Latent Dirichlet Markov Random Fields for Semi-supervised Image Segmentation and Object Recognition

Lester Mackey
Department of Electrical Engineering and Computer Science
University of California, Berkeley
Berkeley, CA 94720
`lmackey AT eecs DOT berkeley DOT edu`

December 18, 2007

Abstract

Topic models such as Latent Dirichlet Allocation (LDA) and probabilistic Latent Semantic Analysis have shown great success in segmenting and recognizing the component objects of images. However, such models frequently ignore the spatial relationships among image regions and hence fail to capture important local correlations. In this paper, we introduce the Latent Dirichlet Markov Random Field (LDMRF), a model which improves the spatial coherence of LDA by introducing a Markov random field prior over hidden object class labels. We evaluate the model on a number of semi-supervised joint segmentation and object recognition tasks and compare its performance with LDA. We further demonstrate the advantages of multi-modal feature selection for LDA and LDMRF. Finally, we propose a new combination of variational inference procedures for approximate inference in the LDMRF model and demonstrate its efficacy.

1 Introduction

Two problems of wide interest in the vision community are *image segmentation*, dividing an image into its distinct, semantically meaningful regions, and *object recognition*, labeling the regions of images according to their semantic object classes. Solutions to these problems are at the core of applications like content-based image retrieval, video surveying, and object tracking. While segmentation and recognition are often studied independently in the literature ([1, 2, 3, 4, 9, 10, 5, 6]), many recent works have focused on the joint problem of simultaneously recognizing and localizing objects within images ([11, 12, 7, 16, 13, 15, 17, 14, 18, 19, 20]).

Studies of simultaneous recognition and segmentation frequently focus on the extremes of the supervision spectrum, either considering completely supervised image models ([7, 16, 13, 15]), with class labels for every pixel or region of an image, or completely unsupervised image models ([17, 14, 18]) with no auxiliary annotations or segmentation information. Both extremes have their disadvantages. The supervised setting offers a predefined interpretation of learned classes and the potential for greater accuracy. However, such supervised learning comes at the cost of labeling each pixel or region of every training image, a task too expensive for many image collections. Unsupervised learning dispenses with the need for manual labeling but suffers from identifiability issues. That is, when an unsupervised model labels image regions, human intervention (or a separate procedure) is needed to map those labels back to meaningful object classes.

One middle ground explored by several authors ([19, 20]) is that of “weak supervision,” labeling each image with only the names of the categories it contains. This drastically reduces the amount of external information needed but still requires some degree of annotation for every image. Moreover, both weakly supervised and unsupervised models ignore detailed annotations when they are available. A second middle ground, semi-supervised learning, escapes these problems by accommodating both labeled and unlabeled training data. Models designed for the semi-supervised setting thus benefit both from the accuracy and identifiability granted by labeled images and from the relative abundance of unlabeled images. In this paper, we will focus on the task of joint segmentation and recognition in a semi-supervised setting.

Latent topic models such as Probabilistic Latent Semantic Analysis (pLSA) [22] and Latent Dirichlet Allocation (LDA) [21] naturally lend themselves to our task, as they have shown impressive performance on unsupervised classification and recognition tasks ([18]) and are readily extensible to the semi-supervised setting. However, topic models suffer from poor spatial coherence: they ignore local label correlations by assuming a conditional independence of labels given the image.

In [20], the authors demonstrate that the segmentation and recognition accuracy of the pLSA model can be improved by introducing a Markov Random Field (MRF) to enforce spatial coherence on the labels of an image. The authors analyze their Markov Random Field Aspect Model (pLSA-MRF) in fully supervised and weakly supervised settings.

In this paper, we give similar treatment to LDA, extending the topic model with a Markov random field prior over its hidden labels. The resulting Latent Dirichlet Markov Random Field (LDMRF) exhibits significant boosts in classification and segmentation performance over standard LDA. In addition, LDMRF inherits the benefits of LDA over pLSA. Whereas the number of parameters of pLSA-MRF grows linearly with the number of training images, the number of LDMRF parameters does not grow with the training set size and hence is less susceptible to overfitting. Moreover, LDMRF provides a well-defined model of previously unseen images, whereas pLSA-MRF relies on a heuristic ”folding-in” procedure ([22]) to analyze new images.

The LDMRF model also poses new challenges to learning, due to its doubly intractable structure. As a result, we propose a new combination of variational inference procedure for the model and demonstrate its efficacy in the semi-supervised setting. Further, we achieve increased joint segmentation and classification accuracy by extending the set of image features used in [20] to train pLSA-MRF.

This paper is structured as follows. Section 2 describes the latent Dirichlet Markov random field model. Feature selection is discussed in Section 3. Section 4 describes our approach to inference and parameter estimation. Section 5 introduces the data set used for evaluation. Experiments and results are detailed in Section 6. Finally, we summarize our contributions in Section 7.

2 The Latent Dirichlet Markov Random Field Model

2.1 LDA

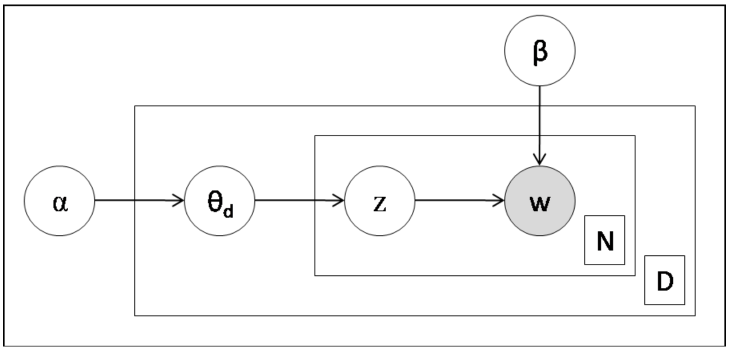


Figure 1: Graphical model representation of LDA.

Latent Dirichlet allocation was originally developed for text document modeling, and we will use the terminology of that field to describe the model. LDA is a generative model for documents. It asserts that every document is a finite mixture over latent topics; each topic is in turn a mixture over words. A graphical model representation is depicted in Figure 1.

Associated with every document, d, is a vector of topic weights, θ_d, drawn independently from a Dirichlet distribution parameterized by a global hyperparameter, α. θ_d represents the mixing proportions of topics specific to that document. Thus for any topic z, $\theta_{dz} = p(z|d)$.

An underlying topic is drawn for each word of a document, and all topics of a document are assumed to be conditionally independent of one another given the document vector. This is often called the "bag-of-words" or "bag-of-features" assumption.

Associated with every topic, z, is a topic vector, β_z, representing the probability of each word being generated by that topic. Thus for any word, w, $\beta_{zw} = p(w|z)$. Note that the topic vectors are global parameters shared by all documents. We will represent the vectors as a single matrix, β, where topics index the rows of the matrix.

In the terms of our problem, each image is a document, and one word is assigned to every n x n pixel patch extracted from a regular grid covering the image. Words are thus feature descriptors representing the patches of an image. The topic associated with each feature is a label corresponding to the object class which produced that feature. For example, one word of an image might be a descriptor for a small square of blue pixels, and the associated topic might be "sky".

In assuming that image topics are conditionally independent, LDA ignores the known spatial relationship among image patches. This bag-of-features assumption is responsible for the spatial incoherence of LDA in the image setting.

2.2 Improving spatial coherence with an MRF prior

To improve the spatial coherence of our image model, we move from a bag-of-features to a "grid-of-features" representation. The latent Dirichlet Markov random field, depicted as a graphical model in Figure 2, introduces explicit couplings between the labels of neighboring patches of an image. This allows the model to capture local correlations that would be missed under the conditional independence assumption of LDA. The transition from LDA to LDMRF is equivalent to placing a Markov random field prior on the vector of labels, z:

$$p(z|\theta_d) \propto exp(\sum_i log\theta_{dz_i} + \sum_{N(i,j)} f(z_i, z_j)) \tag{1}$$

Here, i runs over all patches in the image and $N(i,j)$ runs over all pairs of neighboring patches. The terms $f(z_i, z_j)$ are edge potentials between neighboring nodes. As in [20], we parameterize these potentials with a single parameter σ, according to the Potts model

$$f(z_i, z_j) = \sigma \cdot [z_i = z_j] \tag{2}$$

where $[\cdot]$ represents the indicator function and σ is set empirically.

A positive value of σ awards configurations in which neighboring patches have the same label; a negative value discourages these configurations. Note that if we set σ to 0 we obtain the original decoupled LDA model.

We have experimented with four-neighbor connectivity (connecting each topic node to its neighbors in the four cardinal directions) and eight-neighbor connectivity (connecting each topic node to its eight surrounding neighbors). We report results from the eight-neighbor connectivity model.

3 Feature Selection

To capture different aspects of the image data, we use words drawn from multiple modalities to describe each image patch. The authors of [20] suggest three complementary modalities: texture, hue, and location. In this work, we introduce a fourth: opponent angle.

To describe hue, we use the robust hue descriptor of [25], which grants invariance to illuminant variations, lighting geometry, and specularities. For texture description we use "dense SIFT" features

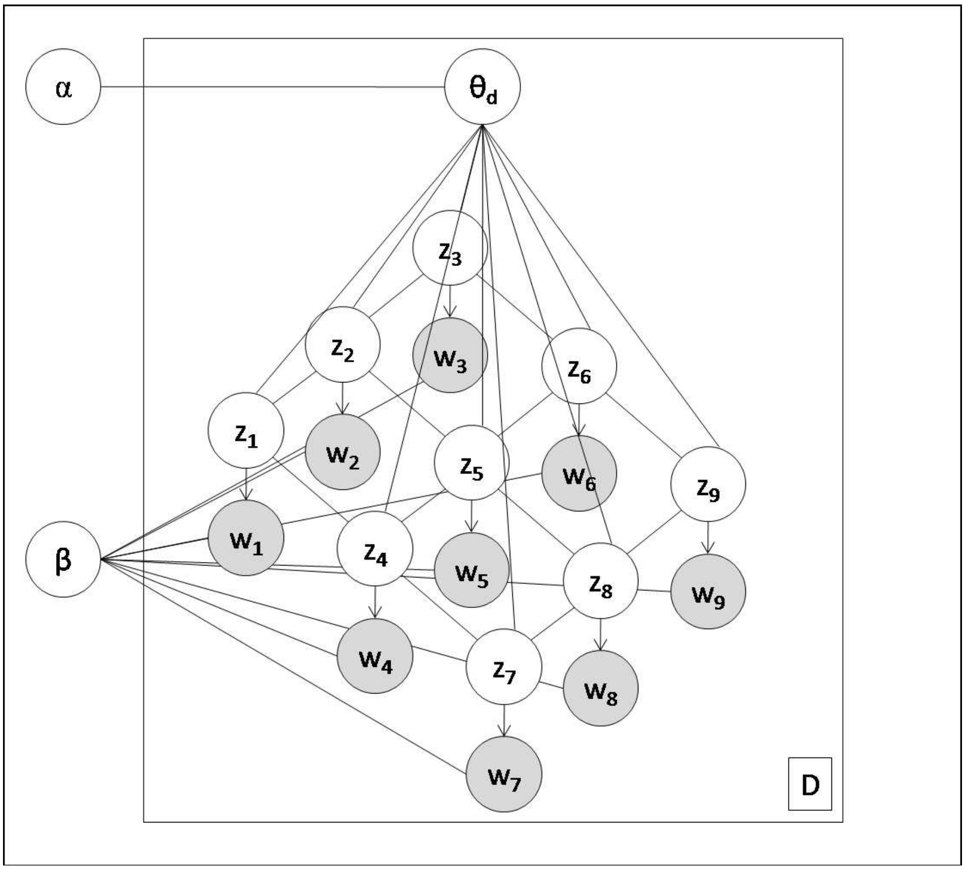


Figure 2: LDMRF with 4-neighbor connectivity: w-nodes are patch-level image features, z-nodes are latent object classes.

[24, 23], histograms of oriented gradients computed not at local keypoints but rather at a single scale over each patch. To describe coarse location, we cover each image with a regular c x c grid of cells and assign each patch the index of the covering cell. We have experimented with $c = 5$ and $c = 10$ and achieve better performance with the latter. The opponent angle descriptor of [25] captures a second characterization of image patch color. These features are invariant to specularities, illuminant variations, and diffuse lighting conditions.

To build a discrete visual vocabulary from these raw descriptors, we vector quantize the dense SIFT, hue, and opponent angle descriptors using k-means, producing 1000, 100, and 100 clusters respectively. The word describing each patch is formed by concatenating the corresponding modality components into a single vector. However, learning a vocabulary with $10^7 \cdot c^2$ possible values is infeasible, so we introduce the following assumption into our model (as in [20]): the multiple modality word components describing a patch are conditionally independent given the latent topic of the patch. This assumption is equivalent to the following conditional probability factorization:

$$p(w|z) = \prod_{m=1}^{M} p(w^m|z) \tag{3}$$

where M is the number of modalities and w^m is the mth component (mth modality value) of word w.

4 Inference and Parameter Estimation

4.1 Learning the model

Training the LDMRF model is equivalent to estimating the model parameters from a set of training images. Given that we are working in a semi-supervised setting, expectation-maximization (E-M) is a natural choice for parameter estimation. E-M iterates between an E-step, estimating the values of unknown variables (i.e. document vectors and the patch labels of unlabeled images), and an M-step, computing maximum likelihood estimates of model parameters given the estimated variable values.

The E-step of the E-M algorithm requires that we compute the posterior distribution of the unknown variables given the observed image features:

$$p(\theta, z|w, \alpha, \beta) = \frac{p(\theta, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}. \tag{4}$$

However, this distribution is doubly intractable to compute, first because of the implicit coupling between θ and β ([21]) and second because of the intractable nature of inference in MRFs. Fortunately, approximate inference can be achieved through Markov Chain Monte Carlo methods or variational inference procedures. In this paper, we propose a method for variational inference.

4.2 Variational inference

To eliminate the first source of intractability, we introduce a variational distribution in which θ_d is decoupled from z:

$$q(\theta_d, z|\gamma_d, \phi) = q(\theta_d|\gamma_d)q(z|\phi). \tag{5}$$

Here ϕ and γ_d are image-specific variational parameters that depend on the observed words of the image. γ_d is a Dirichlet parameter for θ_d under the variational distribution. ϕ_i is the vector of node potentials for class label z_i, so that

$$p(z|\phi) \propto exp(\sum_i log\phi_{iz_i} + \sum_{N(i,j)} f(z_i, z_j)). \tag{6}$$

This family of distributions, shown in graphical form in Figure 3, is related to the family introduced in [21] for approximate inference in the LDA model. Indeed, if we remove the edges between z nodes we recover the original LDA variational family. However, the presence of these edges implies that exact inference is intractable even in our variational model. We deal with this second instance of intractability by embedding a second variational approximation into the standard LDA

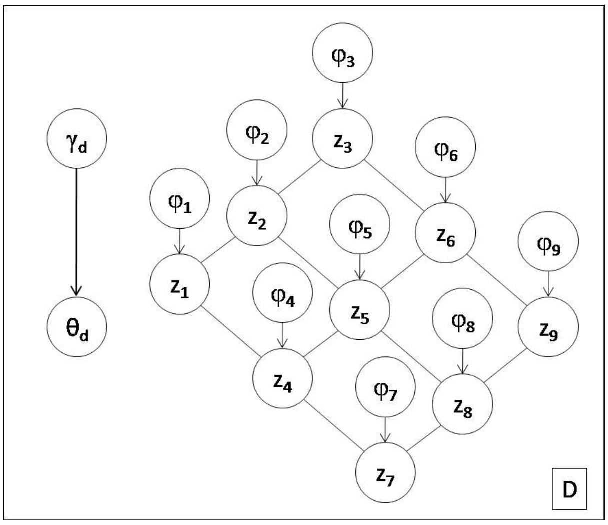


Figure 3: Variational distribution for approximating LDMRF posterior.

variational inference algorithm detailed in [21]. The original algorithm updates the ϕ parameters assuming uncoupled label nodes; we instead approximate these posterior node potentials with the pseudomarginals returned by running Tree-Reweighted Belief Propagation[1] (TRWBP) [26] on the variational MRF. We have experimented with a number of constant edge appearance probabilities ($\mu_{ij} = a, \forall i, j$ s.t. $N(i, j)$) , and results are reported for $\mu_{ij} = 1$ (note that this corresponds to standard loopy belief propagation).

Our full inference procedure learns the parameters of the variational model by alternating TRWBP updates of ϕ and the standard variational updates of γ until convergence. We use the resulting variational distribution as a surrogate for the posterior. To learn the original LDMRF parameters, we interleave the variational E-step of approximating the posterior with the M-step of updating α and β as described in [21].

4.3 Inferring class labels

To obtain a segmentation and labeling for a test image, the final task is to infer the most probable configuration of labels given the estimated parameters and the test words. Note that there is a distinction between inferring the most probable label for each patch and inferring the most probable configuration of labels. When labels are especially correlated or anticorrelated, favored configurations may differ significantly from those derived by examining each node individually. Thus, we cannot simply maximize the pseudomarginals returned by TRWBP for our final inference, for these estimate only locally probable configurations. Instead, at test time, we culminate the inference procedure described above with a computation of approximate max-marginals using the Tree-Reweighted Max Product algorithm[2] (TRWMP) [27]. We maximize over these max-marginals to obtain our predicted maximum a posteriori class label configuration.

5 Evaluation Data

We use the Microsoft Research Cambridge pixel-wise labeled image database v1[3] in our experiments. The data set consists of 240 images, each of size 213 x 320 pixels (some were rotated to conform to this standard). Each image has an associated pixel-wise ground truth labeling. Each pixel is labeled as belonging to one of 13 semantic classes or to the *void* class. Pixels have a ground truth label of *void* when they do not belong to any semantic class or when they lie on the boundaries between classes in an image (to ease the burden of manual segmentation). It is noted on the associated website that there are not sufficiently many instances of *horse*, *mountain*, *sheep*, or *water* to learn these classes, so, as in [20], we treat these ground truth labels as *void* as well. Thus, our general task is to learn and segment the remaining 9 semantic object classes.

In the experiments which follow, we use 20 x 20 pixel patches spaced at 10 pixel intervals across the image, and we assume a eight-neighbor connectivity LDMRF model. As in [20], we assign ground truth patch labels based on the most frequent pixel label within a patch. For prediction, we assign to each pixel the label predicted for the nearest patch center.

Unless otherwise specified, our patch-level classification accuracies are averaged over twenty randomly generated 90% training / 10% test divisions of the data set. We utilize semi-supervised training in each experiment, so our model will only view the ground-truth labels of a specified percentage of all training images. We measure overall accuracy by averaging over class-level rates (an average over patch level rates would be unfairly biased towards the most frequent classes).

6 Experiments and Results

6.1 Experiment 1: Merits of multiple modalities

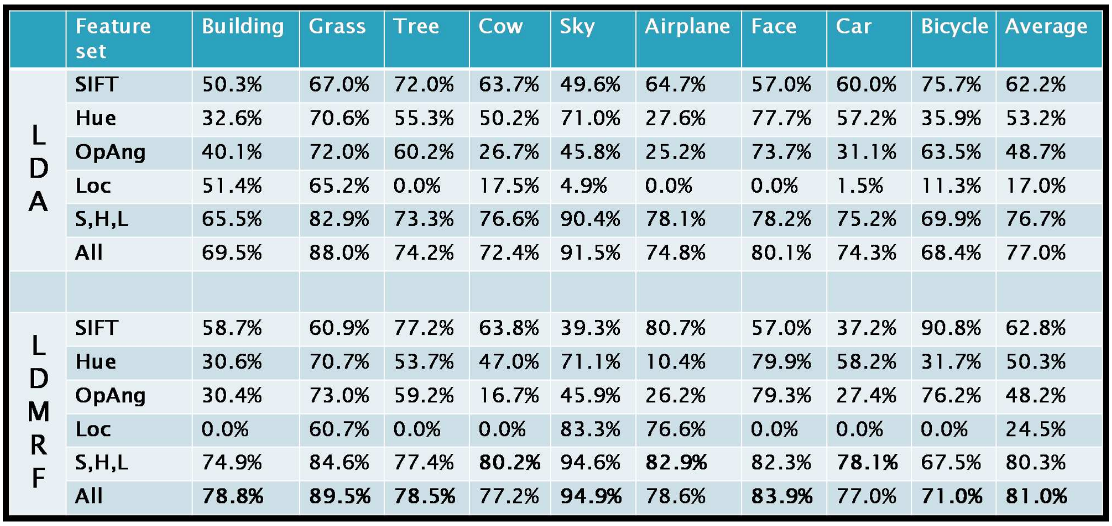
In our first experiment we examine the merits of a multi-modal model, that is, a model which extracts features from multiple modalities. We train four single modality versions of LDMRF, one trimodal

[1] For a detailed description of the TRWBP algorithm, see [26].

[2] For a detailed description of the TRWMP algorithm, see [27].

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

version, incorporating only the three modalities suggested in [20], and the full model, utilizing all four modalities. We implement LDA using the same feature bases for comparison. The resulting per-semantic-class classification rates averaged over twenty data splits are reported in Figure 4. We use 99% of all training labels in this experiment.



	Feature set	Building	Grass	Tree	Cow	Sky	Airplane	Face	Car	Bicycle	Average
LDA	SIFT	50.3%	67.0%	72.0%	63.7%	49.6%	64.7%	57.0%	60.0%	75.7%	62.2%
	Hue	32.6%	70.6%	55.3%	50.2%	71.0%	27.6%	77.7%	57.2%	35.9%	53.2%
	OpAng	40.1%	72.0%	60.2%	26.7%	45.8%	25.2%	73.7%	31.1%	63.5%	48.7%
	Loc	51.4%	65.2%	0.0%	17.5%	4.9%	0.0%	0.0%	1.5%	11.3%	17.0%
	S,H,L	65.5%	82.9%	73.3%	76.6%	90.4%	78.1%	78.2%	75.2%	69.9%	76.7%
	All	69.5%	88.0%	74.2%	72.4%	91.5%	74.8%	80.1%	74.3%	68.4%	77.0%
LDMRF	SIFT	58.7%	60.9%	77.2%	63.8%	39.3%	80.7%	57.0%	37.2%	90.8%	62.8%
	Hue	30.6%	70.7%	53.7%	47.0%	71.1%	10.4%	79.9%	58.2%	31.7%	50.3%
	OpAng	30.4%	73.0%	59.2%	16.7%	45.9%	26.2%	79.3%	27.4%	76.2%	48.2%
	Loc	0.0%	60.7%	0.0%	0.0%	83.3%	76.6%	0.0%	0.0%	0.0%	24.5%
	S,H,L	74.9%	84.6%	77.4%	**80.2%**	94.6%	**82.9%**	82.3%	**78.1%**	67.5%	80.3%
	All	**78.8%**	**89.5%**	**78.5%**	77.2%	**94.9%**	78.6%	**83.9%**	77.0%	**71.0%**	**81.0%**

Figure 4: Supervised patch-level classification accuracies for LDA and LDMRF trained over various modalities. S = SIFT, H = hue, L = location.

It is clear from both the LDA and LDMRF results that the SIFT-based texture is the most discriminant of our modalities. Grid location on the other hand leads to decidedly poor object recognition. This coarse descriptor detects most classes with 0% percent accuracy and only recognizes the most readily localizable classes, like *sky*.

The results also show the merit of incorporating multiple modalities into our model. The trimodal model of hue, location, and SIFT significantly outperforms the unimodal SIFT model, and introducing opponent angle extends accuracy even further. These results give evidence that our other modalities are capturing complementary information, not discernible through the SIFT vocabulary alone.

Finally, although the LDMRF and LDA models perform comparably when trained on most single modalities, the LDMRF significantly outperforms LDA in multi-modal settings. Indeed, the gains in accuracy for the multi-modal models are comparable to those reported for pLSA-MRF over pLSA in [20] (the authors do not report unimodal pLSA-MRF performance). Interestingly, LDMRF realizes the greatest accuracy increase on the least discriminant feature modality: location.

Several example segmentations produced by quadmodal LDA and quadmodal LDMRF are displayed in Figure 5.

6.2 Experiment 2: Robustness to fewer labeled examples

In our second experiment, we test the robustness of quadmodal LDA and LDMRF to decreased proportions of labeled training examples. We fix the training set size, and on each evaluation run, we allow our models to view a varying fraction of the labels of the training set images. The remaining training images are incorporated as unlabeled training data. We apply this procedure to a range of percentages of labeled data and evaluate the model on twenty random training-test splits per percentage. The results are summarized in Figure 6.

We observe a similar pattern of decline in the LDA and LDMRF models: for most object categories, accuracy falls slowly but steadily as the percentage of supervised training examples falls from 100% to 30%. Nevertheless, the LDMRF model continues to outperform the LDA model in this semi-supervised environment. Indeed, LDMRF trained with only 60% labeled data achieves

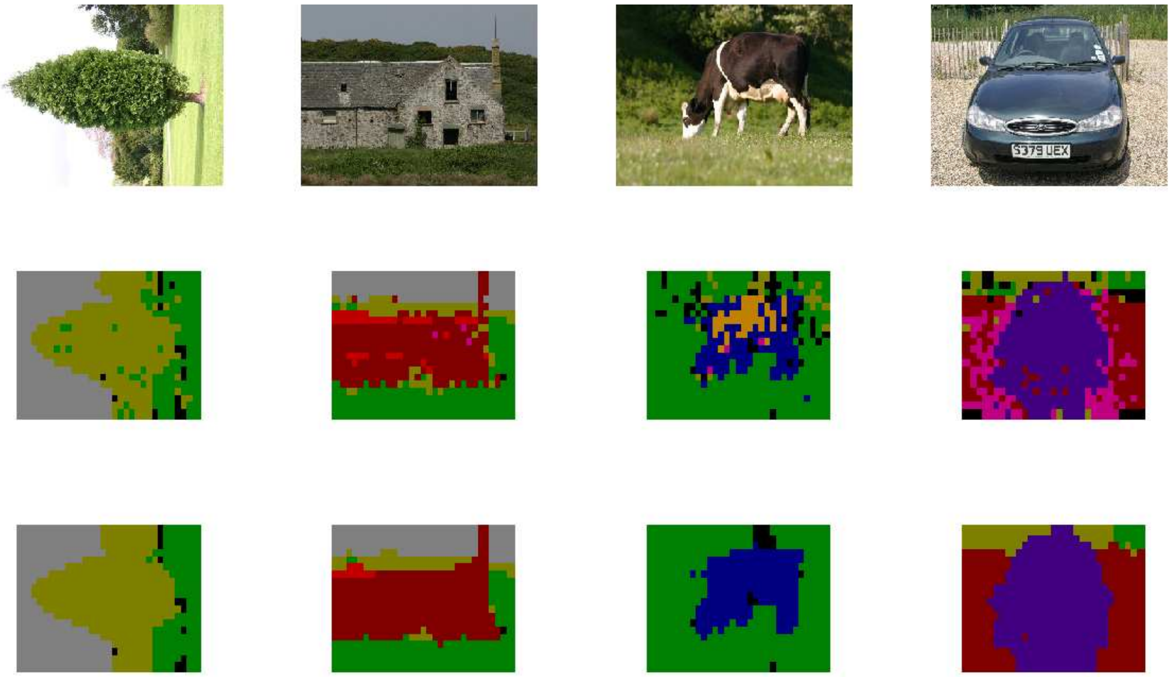
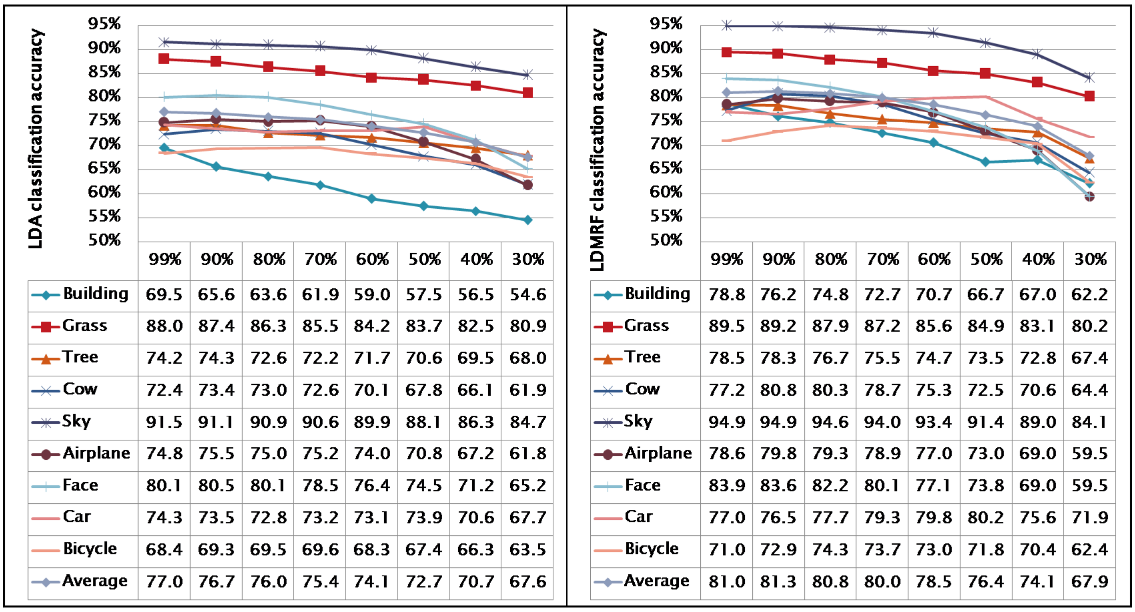


Figure 5: Example test images from the MSRC Dataset (top) and segmentations produced by quadmodal LDA (middle) and quadmodal LDMRF (bottom). (Best viewed in color.)



	99%	90%	80%	70%	60%	50%	40%	30%
Building	69.5	65.6	63.6	61.9	59.0	57.5	56.5	54.6
Grass	88.0	87.4	86.3	85.5	84.2	83.7	82.5	80.9
Tree	74.2	74.3	72.6	72.2	71.7	70.6	69.5	68.0
Cow	72.4	73.4	73.0	72.6	70.1	67.8	66.1	61.9
Sky	91.5	91.1	90.9	90.6	89.9	88.1	86.3	84.7
Airplane	74.8	75.5	75.0	75.2	74.0	70.8	67.2	61.8
Face	80.1	80.5	80.1	78.5	76.4	74.5	71.2	65.2
Car	74.3	73.5	72.8	73.2	73.1	73.9	70.6	67.7
Bicycle	68.4	69.3	69.5	69.6	68.3	67.4	66.3	63.5
Average	77.0	76.7	76.0	75.4	74.1	72.7	70.7	67.6

	99%	90%	80%	70%	60%	50%	40%	30%
Building	78.8	76.2	74.8	72.7	70.7	66.7	67.0	62.2
Grass	89.5	89.2	87.9	87.2	85.6	84.9	83.1	80.2
Tree	78.5	78.3	76.7	75.5	74.7	73.5	72.8	67.4
Cow	77.2	80.8	80.3	78.7	75.3	72.5	70.6	64.4
Sky	94.9	94.9	94.6	94.0	93.4	91.4	89.0	84.1
Airplane	78.6	79.8	79.3	78.9	77.0	73.0	69.0	59.5
Face	83.9	83.6	82.2	80.1	77.1	73.8	69.0	59.5
Car	77.0	76.5	77.7	79.3	79.8	80.2	75.6	71.9
Bicycle	71.0	72.9	74.3	73.7	73.0	71.8	70.4	62.4
Average	81.0	81.3	80.8	80.0	78.5	76.4	74.1	67.9

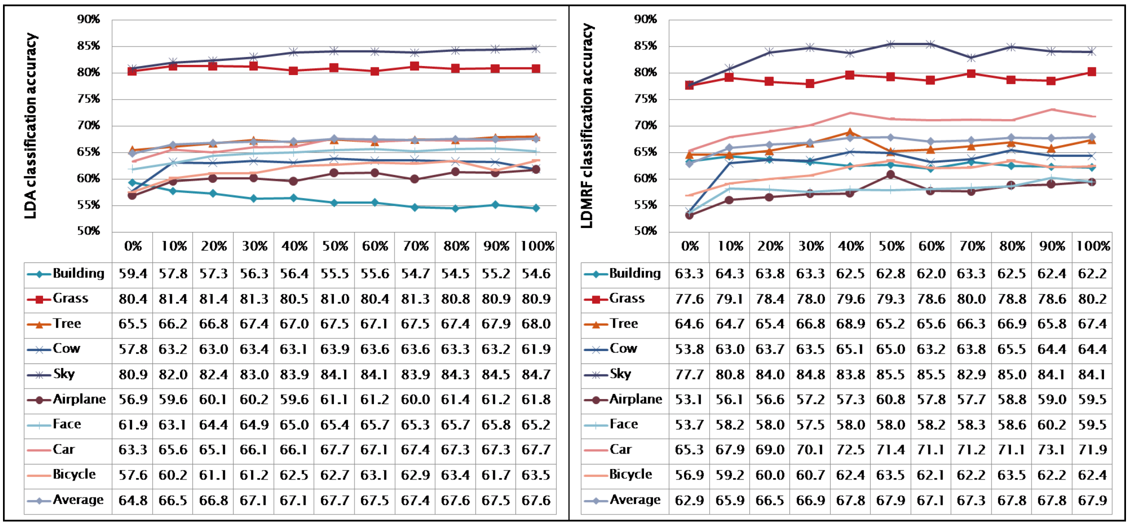
Figure 6: Variation of LDA (left) and LDMRF (right) classification accuracy with percentage of labeled training examples

higher accuracy than the 99% supervised LDA model. Further, the LDMRF model maintains a high degree of discrimination in all test cases, classifying with 74% accuracy with access to only 40% of the training labels. For comparison, recall that LDMRF unimodal classification with SIFT achieves 63% accuracy on the 99% labeled training set. Thus, multi-modality once again shows its benefit.

A more surprising occurrence revealed by Figure 6 is the large jump in *cow*, *airplane*, and *bicycle* classification accuracy realized by both models in the transition from 99% to 90% labeled training data. It is possible that the models are overfitting the training data in the 99% setting and that the greater number of unknown variables in the 90% setting allows for a more regularized and hence more robust model.

6.3 Experiment 3: Enhanced accuracy using additional unlabeled training images

Our third experiment demonstrates how accuracy improves in LDA and LDMRF when additional unlabeled training images are made available. We begin with our standard training set size and select a random 30% to serve as labeled data for our models. We ignore the remaining training data and evaluate the resulting models. We then repeat this procedure, adding in anywhere from 10% to 100% of the remaining training data as additional unlabeled examples on top of the labeled examples. The results of this experiment are displayed in Figure 7.



	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Building	59.4	57.8	57.3	56.3	56.4	55.5	55.6	54.7	54.5	55.2	54.6
Grass	80.4	81.4	81.4	81.3	80.5	81.0	80.4	81.3	80.8	80.9	80.9
Tree	65.5	66.2	66.8	67.4	67.0	67.5	67.1	67.5	67.4	67.9	68.0
Cow	57.8	63.2	63.0	63.4	63.1	63.9	63.6	63.6	63.3	63.2	61.9
Sky	80.9	82.0	82.4	83.0	83.9	84.1	84.1	83.9	84.3	84.5	84.7
Airplane	56.9	59.6	60.1	60.2	59.6	61.1	61.2	60.0	61.4	61.2	61.8
Face	61.9	63.1	64.4	64.9	65.0	65.4	65.7	65.3	65.7	65.8	65.2
Car	63.3	65.6	65.1	66.1	66.1	67.7	67.1	67.4	67.3	67.3	67.7
Bicycle	57.6	60.2	61.1	61.2	62.5	62.7	63.1	62.9	63.4	61.7	63.5
Average	64.8	66.5	66.8	67.1	67.1	67.7	67.5	67.4	67.6	67.5	67.6

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Building	63.3	64.3	63.8	63.3	62.5	62.8	62.0	63.3	62.5	62.4	62.2
Grass	77.6	79.1	78.4	78.0	79.6	79.3	78.6	80.0	78.8	78.6	80.2
Tree	64.6	64.7	65.4	66.8	68.9	65.2	65.6	66.3	66.9	65.8	67.4
Cow	53.8	63.0	63.7	63.5	65.1	65.0	63.2	63.8	65.5	64.4	64.4
Sky	77.7	80.8	84.0	84.8	83.8	85.5	85.5	82.9	85.0	84.1	84.1
Airplane	53.1	56.1	56.6	57.2	57.3	60.8	57.8	57.7	58.8	59.0	59.5
Face	53.7	58.2	58.0	57.5	58.0	58.0	58.2	58.3	58.6	60.2	59.5
Car	65.3	67.9	69.0	70.1	72.5	71.4	71.1	71.2	71.1	73.1	71.9
Bicycle	56.9	59.2	60.0	60.7	62.4	63.5	62.1	62.2	63.5	62.2	62.4
Average	62.9	65.9	66.5	66.9	67.8	67.9	67.1	67.3	67.8	67.8	67.9

Figure 7: Variation of LDA (left) and LDMRF (right) classification accuracy with percentage of unlabeled training examples added. (See text for details.)

Again we observe similar behavior by the LDA and LDMRF models: both models enjoy moderate growth in classification accuracy as unsupervised images are added. We were limited in this study by the size of the dataset, but an interesting focus of future work would be the limits of this process. Given sufficiently many unsupervised images, would the model eventually become saturated and unable to increase its performance? If so, what is its asymptotic performance level? A metric like asymptotic performance level could prove useful in model selection or in ranking the utility of various learning procedures. Moreover, such a study could place an upper bound on how much data is required to maximize performance.

Another notable aspect of our experiment 3 results is the overall performance of LDMRF under the impoverished conditions of 30% labeled and 0% additional unlabeled data. At 63% accuracy, the multimodal LDMRF model still achieves classification rates on par with the 99% supervised unimodal SIFT LDMRF.

Finally, it is noteworthy that while recognition accuracies for all other classes show a net increase as unlabeled images are added, recognition rates for buildings decrease significantly under both the LDA and LDMRF models. This phenomenon may relate to the relative frequency with which buildings co-occur with objects of other categories. In the MSRC dataset, buildings appear frequently in images focused on bicycles, cars, airplanes, and even cows. In addition, the images focused on buildings frequently feature sky, grass, and trees. Thus, the building category appears in close proximity to nearly every other category. Moreover, unlike sky or grass, the building category is not readily localizable to one region of an image. Taken together, these conditions make building a difficult category to learn and an easy category to confuse in a a principally unsupervised setting.

7 Conclusion

We have introduced the latent Dircihlet Markov random field, a new model for spatially coherent image segmentation and object recognition, and compared its performance with LDA in a number of semi-supervised segmentation and classification tasks. We have additionally augmented the feature base suggested by [20] for multi-modal training and demonstrated the advantages of drawing features from multiple modalities. Finally, we have proposed a new combination of variational procedures for inference and learning in the LDMRF and demonstrated its efficacy and accuracy in learning joint segmentations and classifications.

Acknowledgments

I would like to thank Erik Sudderth for his conversations and advice, Joost van de Weijer for his color descriptor code, Rob Fergus for his k-means implementation, and Gyuri Dork for the use of his dense SIFT executable.

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. PAMI, 24(4):509522, 2002.

[2] A. C. Berg, T. L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. 2005 Conference on Computer Vision and Pattern Recognition (CVPR 2005), June 2005, pp. 26-33.

[3] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In Proc. of IEEE CVPR, Madison, WI, June 2003.

[4] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, New York, June 2006, vol. II, pp. 2169-2178.

[5] J. Shotton, A. Blake, and R. Cipolla. Contour-Based Learning for Object Detection, ICCV05(I: 503-510).

[6] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In Proc. ICCV, 2005.

[7] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop on Perceptual Organization in Computer Vision, June 2004.

[8] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In Proceedings of the IEEE International Conference on Computer Vision, 2003.

[9] A. Levin and Y. Weiss. Learning to combine bottom-up and top-down segmentation. In ECCV, 2006.

[10] M. K. Schneider. Multiscale methods for the segmentation of images. M.S. thesis, Mass. Inst. Technol., Cambridge, May 1996.

[11] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio. Robust Object Recognition with Cortex-Like Mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence, March 2007, 29(3), 411426.

[12] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In Proc. ICCV, 2007.

[13] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labelling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 695702, 2004.

[14] F. Schroff, A. Criminisi, and A. Zisserman. Single-histogram class models for image segmentation. In Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing, 2006.

[15] 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 115, 2006.

[16] J. Winn and J. Shotton. The Layout Consistent Random Field for Recognizing and Segmenting Partially Occluded Objects. In CVPR 2006.

[17] B. Russell, A. Efros, J. Sivic, W. Freeman, and A. Zisserman. Using Multiple Segmentations to Discover Objects and their Extent in Image Collections. CVPR, 2006.

[18] J. Sivic, B. C. Russell, A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. MIT AI Lab Memo AIM-2005-005, MIT, 2005.

[19] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In Proceedings of the Workshop on Statistical Learning in Computer Vision, Prague, Czech Republic, May 2004.

[20] J. Verbeek and B. Triggs. Region classification with Markov field aspect models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[21] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. Journal of Machine Learning Research, 3:9931022, 2003.

[22] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. Machine Learning, 43:177196, 2001.

[23] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, volume II, pages 886893, 2005.

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

[25] J. van de Weijer and C. Schmid. Coloring local feature extraction. In Proceedings of the European Conference on Computer Vision, pages 334348, 2006.

[26] M. J. Wainwright, T. Jaakkola, and A. S. Willsky. Tree-reweighted belief propagation and approximate ml estimation by pseudo-moment matching. In 9th Workshop on Artificial Intelligence and Statistics, 2003.

[27] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Exact MAP estimates by (hyper)tree agreement. In NIPS, volume 15, December 2002.