.9 (13.0)	61.0 (4.5)
SCP	**70.5** (9.1)	**88.3** (7.9)	62.5 (15.3)	**77.8** (15.4)	**93.5** (3.0)	**86.7** (6.7)	**82.5** (7.5)	**76.2** (8.7)	71.3 (8.7)	**78.8** (3.5)

Table 1. Patch-level classification accuracies under PLSA on topic vectors learned from labeled patches, for various combinations of the three modalities *S*IFT, *C*olor and *P*osition.

probability maps as in Figure 1. We could also apply the algorithms at pixel level by extracting a patch around each pixel, but this would be computationally expensive.

Each experiment is an average over a specified number of random 90% training / 10% test splits of the data set, with the additional constraint that there must be at least 4 images from each class in each test and training partition. We report averages and standard deviations of patch-level classification accuracies for each class. Some classes are much more frequent than others (21% of pixels are *grass*, while only 3% are *aeroplane* or *face*), so overall accuracies are computed by averaging class-level classification rates, not patch-level ones.

5.1. Combining modalities to improve classification

We first consider how performance depends on the features used: SIFT, color and position. To ensure optimal parameter settings, we learned the topic vectors $p(w|t)$ from labeled training patches. For each test image we estimated the class mixing weights $\boldsymbol{\theta}_d$ using PLSA and found the maximum a posteriori (MAP) label for each patch. The resulting classification accuracies, averaged over 20 random training/test divisions, are shown in Table 1.

Overall SIFT is the most discriminant, but color and to some extent position still provide useful complementary information. We used all three modalities in the experiments below because this gives the best performance both on average and for 7 of the 9 individual classes. Position alone gives zero accuracy for most object classes because their broad spatial distributions are locally dominated by those of the three best-localized classes *grass*, *sky* and *aeroplane*.

5.2. Using spatial priors to improve classification

We now compare PLSA, tree-structured and MRF models for patch-level classification under two training scenarios that are designed to explore the influence of accurate topic vectors: supervised training in which the topic vectors are estimated from Patch-level labels ('P'); and weakly supervised training in which only Image-level labels ('I') are used (*i.e.* training patches are constrained to have a label from their image's label set, but otherwise free). We give results for two different labeling tasks: *Unsupervised labeling* ('U') – nothing is known about the test image and its patches can have any label, and *Weakly supervised labeling* ('W') – a set of image-level labels is given for the test image and each patch must be labeled with one of these (*e.g.*, in Figure 1, each patch must be labeled *building*, *grass*, *tree*, *sky* or *void*).

Table 2 summarizes the patch-level classification rates for the different models and settings. The spatial models have higher accuracy than PLSA for almost all classes and settings, with PLSA-MRF being on average slightly more accurate than PLSA-TREE. Indeed, for unsupervised labelling ('U'), PLSA-MRF with topic vectors learned using PLSA from image-level labels alone has an average accuracy of 78.1%, which is close to the 78.5% attained by PLSA learned from fully labeled patches. These findings are in line with [17], which reports improved classification rates for an aspect model over an irregularly structured MRF built by connecting interest point based patches to their 5 nearest neighbors.

The PLSA-TREE and PLSA-MRF models have similar optimal parameter settings for unsupervised 'U' and weakly-supervised 'W' labelling ($\sigma \approx 0.7$ for the PLSA-MRF), while the non-PLSA TREE and MRF models require different settings for the two tasks: for 'W' the set of available labels is typically quite restricted (88% of the images have four or fewer labels including *void*) and relatively weak spatial couplings are best ($\sigma \approx 0.7$ for the MRF), while for 'U' all 10 labels compete and significantly stronger couplings ($\sigma \approx 2.6$) are needed to prevent the appearance of many small regions with incorrect labels. Including the image-wide PLSA potential $\boldsymbol{\theta}_d$ effectively reweights the observation likelihoods from $p(w|t)$ to $p(w|t)\,\boldsymbol{\theta}_{dt}$, thus suppressing classes for which there is little global evidence and providing image-level regularization that improves the overall accuracy.

5.3. Improved topic learning though MRF inference

The above experiments used the spatial priors only during test image labelling: the topic vectors and the image-

	Model	Building	Grass	Tree	Cow	Sky	Aeroplane	Face	Car	Bicycle	Average
P/U	PLSA	69.8 (10.5)	87.0 (7.3)	63.3 (16.2)	76.3 (14.9)	93.1 (3.2)	85.5 (6.4)	82.7 (6.9)	75.9 (7.8)	72.9 (7.3)	78.5 (3.1)
	PLSA-TREE	73.4 (11.3)	89.0 (7.3)	64.4 (17.4)	75.9 (15.5)	95.1 (3.2)	90.6 (5.5)	87.9 (5.6)	80.4 (9.6)	78.3 (9.6)	81.7 (3.4)
	PLSA-MRF	74.0 (11.6)	88.7 (7.5)	64.4 (17.8)	77.4 (16.3)	95.7 (2.7)	92.2 (4.9)	88.8 (6.3)	81.1 (10.0)	78.7 (10.2)	82.3 (3.4)
I/U	PLSA	40.8 (10.7)	69.6 (9.7)	56.9 (17.3)	79.1 (12.0)	86.1 (6.2)	88.0 (5.8)	91.9 (5.7)	75.8 (9.8)	77.7 (7.5)	74.0 (2.8)
	PLSA-TREE	40.6 (13.5)	71.4 (10.1)	57.6 (18.8)	82.3 (12.7)	88.0 (6.2)	94.4 (4.4)	94.9 (4.8)	83.1 (10.1)	85.5 (7.9)	77.5 (2.9)
	PLSA-MRF	40.1 (14.4)	70.4 (10.3)	56.6 (19.6)	84.2 (13.1)	87.7 (6.4)	95.6 (3.6)	95.5 (4.7)	85.5 (10.2)	87.4 (7.8)	78.1 (3.1)
I/U*	PLSA*	47.8 (10.9)	73.0 (8.9)	62.3 (15.9)	79.7 (10.7)	90.2 (5.2)	90.0 (6.9)	89.7 (6.2)	79.2 (7.8)	77.3 (10.7)	76.6 (3.1)
	MRF*	47.4 (11.3)	75.0 (6.1)	63.2 (13.7)	71.2 (11.2)	92.8 (4.2)	87.1 (8.9)	85.9 (5.0)	75.9 (8.9)	72.1 (11.2)	74.5 (3.1)
	PLSA-MRF*	49.4 (14.0)	73.5 (9.5)	62.9 (17.4)	83.5 (11.0)	91.5 (5.6)	95.5 (4.4)	93.3 (6.0)	85.8 (10.8)	86.4 (8.6)	80.2 (3.3)
P/W	PLSA	80.6 (6.3)	89.0 (6.6)	67.4 (14.6)	80.4 (12.8)	93.4 (2.9)	90.0 (4.6)	89.4 (5.0)	80.6 (7.6)	84.2 (7.2)	83.9 (2.3)
	PLSA-TREE	83.9 (6.5)	90.1 (7.0)	68.5 (15.1)	79.1 (13.8)	95.0 (3.0)	93.6 (3.6)	89.9 (5.5)	82.8 (8.9)	87.1 (7.4)	85.6 (2.5)
	PLSA-MRF	84.1 (6.9)	89.7 (7.3)	69.2 (15.3)	79.5 (14.7)	95.4 (2.8)	94.5 (3.1)	90.7 (5.5)	83.1 (9.7)	87.1 (7.7)	85.9 (2.6)
I/W	PLSA	57.1 (10.1)	72.5 (10.1)	61.8 (15.1)	84.8 (11.4)	84.7 (6.2)	91.5 (4.3)	94.7 (4.9)	85.1 (7.2)	87.7 (7.1)	80.0 (2.6)
	PLSA-TREE	56.9 (12.2)	73.9 (10.8)	63.2 (15.9)	85.0 (12.3)	86.6 (6.2)	95.5 (3.4)	96.8 (4.4)	89.4 (7.5)	92.0 (6.2)	82.2 (2.8)
	PLSA-MRF	55.9 (13.1)	73.0 (10.8)	62.6 (16.6)	85.7 (12.3)	86.1 (6.5)	96.4 (2.9)	96.9 (4.5)	90.4 (7.8)	92.4 (6.3)	82.2 (2.9)

Table 2. Patch-level classification accuracies for the PLSA, PLSA-TREE and PLSA-MRF models, for topic vectors learned from Patch-level labels ('P'), or Image-level labels ('I'). For Unsupervised testing ('U') patch labels can belong to any class, while for Weakly supervised testing ('W') they must belong to the test image's predefined set of keywords. The topic vectors were learned using either standard PLSA (unstarred models) or PLSA-MRF (starred models). The figures in parentheses are standard deviations over 200 random 90% training / 10% test partitions.

level topic mixtures $\boldsymbol{\theta}_d$ were estimated using standard PLSA. To integrate topic vector estimation into the MRF training process, we modified the EM based PLSA algorithm to incorporate the MRF prior in the E-step, using EP to approximate the posterior marginals $p(z_i|\{w_n\}_{n=1}^{n_d})$. We used the learned topic vectors to test three classification methods: (1) standard PLSA; (2) topic-level *MRF* without the image-level PLSA potential; (3) the combined *PLSA-MRF* model as above. Patch-level classification accuracies for test images are reported in Table 2 under I/U*. For both PLSA and PLSA-MRF, the PLSA-MRF based topic vectors give classification accuracies at least comparable to, and often a few percent better than, the PLSA based ones, and the PLSA-MRF model does better than either PLSA or MRF alone. Using PLSA-MRF both to learn topics from keywords and to classify test patches gives an average classification rate of 80.2%, which is quite close to the 82.3% of PLSA-MRF, and better than the 78.5% of plain PLSA, for topics learned from labeled patches. Figure 3 shows some examples of segmentations obtained using respectively PLSA and PLSA-MRF for both topic learning and classification.

5.4. Comparison to Textonboost with 21-class model

Now we compare the combined PLSA-MRF model with Textonboost [21] on the 591 image MSRC data set. For each image a segmentation using 23 semantic classes and *void* is available. Very few pixels belong to *horse* and *mountain* so we treated these as *void*, leaving a total of 21 semantic classes. Textonboost [21] was learned from labelled pixels so for a fair comparison we estimate the PLSA-MRF topics from labeled patches (20×20 pixel patches spaced by 10 pixels as before). The topics were estimated from 276 randomly selected images, with the remaining ones being used for evaluation. (The training/test split is probably different from that used in [21]).

Again $\sigma \approx 0.7$ gave the best classification results. The classification accuracies, per class and averaged over the classes, are reported in Table 3. The two models achieve comparable average accuracies over the 21 classes, but their error rates on individual classes are often quite different (more than 10% for 15 of the 21 classes). As before there are large differences in class frequencies, with *grass* covering 20% of the pixels and *boat* covering only 0.9%. Averaging classification performance over all pixels, Textonboost achieves 72.2% while PLSA-MRF attains 73.5%. The improvement is modest but still suggestive. Example segmentations are shown in Figure 4. When trained using only image-level labels, PLSA-MRF still attains 60.6% classification accuracy over all pixels.

Note that estimating topic vectors from labeled patches is trivial and takes less than 2 seconds, while 11 minutes are required for feature extraction for the 276 images. Most of the training time is spent quantizing the 200k patch features – this takes up to an hour depending on the convergence criterion. Training Textonboost is much slower: 42 hours were required using 276 images [21]. Textonboost required 3 minutes to process a test image, while PLSA-MRF requires 2 seconds per image for feature extraction and 0.3–2 seconds for inference depending on the number of EP iterations.

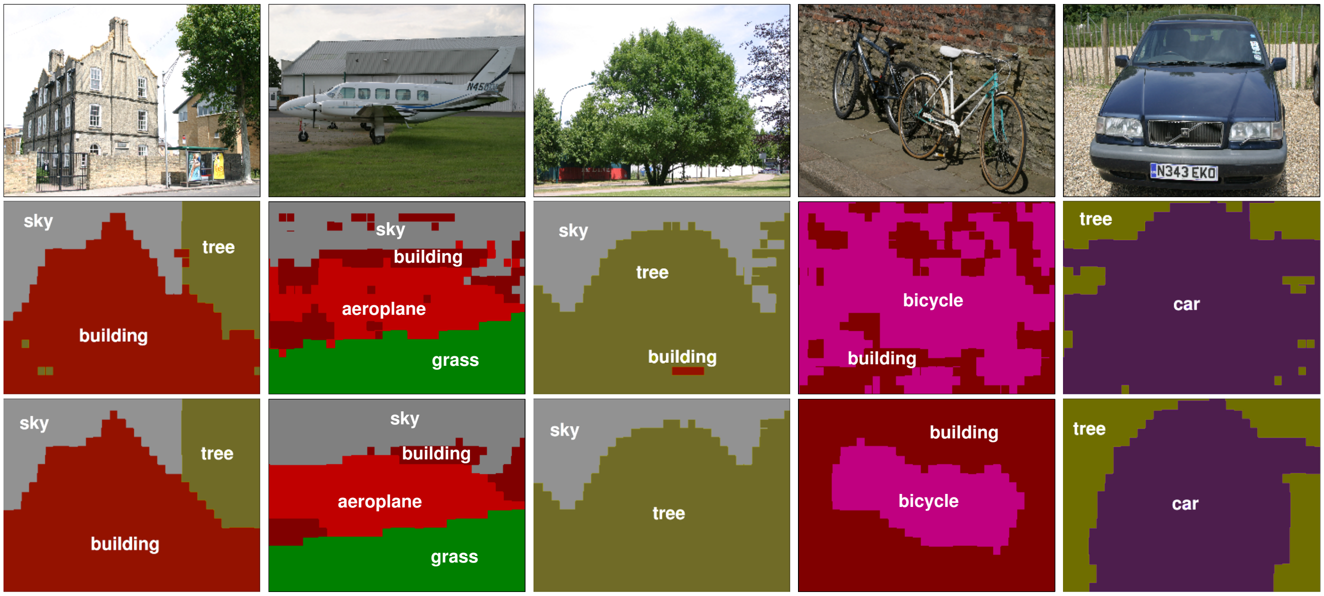

Figure 3. (Best viewed in color). Example images (top), and segmentations using PLSA (middle) and PLSA-MRF (bottom), with topics learned from image labels.

6. Conclusions

We addressed the problem of learning to label image regions with semantic classes from training images that are labeled with image-level keywords rather than with detailed pixel-level segmentations. We showed that aspect models – modelling each image as a mixture of latent 'aspects' or 'topics' – are well suited to this because image-level labels can be taken into account simply by restricting the topics that are used to model the image. We also extended PLSA to incorporate multiple observation modalities for each patch. The combination of SIFT, color, and position cues greatly improves the discrimination: for topics learned from labeled patches, the average classification accuracy increases from 61.3% using only SIFT to 78.8% using all three cues.

We then presented two models that improve the spatial accuracy of the labelling by combining the *global* (image-level) label coupling of PLSA with *local* spatial interactions: PLSA-TREE based on averaging over tree structured couplings; and PLSA-MRF based on an 8-neighbour MRF with Expectation Propagation based inference. In comparison to the 78.5% classification accuracy obtained with PLSA trained from segmented images, PLSA-MRF achieves 80.2% when trained using only image-level labels and 82.3% using patch-level ones. The performance of PLSA-TREE is only slightly lower.

On the 21-class MSRC dataset our models achieve pixel classification accuracies at least as good as those of Textonboost at much lower computational cost.

References

[1] A. Agarwal and B. Triggs. Hyperfeatures - multilevel local coding for visual recognition. In *Proceedings of the European Conference on Computer Vision*, pages 30–43, 2006.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[3] C. Bishop. *Pattern recognition and machine learning*. Spinger-Verlag, 2006.

[4] D. Blei and M. Jordan. Modeling annotated data. In *Proceedings of the 26th annual intetational ACM SIGIR conference*, 2003.

[5] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[6] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *Proceedings of the European Conference on Computer Vision*, pages 517–530, 2006.

[7] P. Carbonetto, N. de Freitas, and K. Barnard. A statistical model for general contextual object recognition. In *Proceedings of the European Conference on Computer Vision*, 2004.

[8] F. Hamze and N. de Freitas. From fields to trees. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence*, volume 20, pages 243–250, 2004.

[9] X. He, R. Zemel, and M. Carreira-Perpiñán. Multiscale conditional random fields for image labelling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 695–702, 2004.

[10] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1/2):177–196, 2001.

[11] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *Proceedings of the*

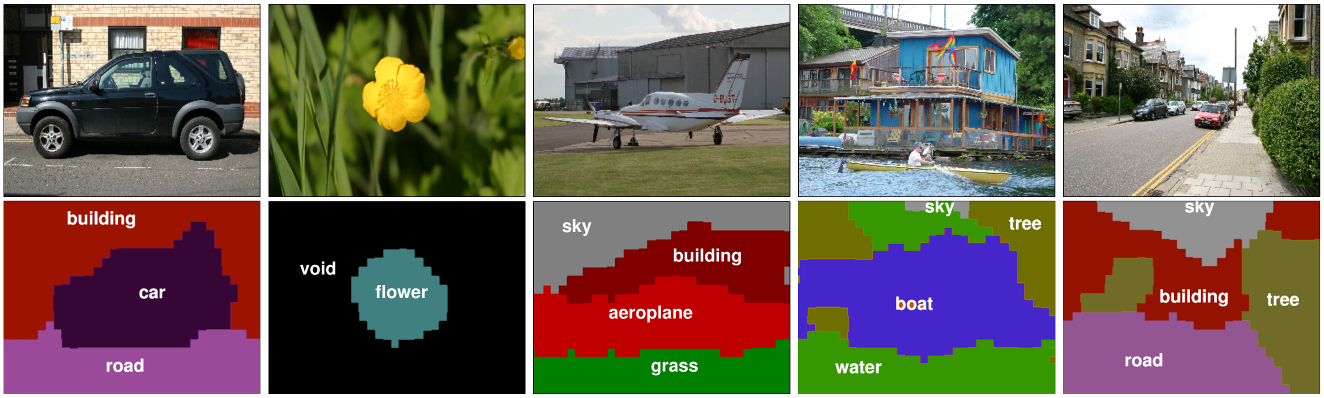

Figure 4. Example images and segmentations using PLSA-MRF on the 21 class data set, with topic vectors learned from labeled patches.

	Building	Grass	Tree	Cow	Sheep	Sky	Aeroplane	Water	Face	Car	Bicycle	Flower	Sign	Bird	Book	Chair	Road	Cat	Dog	Body	Boat	Average
Textonboost	**62**	**98**	**86**	58	50	83	60	53	**74**	63	**75**	63	**35**	**19**	**92**	15	86	**54**	19	**62**	7	58
PLSA-MRF/P	52	87	68	**73**	**84**	**94**	**88**	**73**	70	**68**	74	**89**	33	**19**	78	**34**	**89**	46	**49**	54	**31**	**64**
PLSA-MRF/I	45	64	71	75	74	86	81	47	1	73	55	88	6	6	63	18	80	27	26	55	8	50

Table 3. Classification rates for each of the 21 classes and their average over classes, using Textonboost and our PLSA-MRF model. Textonboost was trained using segmented images, so for a fair comparison we used PLSA-MRF to estimate topics both from labeled patches ('P') and from image-level labels ('I').

IEEE International Conference on Computer Vision, pages 1284–1291, 2005.

[12] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*, volume 18, pages 282–289, 2001.

[13] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization. In *Proceedings of the British Machine Vision Conference*, pages 959–968, 2006.

[14] S. Lazebnik, C. Schmid, and J. Ponce. Affine-invariant local descriptors and neighborhood statistics for texture recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 649–655, 2003.

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[16] T. Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft research, Cambridge, 2005.

[17] F. Monay, P. Quelhas, J.-M. Odobez, and D. Gatica-Perez. Integrating co-occurrence and spatial contexts on patch-based scene segmentation. In *Beyond Patches Workshop, in conjunction with CVPR*, 2006.

[18] J. Puzicha, T. Hofmann, and J. Buhmann. Histogram clustering for unsupervised segmentation and image retrieval. *Pattern Recognition Letters*, 20:899–909, 1999.

[19] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. Van-Gool. Modeling scenes with local descriptors and latent aspects. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 883–890, 2005.

[20] B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1605–1614, 2006.

[21] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 1–15, 2006.

[22] J. Sivic, B. Russell, A. Efros, A. Zisserman, and B. Freeman. Discovering objects and their location in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 370–377, 2005.

[23] J. van de Weijer and C. Schmid. Coloring local feature extraction. In *Proceedings of the European Conference on Computer Vision*, pages 334–348, 2006.

[24] M. Welling, T. Minka, and Y.-W. Teh. Structured region graphs: morphing EP into GBP. In *Proceedings of the Annual Conference on Uncertainty in Artificial Intelligence*, volume 21, 2005.