Image Segmentation with Topic Random Field

Bin Zhao[1], Li Fei-Fei[2], and Eric P. Xing[1]

[1]School of Computer Science, Carnegie Mellon University
[2]Computer Science Department, Stanford University

Abstract. Recently, there has been increasing interests in applying aspect models (e.g., PLSA and LDA) in image segmentation. However, these models ignore spatial relationships among local topic labels in an image and suffers from information loss by representing image feature using the index of its closest match in the codebook. In this paper, we propose Topic Random Field (TRF) to tackle these two problems. Specifically, TRF defines a Markov Random Field over hidden labels of an image, to enforce the spatial coherence between topic labels for neighboring regions. Moreover, TRF utilizes a noise channel to model the generation of local image features, and avoids the off-line process of building visual codebook. We provide details of variational inference and parameter learning for TRF. Experimental evaluations on three image data sets show that TRF achieves better segmentation performance.

1 Introduction

Image segmentation represents a fundamental problem in computer vision, which aims to cluster pixels in an image into distinct, semantically coherent and salient regions [1,2,3]. Solutions to image segmentation serves as the basis for a wide range of applications including object recognition, content-based image retrieval, video surveillance and object tracking [4].

Although geometry-based methods such as normalized cuts [1] remain an effective approach to image segmentation, motivated by the success of probabilistic aspect models, such as the probabilistic latent semantic analysis (PLSA) [5] and the latent Dirichlet allocation (LDA) [6], in text analysis and information retrieval, there has been a growing interest in applying such models for semantically-driven segmentation of natural images [7,8,9,10,11,12]. Among various advantages offered by these approaches, is their affordances for unsupervised training of representations of the latent aspects underlying a content-rich corpora, often known as *topics*, which can help define a semantically meaningful "content space" in which an image can lie. Thus the segmental results derived from an aspect model (also known as topic model) can be more reliant on content coherence, rather than mere spatial contiguity as in the spectrum methods. Other advantages include flexibility in capturing content granularity [7], and computational efficiency based on efficient approximate inference.

To apply those aspect models originally proposed for text data, it is necessary to first build a connection between an image and a text document. While text

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part V, LNCS 6315, pp. 785–798, 2010.

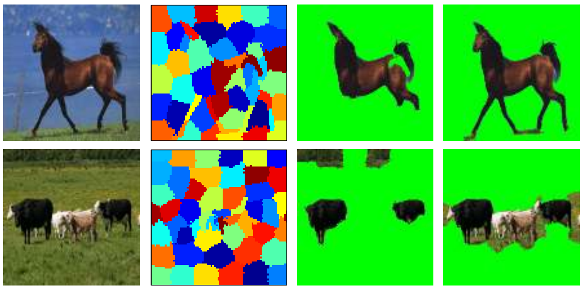


Fig. 1. (Best viewed in color) Comparison of the spatial latent Dirichlet allocation (spatial LDA) model [13], one of the state-of-the-art aspect model for image analysis, and our proposed model, topic random field (TRF). First column shows the input images. Second column shows the input regions to spatial LDA and TRF provided by an over-segmentation method (normalized cut in this paper). The last two columns show the segmentation results of spatial LDA and TRF respectively. The first row indicates that by defining MRF over latent topics, TRF enforces spatial coherency over adjacent regions, while spatial LDA separates adjacent and semantically similar regions into two segments. The second row shows that using noise channel instead of codebook enables TRF to group two visually non-identical objects (black cow and white cow) from the same semantic class into the same category, while the spatial LDA model categorizes these two objects into different semantic classes.

documents are naturally composed from a vocabulary of distinctive words, an image is made from a collection of pixels, and there is no such obvious word-level representation for images. Conventionally, researchers extract various local features, for example, interest points detected by scale invariant saliency detector [14] as used in [13], and transform these local features to “visual words”, which play the same role as textual words in text analysis. Typically, after extracting local features from training images, a clustering is performed on the entire set of local features. Then a dictionary is constructed, with “words” being the centroids of the feature clusters. Based on this dictionary, each feature extracted from the image is then represented by the index of the most similar item (i.e. a visual word defined by the feature centroid) in the dictionary. Finally, analogous to text data, an image is represented as a collection of visual words, obtained by assigning every local feature an index in the visual dictionary.

Despite the success of modern low-level visual feature detectors, the aspect model built upon those local features suffers from several weaknesses. First, most existing aspect models regard an image as a bag of visual words, ignoring the spatial relationship between them. Although the spatial relationship between words in text documents might not severely affect content distillation, the spatial relationship between visual words are crucial for image understanding. For example, a scrambled collection of patches from a building image does not necessarily evoke the recognition of a building [11]. Most current work on aspect model of images ignores this important issue, hence might have compromised the final accuracy of the segmentation and recognition tasks. This contrasts the spectrum methods for which spatial contiguity is crucial in defining segmental patterns. Second, representing each local image feature by the index of the item that is closest to it in the dictionary can result in severe loss of information. Due to the usual high dimensionality of local features extracted from images, it is impractical to build a large size dictionary that could enumerate all possible local features. Therefore, it is highly possible that even the closest matching visual word in the dictionary for a particular local feature instance can be quite different from the feature instance itself, and the matched visual word might even represent a mismatching content, thereby causing ambiguity in feature-instance versus visual-word matching. This phenomena has never been an issue in text modeling, where a word-instantiation in a document can be always unambiguously mapped to a word in the dictionary. We suspect that these two problems could seriously hinder the application of aspect models on image data.

In this paper, we propose a *Topic Random Field* (TRF) model for image segmentation, which improves over the basic LDA-style models, by defining a Markov Random Field (MRF) over hidden topic assignment of super-pixels in an image to enforce the spatial coherence between neighboring regions; and by employing a noise channel between visual words in the dictionary and instantiated super-pixels in the real image to better model the variance of local features. Specifically, instead of assuming that the latent topic assignments of every super-pixels in an image are generated independently according to a multinomial distribution, a TRF defines an MRF over the hidden super-pixels' labels to model their spatial relationship. Moreover, different from previous attempts, which first build a codebook off-line and then generate each local feature instantiation according to a multinomial distribution over word-index, a TRF generates each local feature instantiation as a corrupted or transformed version of a matching visual word in the codebook according to a noise-channel model, which allows explicit modeling and inference of the ambiguity of the matching between feature instantiation and feature prototypes (i.e., visual word). As a result, TRF avoids the problem of information loss during topic learning without building a large size codebook, and is significantly more robust to variability in the instantiations of local features corresponding to the same objects or common visual words due to variations in lighting, transformation, viewing angle, etc.

It should be noted that there has been some attempt in utilizing spatial relationships between topic labels to improve the performance of aspect models on image segmentation [13,11,15,16,17]. Probably the most related work to this paper is the spatial-LDA model [13], which also considers utilizing spatial consistency, by defining latent topic variables on over-segmented regions and enforcing all local patches within the region to share the same latent topic. In fact, we adopt a similar way of defining latent topic variables on over-segmented regions to enforce spatial consistency between local patches within the same region. However, in spatial-LDA model, the authors only consider the spatial consistency between adjacent local patches, while topic labels for over-segmented regions are assumed to be generated independently. Empirical comparison between spatial-LDA and TRF demonstrate the necessity of enforcing spatial consistency between adjacent over-segmented regions. Besides, in [11], the authors demonstrated that the performance of PLSA can be improved by introducing an image-specific MRF to enforce the spatial coherence on the labels of the fine-grained local patches in an image. However, in that model the number of parameters grows linearly with the number of training images and the model is trained fully supervised. On the other hand, the number of parameters in TRF does not grow with the size of the training data because we apply a universally-parameterized MRF within the TRF over all images, which can be trained via a maximum likelihood principle in a fully unsupervised fashion. Applying a universal MRF rather than an image-specific one as in [11] is crucial to avoid overfitting and enable scalability. Moreover, [11] builds an MRF on local patches. Since there might be several hundreds of patches in one image, the resulting MRF is quite large; whereas our approach defines an MRF on over-segmented regions usually with homogeneous object-level contents, whose number in an image is around 50, and enforces the consistency among local semantically similar and adjacent patches by enforcing them to share the same latent topic. Therefore, the MRF in our model is much smaller than the one in [11], yet enforces the same amount of spatial consistency. In our empirical studies, we found that training TRF takes much less time than training the model in [11] on the same data set, which makes TRF more practical for web-scale image analysis.

Unlike the attempt on utilizing spatial relationships between topic labels to improve aspect model, as far as we are concerned, the noise channel presented in this paper is the first attempt in modeling visual feature generation without building a codebook in this topic-model based image analysis. Despite the fact that noise model could tolerate variability in the instantiations of local features due to variations in lighting, transformation, viewing angle, etc, using a noise channel also avoids the hassle of building a codebook off-line.

In summary, the main contributions of this paper can be highlighted as the follows: (1) The *Topic Random Field* provides a probabilistically sound framework for modeling spatial coherency within an aspect model. (2) TRF offers a more principled approach for addressing the ambiguity in feature-instance versus visual-word matching, and for codebook construction via unsupervised maximum likelihood learning during training the TRF (rather then via an off-line preprocessing). (3) The conjoint effect of a spatial MRF on topic labels and a noise-chanel codebook lead to a segmental algorithm that takes into consideration of both semantic and spatial coherence, without any supervision. Figure 1 illustrates topic random field's novelty by comparing the segmentation results of spatial-LDA and TRF.

The rest of this paper is organized as follows. We briefly review the image representation employed in aspect models for image segmentation problems, and describe the visual features we utilize in this paper. We introduce the topic random field model in Section 3. Section 4 presents the details of variational inference and parameter learning for this model. We give experimental results on three image data sets in Section 5, followed by conclusions in Section 6.

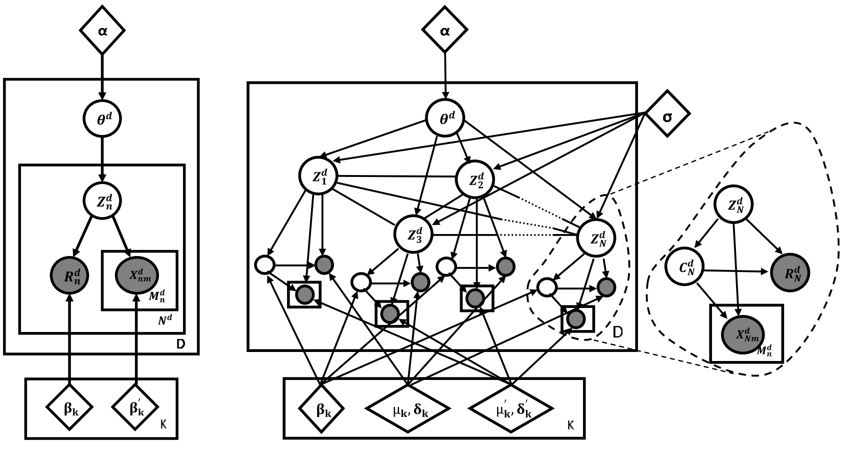
2 Preliminary: Image Representation

Given an image, TRF starts with an initial over-segmentation of the image by partitioning it into multiple homogeneous regions. To ensure that pixels in a region belongs to the same object and avoid obtaining regions larger than the objects we want to segment, we start with an over-segmentation of the images using spectral clustering [1]. For each over-segmented region, we extract 4 types of region-level features: shape, color, location and texture. specifically, the shape features include the centered object mask in a canonical 32×32 frame, the size of the region, and the size of region's bounding box, which results in a 1027 dimensional vector [18]. The color features include the mean RGB value, its standard deviation and a color histogram. The location information extracted from each region is represented by a coarse 8×8 absolute segmentation mask as well as the height of the top-most and bottom-most pixel in the region [18]. Finally, the texture features are average responses of filter banks in each region. Besides region-level features, we also extract pixel-level features within each segmented region. Specifically, we find a number of scale invariant interest points and describe them by SIFT [19].

3 Topic Random Field

In this section, we will introduce the *Topic Random Field* and explain in detail the generative process of this model. As discussed in the first section, TRF improves the spatial LDA model [13], a specially designed topic model for image segmentation, in two perspectives: the incorporation of an MRF over the hidden labels in the image and the introduction of a noise model for generating image features. To better understand the motivation of TRF, we first briefly describe the spatial LDA model as depicted in figure 2(a).

Given an image I^d ($d \in \{1, 2, \ldots, D\}$) and its over-segmented regions $n = 1, 2, \ldots, N^d$, the spatial LDA model defines a latent topic z_n^d to represent the label of region n. Topics in image data have similar meanings as they do in text data: a



(a) Spatial LDA (b) TRF

Fig. 2. Graphical model representation of Topic Random Field and comparison with spatial LDA model

topic represents category identity of an object, e.g., buildings, horses, cars, trees, etc. Suppose there are totally K topics within the image collection, then for each region n, $z_n^d \in \{1,\ldots,K\}$. Each topic z_n^d then generates a region-level feature R_n^d, for example the average filter responses in the region, and M_n^d pixel-level features $\{\mathbf{X}_{nm}^d\}_{m=1}^{M_n^d}$, such as the detected salient points described by SIFT. In order to do image segmentation using topic model, we first need to infer the hidden topic label for each over-segmented region and then group all the regions with the same topic label to form an object.

We will take an example to explain the generative process of spatial LDA: suppose we want to generate a "building" image I^d. First, we will draw a probability vector $\boldsymbol{\theta}^d$ which determines what intermediate topics to select to generate each region of the image. For a building image, $\boldsymbol{\theta}^d$ should privilege topics like "glasses", "walls", etc. Then, to create each region in the image, we determine a group of particular topics $\{z_n^d\}_{n=1}^{N^d}$ out of the mixture of possible topics. For example, if a "glass" topic is selected, this will in turn give preference on some codewords that occur more frequently in glasses. Finally, we draw codewords R_n^d and $\{\mathbf{X}_{nm}^d\}_{m=1}^{M_n^d}$ to describe the appearance of region n. The process of drawing both the topic and codewords will be repeated N^d times, eventually forming an entire bag of visual words that would construct an image of buildings.

3.1 Spatial MRF over Topic Assignments

The basic model ignores the spatial structure of the image, modeling its regions as independent draws from the topic mixing vector $\boldsymbol{\theta}^d$. However, the labels for adjacent regions tend to be strongly correlated in real images. TRF extends spatial LDA by enforcing spatial coherence among neighboring regions. Specifically, to enforce spatial coherence over hidden topic labels in our image model, we move from a multinomial distribution over hidden topics to a Markov Random Field. The topic random field, depicted as a generative model in Figure 2(b), introduces explicit couplings between the labels of adjacent regions in an image. This allows the TRF model the ability to capture local correlations that would be missed under the conditional independence assumption of spatial LDA. The transition from spatial LDA to TRF is equivalent to placing an MRF prior on hidden topic labels $\mathbf{z}^d$:

$$p(\mathbf{z}^d|\boldsymbol{\theta}^d,\sigma) = \tfrac{1}{A(\boldsymbol{\theta}^d,\sigma)} \exp\left[\textstyle\sum_n \sum_k z_{nk}^d \log\theta_k^d + \sum_{n\sim m} \sigma I(z_n^d = z_m^d)\right] \tag{1}$$

where I is the indicator function, n runs through all over-segmented regions in the image, k runs through all possible topics, $n \sim m$ means that z_n^d and z_m^d are connected by an edge in the graphical model, and $A(\boldsymbol{\theta}^d,\sigma)$ is the normalizing factor

$$A(\boldsymbol{\theta}^d,\sigma) = \textstyle\sum_{\mathbf{z}^d} \exp\left[\sum_n \sum_k z_{nk}^d \log\theta_k^d + \sum_{n\sim m} \sigma I(z_n^d = z_m^d)\right] \tag{2}$$

A positive value of σ awards configurations in which neighboring regions have the same label. Moreover, if we set $\sigma = 0$, i.e., assume the hidden topic labels are generated independently, $A(\boldsymbol{\theta}^d,\sigma) = 1$ and $\mathbf{z}^d$ follows a multinomial distribution parameterized by $\boldsymbol{\theta}^d$, and this gives us exactly the spatial LDA model.

Throughout this paper, we assume the Markov Random Field structure is known. Although structure learning over latent variables could be an interesting problem, it is not our intention to tackle this problem in current paper. The Markov Random Field is built by connecting a region with its nearest k neighbors.

3.2 Noise Channel over Codebook

Despite of the empirical success of aspect models on image data [12], one should be careful with the distinction between text data and image data. Representing each word by its index in the dictionary incurs no loss of information, and we could still recover that exact word using the index and the dictionary. However, due to the fact that there is no natural counterpart of words and dictionary in image data, we have to manually build a dictionary. Different from text data, representing each visual feature by the index of its most similar visual word in the dictionary will lose information about that particular local feature, since it is highly possible that there might not be an exact match in the dictionary we built. One would probably argue that we could alleviate this problem by building a large dictionary, to make sure every possible local detector has exact or close enough match in the dictionary. However, different from text word, visual words are usually high dimensional, to ensure each visual word has exact match in the dictionary would render the dictionary so large that no practical inference algorithm could solve the resulting model.

Therefore, the size of the codebook becomes crucial: a small codebook would incur heavy information loss, while a large codebook could render the model too difficult to solve. However, a closer look into the problem reveals that although it is not possible to exactly match every visual feature to a visual word in the dictionary, we could always find an entry in the dictionary such that the visual feature could be represented by this entry plus some noise. For example, given features extracted from a tree image, it is highly possible that we could extract similar features from another tree image. This intuition tells us that we could find several "prototype" visual features for an object, and model features extracted from the same object by these prototype features plus some noise. Therefore, each object is represented by a group of prototype features, and feature extracted from each individual image is the combination of prototype feature and noise.

To ease the description of the model, in the rest of this paper, we use $\mathbf{x}_n^d$ to represent both region-level feature R_n^d and pixel-level features $\{\mathbf{X}_{nm}^d\}_{m=1}^{M_n^d}$. The generative process for visual features could then be modeled as a two-step process: first draw the prototype indicator c_n^d according to a multinomial distribution $p(c_n^d|z_n^d, \boldsymbol{\beta})$, then draw the visual feature $\mathbf{x}_n^d$ using a noise model $p(\mathbf{x}_n^d|c_n^d, z_n^d, \boldsymbol{\mu}, \delta)$, where $\boldsymbol{\mu}$ and δ are parameters. Specifically, in this paper, we employ a Gaussian noise model, where $\boldsymbol{\mu}$ is the mean vector and δ^2 is the variance. Suppose the number of possible prototype features for each object is L_k, then $c_n^d \in \{1, \ldots, L_k\}$. For simplicity, we assume the number of different prototypes for all objects are the same, say L. Then the Gaussian noise model is

$$p(\mathbf{x}_n^d|c_n^d = l, z_n^d = k, \boldsymbol{\mu}, \delta) \propto \exp\left\{-\frac{(\mathbf{x}_n^d - \boldsymbol{\mu}_{kl})^T(\mathbf{x}_n^d - \boldsymbol{\mu}_{kl})}{2\delta_{kl}^2}\right\} \tag{3}$$

Note that by introducing the noise model, we no longer need to build a codebook off-line. The prototype features are learned during the training process, and are stored in the mean vectors $\boldsymbol{\mu}$. This could also be understood as building a "codebook" online, where L is the size of the codebook for each object. The optimal value of L could be determined using Bayesian information criterion [20].

3.3 The Proposed Model

The generative process of *Topic Random Field* is as follows:

- For each image I^d, draw the prior distribution of $\boldsymbol{\theta}^d$ according to a Dirichlet distribution parameterized by $\boldsymbol{\alpha}$;
- Draw hidden topic labels $\{z_1^d, \ldots, z_{N^d}^d\}$ according to Markov random field parameterized by $\boldsymbol{\theta}^d$;
- For each over-segmented region $n \in \{1, \ldots, N^d\}$:
  - Draw a prototype appearance indicator $c_n^d | z_n^d$ ~Mult($\boldsymbol{\beta}$);
  - Draw region-level and pixel-level appearance features according to the noise model $p(\mathbf{x}_n^d | c_n^d, z_n^d, \boldsymbol{\mu}, \delta)$

Putting the generative process together, the joint distribution of $\{\boldsymbol{\theta}^d, \mathbf{z}^d, \mathbf{c}^d, \mathbf{x}^d\}$ given an image I^d can be written as

$$
\begin{aligned}
&p(\boldsymbol{\theta}^d, \mathbf{z}^d, \mathbf{c}^d, \mathbf{x}^d | \boldsymbol{\alpha}, \sigma, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\delta}) \qquad (4)\\
&= p(\boldsymbol{\theta}^d | \boldsymbol{\alpha}) p(\mathbf{z}^d | \boldsymbol{\theta}^d) \prod_{n=1}^{N^d} p(\mathbf{c}_n^d | \mathbf{z}_n^d, \boldsymbol{\beta}) p(\mathbf{x}_n^d | \mathbf{z}_n^d, \mathbf{c}_n^d, \boldsymbol{\mu}, \boldsymbol{\delta})\\
&= \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K (\theta_k^d)^{\alpha_k - 1} \frac{1}{A(\boldsymbol{\theta}^d, \sigma)} \left(\prod_{n=1}^{N^d} \prod_{k=1}^K (\theta_k^d)^{z_{n,k}^d} \right) \exp\left[\sum_{n \sim m} \sigma (\mathbf{z}_n^d)^T \mathbf{z}_m^d \right]\\
&\cdot \prod_{n=1}^{N^d} \left\{ \prod_{k=1}^K \prod_{l=1}^L \left[\beta_{kl} p(\mathbf{x}_n^d | \boldsymbol{\mu}_{kl}, \delta_{kl}) \right]^{z_{n,k}^d c_{n,l}^d} \right\}
\end{aligned}
$$

where we abuse the notation by defining $z_{n,k}^d = 1$ if and only if $z_n^d = k$, and $c_{n,l}^d = 1$ if and only if $c_n^d = l$. $p(\mathbf{x}_n^d | \boldsymbol{\mu}_{kl}, \delta_{kl})$ is the noise model parameterized with $(\boldsymbol{\mu}_{kl}, \delta_{kl})$. After training the model, we label the region r with $(z_r^d)^*$ such that

$$
(z_r^d)^* = \arg\max_{z_r^d} p(\mathbf{x}_r^d | z_r^d) \qquad (5)
$$

The regions with the specific $(z_r^d)^*$ constitute the interested object.

4 Variational Inference and Parameter Learning

The central challenge in using TRF is computing the posterior distribution of hidden variables given an image: $p(\boldsymbol{\theta}^d, \mathbf{c}^d, \mathbf{z}^d | \mathbf{x}^d)$. In general, this distribution is intractable to compute due to the dependence between $\boldsymbol{\theta}^d$, $\mathbf{c}^d$ and $\mathbf{z}^d$, once conditioned on some observations. Various variational inference algorithms have been proposed in the machine learning literature to solve this problem. In this paper, we employ mean field variational inference to efficiently obtain an approximation to this distribution. Specifically, mean field variational inference algorithm forms a factorized distribution of the latent variables, parameterized by free variables known as variational parameters [21].

$$q(\boldsymbol{\theta}^d, \mathbf{z}^d, \mathbf{c}^d | \boldsymbol{\gamma}^d, \boldsymbol{\rho}^d, \boldsymbol{\xi}^d) = q(\boldsymbol{\theta}^d | \boldsymbol{\gamma}^d) \prod_{n=1}^{N^d} q(\mathbf{z}_n^d | \boldsymbol{\rho}_n^d) q(\mathbf{c}_n^d | \boldsymbol{\xi}_n^d) \tag{6}$$

where the Dirichlet parameters $\boldsymbol{\gamma}^d$ and the multinomial parameters $(\boldsymbol{\rho}_1^d, \ldots, \boldsymbol{\rho}_N^d)$, $(\boldsymbol{\xi}_1^d, \ldots, \boldsymbol{\xi}_N^d)$ are variational variables. These parameters are fit by minimizing the Kullback-Leibler (KL) divergence between the approximated and true posterior [21]. We begin with bounding the log likelihood of an image I^d by Jensen's inequality. Specifically, we use variational EM algorithm to do inference and parameter learning for the TRF model. As shown in Algorithm 1, the E-step optimizes the variational parameters $\{\boldsymbol{\gamma}^d, \boldsymbol{\xi}^d, \boldsymbol{\rho}^d\}$ as follows[1]

$$\gamma_k^d = \alpha_k^d + \sum_{n=1}^{N^d} \rho_{nk}^d, \;\; \lambda^d = e^{|E^d|\sigma} \tag{7}$$

$$\xi_{nl}^d \propto \prod_{k=1}^{K} \left\{ \beta_{kl} \left(\frac{1}{2\pi\delta_{kl}^2} \right)^{\frac{m}{2}} \exp[-\frac{(\mathbf{x}_n^d - \boldsymbol{\mu}_{kl})^T (\mathbf{x}_n^d - \boldsymbol{\mu}_{kl})}{2\delta_{kl}^2}] \right\}^{\rho_{nk}^d} \tag{8}$$

$$\rho_{nk}^d \propto \exp[\Psi(\gamma_k^d) - \Psi(\sum_{k=1}^{K} \gamma_k^d) + \sum_{m \in \mathcal{N}(n)} \sigma \rho_{mk}^d]$$
$$\cdot \prod_{l=1}^{L} \left\{ \beta_{kl} \left(\frac{1}{2\pi\delta_{kl}^2} \right)^{\frac{m}{2}} \exp[-\frac{(\mathbf{x}_n^d - \boldsymbol{\mu}_{kl})^T (\mathbf{x}_n^d - \boldsymbol{\mu}_{kl})}{2\delta_{kl}^2}] \right\}^{\xi_{nl}^d} \tag{9}$$

and the M-step optimizes model parameters $\{\boldsymbol{\alpha}, \sigma, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\delta}\}$

$$\beta_{kl} \propto \sum_{d=1}^{D} \sum_{n=1}^{N^d} \xi_{nl}^d \rho_{nk}^d, \;\; \boldsymbol{\mu}_{kl} = \frac{\sum_{d=1}^{D} \sum_{n=1}^{N^d} \xi_{nl}^d \rho_{nk}^d \mathbf{x}_n^d}{\sum_{d=1}^{D} \sum_{n=1}^{N^d} \xi_{nl}^d \rho_{nk}^d} \tag{10}$$

$$\delta_{kl}^2 = \frac{\sum_{d=1}^{D} \sum_{n=1}^{N^d} \xi_{nl}^d \rho_{nk}^d (\mathbf{x}_n^d - \boldsymbol{\mu}_{kl})^T (\mathbf{x}_n^d - \boldsymbol{\mu}_{kl})}{m \sum_{d=1}^{D} \sum_{n=1}^{N^d} \xi_{nl}^d \rho_{nk}^d} \tag{11}$$

$$\sigma = \frac{1}{|E|} \log \frac{\sum_{d=1}^{D} \sum_{k=1}^{K} \sum_{n \sim m} \rho_{nk}^d \rho_{mk}^d}{\sum_{d=1}^{D} \frac{1}{\lambda^d}} \tag{12}$$

5 Experiments

In this section, we show the empirical performance of topic random field for image segmentation, both qualitatively and quantitatively.

5.1 Data Sets

We use three data sets in our experiments, which are selected to cover a wide range of properties. Specifically, those data sets include

- **Weizmann data set [22].** The data set contains 328 images of horses with different poses, sizes, face directions, backgrounds and illumination conditions. Each image has a ground truth segmentation that labels out the horse. There is only one horse in each image, and there is a single object category in the data set: horse.

[1] Here we omit the details due to the space limit. The derivation of variational inference and parameter learning for TRF is provided in the supplemental material.

Algorithm 1. Variational EM for topic random field

```
repeat
    E-step: For each image $I^d$, update $\{\boldsymbol{\gamma}^d, \lambda^d, \boldsymbol{\xi}^d, \boldsymbol{\rho}^d\}$ using equations (7), (8), and (9);
    M-step: Update $\{\sigma, \boldsymbol{\beta}, \boldsymbol{\mu}, \boldsymbol{\delta}\}$ using equations (10), (11), (12), and update $\boldsymbol{\alpha}$ using the linear-time Newton-Raphson algorithm described in [6].
until The increase of log likelihood between two consecutive iterations is less than $\epsilon$
```

- **Microsoft object recognition data set [23].** This data set involves 182 images of cows, facing three different directions: left, right and front. Moreover, some cow pictures also contain multiple instances and significant occlusions. Similar to the Weizmann data set, there is a single object category in the data set, but there might be multiple objects in one image.
- **MSRC pixel-wise labeled image database** [2]**.** There are 240, 213 × 320 pixel images in this data set. Each pixel belongs to one of 13 semantic classes or to the void class.There are multiple objects in one image, and multiple object categories in the data set.

5.2 Experimental Setups and Comparisons

We have conducted comprehensive performance evaluations by testing our method under different circumstances. Specifically, to better understand the effect of introducing MRF on latent topics to enforce spatial consistency and use of noise model to better model image feature generation, we study the model adding only MRF on latent topics and adding only noise model separately, and compare with the TRF model. We use the spatial LDA model [13], which is state-of-the-art aspect model for image segmentation, as baseline and also compare with spectral clustering. The algorithms that we evaluated are listed below.

- **Spatial LDA** [13]. The implementation is the same as in [13]. We use the same region-level and pixel-level features as in our TRF model.
- **LDA+MRF**. This model is based on spatial LDA [13], with the only modification of introducing a Markov random field on the latent topics. Thus, this model could be viewed as the TRF model without noise channel. For each image I^d, we set $L = 20$ and build a Markov random field on $\mathbf{z}^d$ by connecting each $\mathbf{z}_n^d$ with its nearest 4 neighbors.
- **LDA+noise** Similar with LDA+MRF, this model adds a noise channel in the spatial LDA model. Hence, this model could be regarded as the TRF model without Markov random field on the latent topics.
- **TRF**. We build MRF for each image in the same way as LDA+MRF.
- **Normalized cuts. (NCut)** [1]. The implementation code is downloaded from `http://www.cis.upenn.edu/~jshi/software/`.

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

5.3 Image Segmentation Results

Since the Weizman data set provides ground truth segmentations, we could assess the segmentation result quantitatively. Regions sharing the same latent topics z are grouped into the same segment, and the percentage of pixels in agreement with the ground truth segmentation is used to measure the performance of segmentation algorithms. We match the topic that resulted in highest segmentation accuracy as the object, and other topics as background. The segmentation accuracy results are shown in figure 3, from which we could see that both LDA+MRF and LDA+noise model result in higher accuracy than spatial LDA model, and topic random field produces the highest segmentation accuracy. Also, the comparison between LDA+MRF and LDA+noise model shows that the Markov random field defined over latent topic variables improves the accuracy more. It should be noted that our result is not directly comparable to that of the state-of-the-art image segmentation methods, as we did not engineer our image features much. The message here is that spatial consistency and a better model of image feature generation are crucial for the success of aspect models in image analysis.

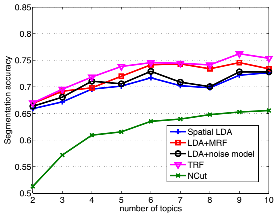


Fig. 3. Segmentation accuracy of normalized cut, spatial LDA, LDA+MRF, LDA+noise, and TRF on the Weizmann horse data set

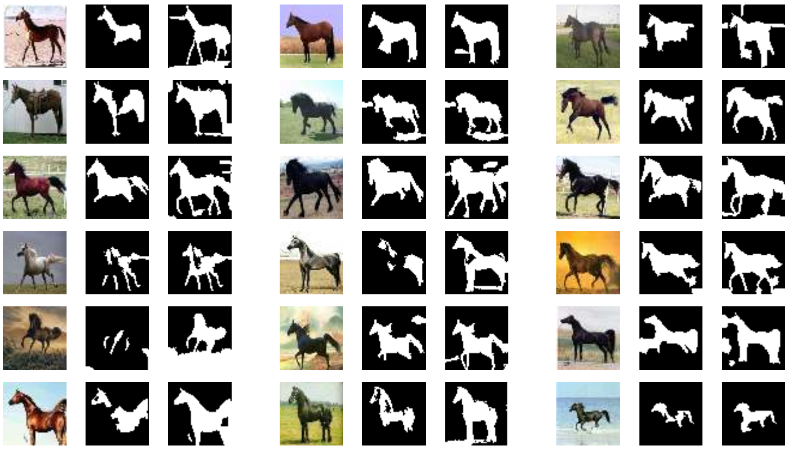


Fig. 4. (Best viewed in color). Segmentation results of horses. From left to right: original image, segmentation result of spatial LDA and TRF. The regions in white are the segmentations of the animals. The regions in black stand for background.

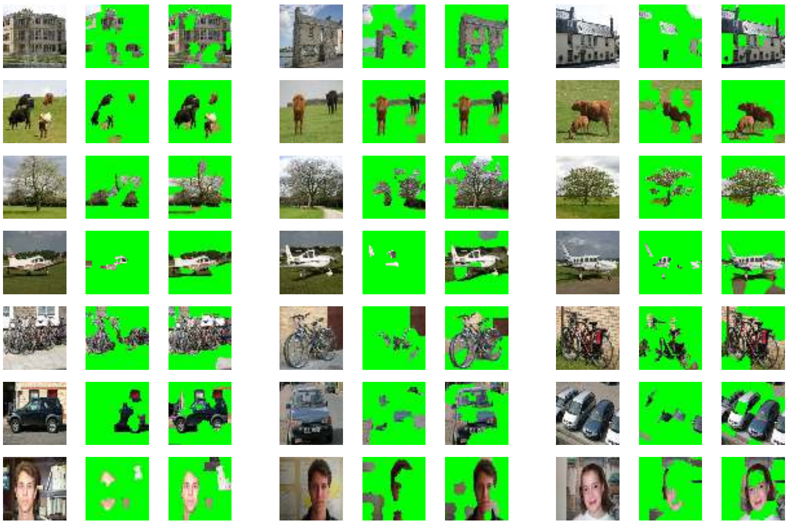


Fig. 5. (Best viewed in color). Segmentation results of the MSRC database. From left to right: original image, segmentation result of spatial LDA and TRF.

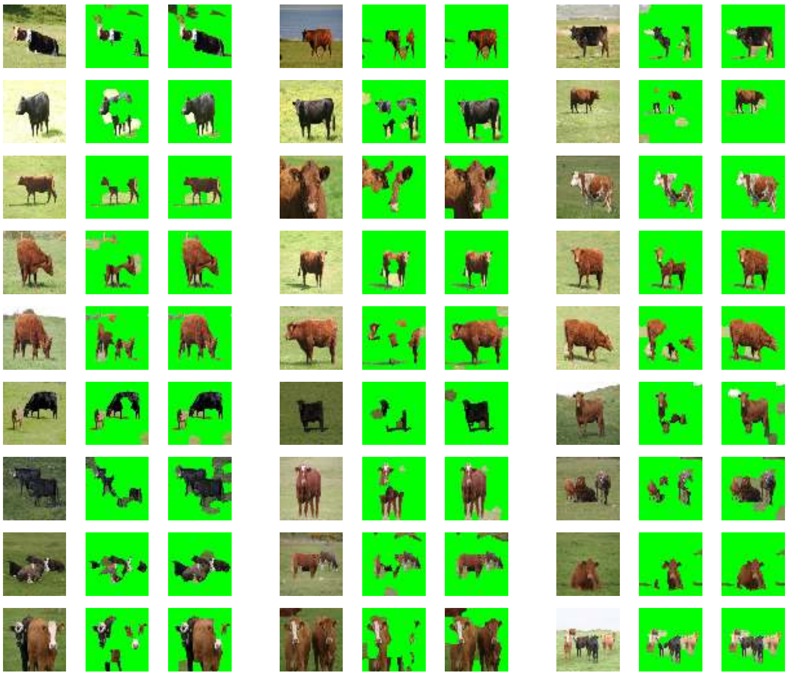


Fig. 6. (Best viewed in color). Segmentation results of cows. From left to right: original image, segmentation result of spatial LDA and TRF.

To better compare the performance of TRF with LDA, we show in figures 4,5,6 the segmentation results on the three data sets, where we have set the number of topics to 4, 12, 4 respectively. From these segmentation results, we could see that one major problem with spatial LDA is that it is more likely to separate parts from the same object into different segments. For example, in the Weizmann horse data, spatial LDA constantly separates the body and legs of a horse into different groups. However, the segmentation results of TRF does not show this phenomenon. Therefore, we argue that enforcing spatial coherence between adjacent regions via MRF avoids separating parts of the same object into different groups. Moreover, from the results on cows data set, we see that spatial LDA is more likely to segment cows with different colors or facing different directions into separate groups. However, by introducing a simple Gaussian noise model for generating image features, TRF is significantly more robust to variability in the instantiations of local features corresponding to the same objects due to variations in lighting, transformation, viewing angle, etc.

6 Conclusions

We propose *Topic Random Field* (TRF) for image segmentation. The TRF model improves over the LDA-style models by defining a Markov Random Field (MRF) over hidden topic assignment of super-pixels in an image to enforce the spatial coherence between neighboring regions, and by employing a noise channel between visual words in the dictionary and instantiated super-pixels in the real image to better model the variance of local features. Empirical studies on three image data sets demonstrate the improvement of our model in image segmentation over the LDA-style model.

Acknowledgments. E. Xing is supported by NSF IIS-0713379, DBI-0546594, Career Award, ONR N000140910758, DARPA NBCH1080007 and Alfred P. Sloan Foundation. L. Fei-Fei is supported by NSF IIS-0845230, Career Award.

References

1. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 22, 888–905 (2000)
2. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 24, 603–619 (2002)
3. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. International Journal of Computer Vision 59, 167–181 (2004)
4. Forsyth, D., Ponce, J.: Computer Vision: A Modern Approach. Prentice Hall Professional Technical Reference (2002)
5. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 42, 177–196 (2001)
6. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)
7. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 524–531 (2005)

8. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google"s image search. In: Proceedings of the Tenth IEEE International Conference on Computer Vision, pp. 1816–1823 (2005)
9. Russell, B., Freeman, W., Efros, A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1605–1614 (2006)
10. Sivic, J., Russell, B., Efros, A., Zisserman, A., Freeman, W.: Discovering object categories in image collections. In: Proceedings of the Tenth International Conference on Computer Vision (2005)
11. Verbeek, J., Triggs, B.: Region classification with markov field aspect models. In: Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
12. Li, L.J., Socher, R., Fei-Fei, L.: Towards total scene understanding:classification, annotation and segmentation in an automatic framework. In: Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2009)
13. Cao, L., Fei-Fei, L.: Spatially coherent latent topic model for concurrent object segmentation and classification. In: Proceedings of IEEE International Conference on Computer Vision (2007)
14. Kadir, T., Brady, M.: Saliency, scale and image description. Int. J. Comput. Vision 45, 83–105 (2001)
15. Sudderth, E., Jordan, M.: Shared segmentation of natural scenes using dependent pitman-yor processes. In: Proceedings of Neural Information Processing Systems (2008)
16. Andreetto, M., Zelnik-Manor, L., Perona, P.: Unsupervised learning of categorical segments in image collections. In: Proceedings of IEEE International Conference on Computer Vision (2008)
17. Levin, A., Weiss, Y.: Learning to combine bottom-up and top-down segmentation. Int. J. Comput. Vision 81, 105–118 (2009)
18. Malisiewicz, T., Efros, A.: Recognition by association via learning per-exemplar distances. In: Proceedings of the 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (2008)
19. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110 (2004)
20. Schwarz, G.: Estimating the dimension of a model. The annals of statistics 6, 461–464 (1978)
21. Wainwright, M., Jordan, M.: Graphical Models, Exponential Families, and Variational Inference. Now Publishers Inc. (2008)
22. Borenstein, E., Ullman, S.: Learning to segment. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3023, pp. 315–328. Springer, Heidelberg (2004)
23. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. In: Proceedings of IEEE International Conference on Computer Vision (2005)