Context Aware Topic Model for Scene Recognition

Zhenxing Niu^b, Gang Hua^b, Xinbo Gao^b, and Qi Tian[‡]
^bXidian University ^bStevens Institute of Technology ^bUniversity of Texas at San Antonio

zhenxingniu@gmail.com ghua@stevens.edu xbgao@mail.xidian.edu.cn qitian@cs.utsa.edu

Abstract

We present a discriminative latent topic model for scene recognition. The capacity of our model is originated from the modeling of two types of visual contexts, i.e., the category specific global spatial layout of different scene elements, and the reinforcement of the visual coherence in uniform local regions. In contrast, most previous methods for scene recognition either only modeled one of these two visual contexts, or just totally ignored both of them. We cast these two coupled visual contexts in a discriminative Latent Dirichlet Allocation framework, namely context aware topic model. Then scene recognition is achieved by Bayesian inference given a target image. Our experiments on several scene recognition benchmarks clearly demonstrated the advantages of the proposed model.

1. Introduction

Visual scene understanding is an active research topic in computer vision. The statistics of the local appearance of an image often provide valuable visual cue for scene recognition. This partly foster the popularity of the bag-of-words (BoW) model, in which local features extracted from an image are first mapped to a set of visual words by vector quantization. An image is then represented as a histogram of visual word occurrences, which naturally encodes the statistics of local features.

The BoW representation largely facilitates the applications of latent topic models invented from document understanding literature for the task of scene recognition, such as probabilistic Latent Semantic Analysis (pLSA) [10] and Latent Dirichlet Allocation (LDA) [3]. For example, Feifei *et al.* [6] exploited LDA for scene category recognition. Sudderth *et al.* [21] presented a hierarchical topic model for part based object and scene category recognition, where the parts can be shared among different visual categories.

Sudderth *et al.*'s model [21] represents a scene image in a three-layer hierarchy, i.e., the top level corresponds to a scene (e.g., "street", "coast", "forest"), the middle level corresponds to a set of scene elements (e.g., "buildings", "sky",

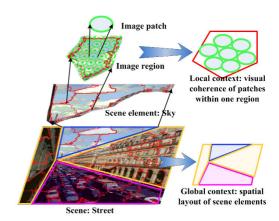


Figure 1. Modeling of four levels of visual information in a scene image: scene (e.g., "street"), scene elements (e.g., "sky", "building", "road"), image regions (shown with red contour), image patches (shown with green circle), where the global context indicates the spatial layout of scene elements, and local context indicates the relationship of image patches within an image region.

"mountain"), and the bottom level corresponds to a set of image features extracted from local image patches. Therefore, it models how a scene image is generated by some scene elements and visual words without the requirement of labeling scene elements.

Though effective, these models are unsupervised in nature, which do not make effective use of the supervised information. This largely motivates recent work on the exploration of supervised topic models [11, 18, 2] for visual recognition. According to different schemes used for leveraging supervised information, supervised topic models can be classified into two categories: *downstream model* and *upstream model* [24].

For downstream model, the supervised response variables are generated from topic assignment variables, such as the sLDA model [2]. For upstream model, the response variables directly or indirectly generate latent topic variables, such as the DiscLDA model [11], which are well-motivated from human vision research. Both downstream and upstream supervised topic model have been utilized for scene recognition in the past. For example, Wang *et al.* [22]

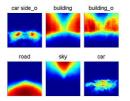
extended sLDA [2] for simultaneous image classification and annotation. Niu *et al.* [18] extended DiscLDA [11] for visual recognition by modeling the spatial arrangement of object parts or scene regions.

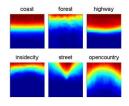
On the other hand, many previous topic models employ the BoW representation directly, which totally ignore the spatial context of the local features. As already revealed in previous works [12, 9, 5, 20, 13, 8, 18, 22], contextual information such as the interaction/relationships among local image features, image regions, and object/scenes provide beneficial information in disambiguating visual words, which often leads to better recognition performance. There are two ways for modeling contextual information: one focuses on modeling the spatial layout of image patches or objects, which is named as global context since the image coordinates of image patches or objects are exploited; the other focuses on modeling the relationship of neighbor image patches or objects, which is named as local context since the distance between image patches or objects is exploited.

To model the global context, the location information of image patches (i.e., local features) is exploited [8, 12, 5, 20, 13, 16, 18]. To list a few, the shape of an object is represented by the mutual position of parts in [8]. Spatial pyramid matching (SPM) [12] and its discriminative variant [9] have demonstrated very good performance on both scene and object recognition tasks. Cao *et al.* [5] generate a series of ordered bag-of-features by projecting local features to different lines or circles. Su *et al.* [20] construct multiple semantic context specific bag-of-words histograms to represent an image. By representing an image in a hierarchical structure. Several previous works [13, 16, 18] have explored to model the contexts of "scene/object"-"scene elements/object parts"-"image patches" for more robust scene recognition or scene parsing.

To model the local context, the relationship of neighbor image patches is exploited [23, 17, 4]. Wang *et al.* [23] propose SLDA for discovering object classes from a collection of images. In SLDA, visual words (e.g., an eye patch and a nose patch), which often occur in the same images and are close in space, are clustered into one topic (e.g., a face). Similar to shape context histogram, a spatial histogram with distance approximately in log scale is exploited for spatial feature in [17]. Cao *et al.* [4] propose a topic model where visual words in a coherent local region share the same topic, which proved to be an very effective local context model.

As pointed out in [18, 16], spatial layout of key scene elements in a specific scene category provides important *global context* for scene understanding. This can be clearly demonstrated by studying the LabelMe natural scene image dataset [16], as shown in Figure 2. As we can observe, different scene category has different global spatial layout of the scene elements (e.g., "car", and "buildings", etc.), and





(a) spatial layout of some scene elements for scene "street"

(b) scene element "sky" for different scene categories

Figure 2. Global context of scene elements. The per-pixel frequency counts of the scene elements are shown by the color of pixels.

the same scene elements (e.g., the sky) also often account for different spatial regions in images from different scene categories.

Intuitively, the global and local spatial contextual information are complementary to each other. Therefore, jointly model both of them would be beneficial for scene understanding. In this paper, we propose such a model, namely context aware topic model which captures both global and local contexts, for scene recognition. To the best of our knowledge, there is few work which had jointly modeled both the global and local contextual information for scene recognition.

As illustrated in Figure 1, our model captures four level of visual information, i.e., a scene image is comprised of some scene elements; each scene element is comprised of some image regions; each region is comprised of some image patches. In the proposed model, the scene elements are represented with latent topic variables, and the assignment of topic is performed at image region level.

In other words, the topic of image region indicates which scene element it comes from. Since topic is assigned to image region, the image patches within one region will share the same topic, which is similar to the model setting of Cao *et al.* [4]. Therefore, our model captures *local spatial context* by reinforcing the visual coherence in uniform local regions.

Moreover, our model captures the category specific spatial layout in a non-parametric fashion, which is more proper than the spatial Gaussian distribution utilized in the S-DiscLDA model [18]. This is justified by the example shown in Figure 2(a). It is obvious that pixels associated with the scene element "building" in the "street" scene appear at two separate areas, which cannot be modeled well by a single Gaussian distribution.

Our model is also very flexible. When the scene-element spatial layout distribution can be obtained off-line, e.g., we can easily build one from the labeled data published by Liu *et al.* [16], we can consume it directly. When it is not available, we can automatically learn it from the data in the training phase. Scene recognition is hence performed by Bayesian inference over the proposed model. Our main

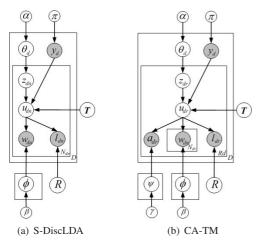


Figure 3. Comparison between S-DiscLDA and CA-TM

contribution is therefore a unified discriminative latent topic model which jointly models both the global and the local spatial context. It leads to a robust scene recognition algorithm that exhibits competitive performance on several scene recognition benchmarks when compared with the state-of-the-art.

2. Context aware topic model

To represent a natural scene image with topic model, we first segment it into regions dr [7] and extract image patches dn within each region; the region visual word a_{dr} is used to represent the appearance of the region, where the corresponding region codebook $\{v_i^r\}_{i=1}^A$ can be obtained offline by vector quantization of region features (e.g., color, texture); the image coordinates l_{dr} of the region center is used to represent the location of region; the patch visual word w_{dn} is used to represent the appearance of each image patch, where the corresponding patch codebook $\{v_i^p\}_{i=1}^W$ can also be obtained offline by vector quantization of patch features such as SIFT.

The graphical model of our proposed context aware topic model (CA-TM) is shown in Figure 3(b). For ease of understanding, we summarize the notations of variables and parameters in the proposed model in Table 1. In the proposed visual representation, scene images are treated as "documents"; scene elements are treated as "topics"; region visual words and patch visual words are treated as "words". We cast our model in a discriminative Latent Dirichlet Allocation framework (i.e., DiscLDA) because of its ability to classification. The generative process of an image with our model can be described as follow:

- 1. For image I_d , draw topic proportions $\theta_d \sim \text{Dir}(\alpha)$;
- 2. For each region dr, select a topic $z_{dr} \sim \text{Multi}(\theta_d)$;
- 3. Given z_{dr} and image label y_d , select another topic u_{dr} through selection matrix T^{y_d} ;
- 4. For each region dr and its corresponding u_{dr} :

Table 1. The notations of variables and distributions.

Notations	Descriptions
d = 1,, D	The index of images.
$dr = 1,, R_d$	The index of image regions.
$dn = 1,, N_{dr}$	The index of image patches within re-
	gion dr .
$y_d = 1,, C$	The scene label of image d .
$u_{dr} = 1,, K_1$	The topic of regions dr (i.e., the scene
	element that image region comes from).
$a_{dr} \in \{v_i^r\}_{i=1}^A$	The visual word of dr (i.e., appearance
	of image region).
$w_{dn} \in \{v_i^p\}_{i=1}^W$	The visual word of dn (i.e., appearance
	of image patch).
$l_{dr} = (x_{dr}, y_{dr})$	The image coordinates of dr (i.e., loca-
	tion of image region).
$p(a_{dr} u_{dr},\Psi)$	the dist. of a_{dr} over region code-
	book $\{v_i^r\}_{i=1}^A$ (i.e., region appearance
	for scene element u_{dr}).
$p(w_{dn} u_{dr},\Phi)$	the dist. of w_{dn} over patch code-
	book $\{v_i^p\}_{i=1}^W$ (i.e., patch appearance for
	scene element u_{dr}).
$p(l_{dr} u_{dr},R)$	the dist. of l_{dr} over image coordinates
	plane (i.e., region location for scene ele-
	ment u_{dr}).

- (a) Draw a region visual word to represent its appearance $a_{dr} \sim \text{Multi}(\Psi)$;
- (b) Draw its location $l_{dr} \sim p(l_{dr}|R, u_{dr})$;
- (c) For each image patch dn within the region dr:
 - i. Draw a patch visual word to represent its appearance $w_{dn} \sim \text{Multi}(\Phi)$

Previously, Niu et al. [18] have presented a S-DiscLDA model for visual recognition, which is also built upon the framework of DiscLDA. For comparison with our proposed model, we present the S-DiscLDA model in Figure 3(a). There are some significant differences between S-DiscLDA and our CA-TM model. Firstly, only the spatial distribution of image patches is encoded by R in S-DiscLDA (i.e., l_{dn} is the location of image patch), while both global and local spatial context are modeled in CA-TM. More specifically, the global context of image region is encoded by R (i.e., l_{dr} is the center of image region); and the local context of image patch within one region is encoded by enforcing them to share the same topic. Secondly, R refers to the parameters of spatial Gaussian distributions in S-DiscLDA, and it refers to the values of a spatial histograms in CA-TM. The latter captures the category specific spatial layout in a non-parametric fashion, which is more proper than Gaussian distribution.

According to Figure 3(b), it is easy to figure out that the joint distribution $p(a_d, l_d, w_d, z_d, u_d)$ given an image I_d in the CA-TM model can be factorized as

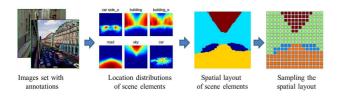


Figure 4. Given global context, we can directly obtain R. Given the spatial layout of scene elements, we get the histograms by sampling the continuous spatial layout.

$$p(\boldsymbol{a_d}, \boldsymbol{l_d}, \boldsymbol{w_d}, \boldsymbol{z_d}, \boldsymbol{u_d} | y_d, \alpha, T, R, \Psi, \Phi)$$

$$= p(\theta_d | \alpha) \prod_{dr=1}^{R_d} \left(p(z_{dr} | \theta_d) p(u_{dr} | z_{dr}, y_d, T) \right)$$

$$\times p(a_{dr} | u_{dr}, \Psi) p(l_{dr} | u_{dr}, R) \prod_{dn=1}^{N_{dr}} p(w_{dn} | u_{dr}, \Phi) \right) \quad (1)$$

3. Model learning

3.1. Problem formation

Given a corpus of scene images with class labels $\{y_d, a_d, w_d, l_d\}_{d=1}^D$, we find the maximum likelihood estimation for the appearance of image patches (w.r.t. Φ), the appearance of image regions (w.r.t. Ψ), and the location distribution of image regions (w.r.t R). More formally, we have

$$\{\Phi^*, R^*, \Psi^*\} = \underset{\Phi, R, \Psi}{\operatorname{argmax}} p(y_d, \boldsymbol{w_d}, \boldsymbol{l_d}, \boldsymbol{a_d} | \Psi, \Phi, R) \quad (2)$$

3.2. CA-TM learning

We propose an iterative algorithm to learn the model parameters, which alternatively estimates the global context and the appearance of scene elements:

- 1. **Initialization-step**: We use many topics to represent a scene and initialize their spatial layout. Then we obtain the appearances of these topics;
- 2. **Iteration-step**: Merge topics and update their spatial layout and appearances:
 - (a) **Context-step**: Given the appearances of topics (i.e., Φ , Ψ), we merge topics with similar appearances, and update their spatial layout.
 - (b) **Appearance-step**: Given the updated spatial layout of topics (i.e., R), we estimate their appearances (i.e., Φ , Ψ) similar to Griffiths and Steyvers's method.

Note that when the global context is available offline (e.g., from the labeled pixels in the LabelMe dataset), we can directly obtain R according to the spatial prior as shown in Figure 4. Then for learning the proposed model, we only need to estimate (Φ, Ψ) by running the Appearance-step.

3.2.1 Learning global context

In the Context-step, we should learn the global context from training data, i.e., learn the spatial layout of scene elements. Since the appearances and spatial layout are learned alternatively, the Initialization-step is necessitated to provide a good start point. For that, we adopt a scheme by merging topics and updating their appearances and spatial layout. Specifically, in the Initialization-step, we densely divide the image into blocks, and each block is initially assigned with a topic. So the number of topics is the same as the number of blocks. Because of the dense partitioning, one scene element may cover several blocks, and thus is represented by several topics. For learning global context, we expect one scene element to be represented by only one topic. So we need to merge topics corresponding to the same scene elements in the Iteration-step.

It is obvious that the topics corresponding to the same scene element usually have similar appearances. Therefore we merge similar topics and use a new topic to represent all of them. The new topic will cover all the cells that was covered by the merged topics. By doing so, the number of topics will decrease, and the spatial layout of topics will be updated. In practice, the appearance of topic k is measured by $\psi(k)$, and the similarity between two topics is hence measured by their Euclidean distance. This way, we can merge similar topics, and only use one topic to represent one scene element.

3.2.2 Learning the appearance

The key step for learning appearance of topics (i.e. Φ and Ψ) is the evaluation of $p(\boldsymbol{z_{dr}}, \boldsymbol{u_{dr}}|\boldsymbol{a_d}, \boldsymbol{l_d}, \boldsymbol{w_d}, y_d, R)$. We resort to Gibbs sampling to estimate this posterior distribution, i.e., we generate samples from

$$p(z_{dr}, u_{dr} | \mathbf{z}_{-dr}, \mathbf{u}_{-dr}, \mathbf{a}_{d}, \mathbf{l}_{d}, \mathbf{w}_{d}, y_{d}, R)$$

$$= \frac{p(\mathbf{z}_{d})}{p(\mathbf{z}_{-dr})} \cdot T^{y}_{u_{dr}z_{dr}} \cdot \frac{p(\mathbf{a}_{d} | \mathbf{u}_{d})}{p(\mathbf{a}_{-dr} | \mathbf{u}_{-dr})} \cdot p(l_{dr} | u_{dr}, R) \cdot \frac{p(\mathbf{w}_{d} | \mathbf{u}_{d}, \beta)}{p(\mathbf{w}_{-dr} | \mathbf{u}_{-dr}, \beta)}$$
(3)

All the items except the last one in Equation 3 are similar to standard LDA model, so we present its detailed form in Equation 4, i.e.,

$$\frac{p(\boldsymbol{w_d}|\boldsymbol{u_d},\beta)}{p(\boldsymbol{w_{-dr}}|\boldsymbol{u_{-dr}},\beta)} = \frac{\prod\limits_{v \in \boldsymbol{w_{dr}}} P_{m(v,u_{dr})+\beta-1}^{q(dr,v)}}{P_{m(u_{dr})+\beta-1}^{N_{dr}}}$$
(4)

where q(dr, v) stands for the number of image patches with visual word v and within region dr. $m(v, u_{dr})$ stands for the number of times that topic u_{dr} is assigned to visual word v,

and $m(u_{dr})$ stands for the number of times that topic u_{dr} is assigned to words. P stands for P-permutation.

With the samples from the posterior distribution, we can estimate Φ and Ψ . The estimation of Ψ is the same as in standard LDA, we only give the estimation of Φ here. Since $p(\phi_k|\boldsymbol{u},\boldsymbol{w}) \sim \text{Dir}(\{\beta+m(v,k)\}_{v=1}^W)$, we can estimate Φ by normalizing the Dirichlet paraments as

$$\phi_k(v) = \frac{m(v,k) + \beta}{m(k) + W\beta} \tag{5}$$

4. Inference and recognition

4.1. Problem formation

To recognize the scene category of an input image, we first represent it with visual words (i.e., a_{dr}, w_{dn}) and overly segment the image to extract region locations (i.e., l_{dr}). With model parameters (i.e., Ψ, Φ, R) obtained in learning phase, we can recognize the scene category of the input image by maximum a posteriori (MAP) estimation of the image label

$$y^* = \underset{y_d \in \{1, 2, \dots, C\}}{\operatorname{argmax}} \left(p(y_d | \boldsymbol{w_d}, \boldsymbol{l_d}, \boldsymbol{a_d}, T, \Psi, \Phi, R) \right)$$
 (6)

4.2. Inference via bridge sampling

We leverage bridge sampling to estimate the maximum a posterior. Specifically, denote $q_c(\mathbf{z_d}) = p(\mathbf{w_d}, \mathbf{l_d}, \mathbf{a_d}, \mathbf{z_d} | y_c)$, we have

$$\frac{p(y_c|\boldsymbol{w_d}, \boldsymbol{l_d}, \boldsymbol{a_d})}{p(y_1|\boldsymbol{w_d}, \boldsymbol{l_d}, \boldsymbol{a_d})} = \frac{p(y_c)}{p(y_1)} \frac{Z_c}{Z_1}$$
where $Z_c = \int q_c(\boldsymbol{z_d}) d\boldsymbol{z_d}$ (7)

So instead of directly estimating the posterior probability, we can estimate the ratio $Z_c/Z_1, c=2,3,...,C$ using bridge sampling. Specifically, we need to sample $\boldsymbol{z_d^{(i)c}}$ from $p(\boldsymbol{z_d}|y_c,\boldsymbol{w_d},l_d,\boldsymbol{a_d})$ and estimate the ratio as

$$\frac{Z_c}{Z_1} \approx \frac{\sum_{i=1}^{M} h_{c1}(\boldsymbol{z_d^{(i)1}})}{\sum_{i=1}^{M} h_{1c}(\boldsymbol{z_d^{(i)c}})}$$
where $h_{ab} = \sqrt{\prod_{dr} \frac{p(\boldsymbol{w_{dr}}, l_{dr}, a_{dr}|z_{dr}, y_a)}{p(\boldsymbol{w_{dr}}, l_{dr}, a_{dr}|z_{dr}, y_b)}}$
(8)

where

$$p(\mathbf{w}_{dr}, l_{dr}, a_{dr} | z_{dr}, y)$$

$$= \sum_{u_{dr}} \left[T_{u_{dr}z_{dr}}^{y} \Psi_{a_{dr}u_{dr}} p(l_{dr} | u_{dr}, R) \prod_{dn=1}^{N_{dr}} \Phi_{w_{dn}u_{dr}} \right]$$
(9)

To sample $z_d^{(i)c}$, we resort to Gibbs sampling to sample from the conditional probability

$$p(z_{dr}|\mathbf{z}_{-dr}, \mathbf{a}_{d}, \mathbf{l}_{d}, \mathbf{w}_{d}, y_{d})$$

$$= \frac{p(\mathbf{z}_{d})}{p(\mathbf{z}_{-dr})} \cdot \frac{p(\mathbf{a}_{d}|\mathbf{z}_{d}, y_{d})}{p(\mathbf{a}_{-dr}|\mathbf{z}_{-dr}, y_{d})} \cdot \frac{p(\mathbf{l}_{d}|\mathbf{z}_{d}, y_{d})}{p(\mathbf{l}_{-dr}|\mathbf{z}_{-dr}, y_{d})} \cdot \frac{p(\mathbf{w}_{d}|\mathbf{z}_{d}, y_{d})}{p(\mathbf{w}_{-dr}|\mathbf{z}_{-dr}, y_{d})}$$
(10)

All the factors except the last one in Equation 10 are similar to standard LDA model. Its exact form is

$$\frac{p(\mathbf{w_d}|\mathbf{z_d}, y_d)}{p(\mathbf{w_{-dr}}|\mathbf{z_{-dr}}, y_d)} = \sum_{u_{dr}} \left[T_{u_{dr}z_{dr}}^{y_d} \prod_{dn=1}^{N_{dr}} \Phi_{w_{dn}u_{dr}} \right].$$
(11)

5. Experiments

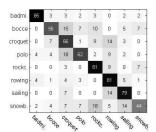
We evaluate the CA-TM model on three scene datasets, ranging from generic natural scene images (Scene 15 and LabelMe 8-class scene dataset) to complex event and activity images (UIUC-Sports). Scene classification performance is evaluated by averaging multi-way classification accuracy over all scene classes in each dataset. We list below the experimental setting for each dataset:

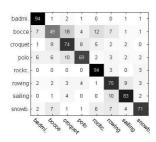
- Scene 15: This is a dataset of 15 natural scene classes, which contains 4482 gray-scale images with different size. We normalize them in size of 256 × 256 pixels, and use 100 images in each class for training and rest for testing. The gray histogram features are extracted in each image region [14].
- LabelMe: This is a dataset of 8 natural scene classes, which contains 2688 color images of the same size (256 × 256).
- UIUC-Sports: This is a dataset of 8 complex event classes, which contains 1574 color images with different sizes. We normalize them in size of 250 × 250 pixels.

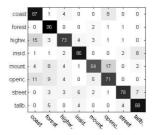
For the last two datasets, we randomly split them into training and testing datasets with equal size; The color histogram features are extracted in each image region [14]. For all dataset, we vector quantize region features into a region codebook of size 100. Each image is first over-segmented into 90 small coherent regions by using the code provided by Felzenszwalb *et al.* [7]. We then discard small regions which are smaller than certain size. We then densely extract SIFT features from 10×10 image patches within each region. These SIFT features are quantized to form an image patch codebook of size 500.

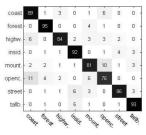
5.1. Scene classification

Figure 6 summarizes the scene classification results of our CA-TM model and other state-of-the-art methods.









(a) UIUC-Sports: 'DiscLDA+GC' avg. accuracy: 70%

(b) UIUC-Sports: CA-TM. avg. accuracy: 78%

(c) LabelMe: 'DiscLDA+GC'. avg. accuracy: 81%

(d) LabelMe: CA-TM. avg. accuracy: 87%

Figure 5. Comparisons using confusion matrix. (a) 'DiscLDA+GC' on the UIUC-Sports dataset. (b) CA-TM on the UIUC-Sports dataset. (c) 'DiscLDA+GC' on the LabelMe dataset. (d) CA-TM on the LabelMe dataset.

These methods can be categorized into two classes: 1) methods based on unsupervised topic model; 2) methods based on supervised topic model. For the first category, topic model usually is used as a generative model for classification, or is combined with an off-the-shelf classifiers such as liner SVM. In Fei-Fei and Perona's method [6] (i.e., the 1st bar), each scene category is modeled as one LDA, and an input image is classified by selecting the model which maximizes the likelihood function. Cao et al. [4](i.e., the 2nd bar) and Li et al. [14](i.e., the 4th bar) extend the model of [6] by encoding local spatial context. Li and Fei-Fei [13](i.e., the 9th bar) further improve the performance by jointly categorizing of events, scenes, and objects. Li et al. [15] (i.e., the 10th bar) represent image as a response map of Object Bank (i.e., a large number of pretrained generic object detectors), and use linear SVM classifier for scene recognition, which achieved the state-of-theart results on UIUC-Sports dataset.

For the second category, supervised topic model can recognize scene image directly. The Corr-LDA [1](i.e., the 3rd bar) and sLDA [22] (i.e., the 5th bar) can be used for scene classification. S-DiscLDA [18] (i.e., the 6th bar) presented a part based model for visual recognition. Zhu $et\ al.$ [24] (i.e., the 8th bar) proposed a joint max-margin and max-likelihood learning method as a upstream supervised topic model for scene understanding.

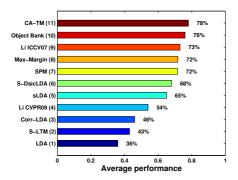


Figure 6. Comparison of classification results on UIUC-Sports dataset.

As shown in Figure 6, in general the supervised topic models achieve better performance than unsupervised ones. And upstream supervised models achieve better performance than downstream supervised ones. In particular, the proposed CA-TM model performed the best when compared with all other algorithms summarized above.

The comparison of classification results on Scene 15 dataset is shown in Figure 7. The proposed CA-TM model performed better than BoW [12], SPM [12], and Object Bank [15], and is only inferior to a recent work which leveraged semantic context trained from additional data [20]. By using semantic contexts that trained from other datasets (e.g., PASCAL 2007), Su et al. [20] constructed multiple semantic context specific BoW histogram (i.e., C-BoW), and achieve the best results on Scene 15 dataset. In this sense, Su et al.'s model [20] leveraged additional knowledge from external dataset, while our proposed model did not. Besides, the semantic context they modeled should be complementary to the local and global spatial contexts we cast into our model. Hence we expect that combining their approach with ours will give further improvement in scene recognition accuracy, which is part of our future exploration.

5.2. Influence of global and local context

In CA-TM, the global and local contextual information were jointly modeled in a supervised topic model. So we should explore how the recognition performance is influenced by them. We demonstrate it by comparing

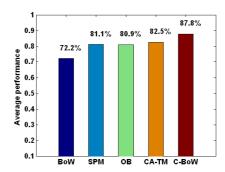


Figure 7. Comparison of classification results on Scene 15 dataset.

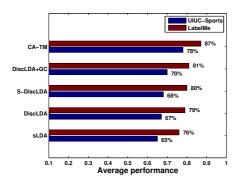


Figure 8. Comparison of classification results between *DiscLDA*, '*DiscLDA+GC*' and *CA-TM* models on both UIUC-Sports and LabelMe dataset.

three models on both UIUC-Sports and LabelMe datasets, which are 1) DiscLDA without any contextual information; 2) modeling only the global context into DiscLDA ('DiscLDA+GC'); and 3) jointly modeling both the global and the local context into DiscLDA ('CA-TM').

As shown in Figure 8, the modeling of global context can improve the performance from 67% to 70% on UIUC-Sports dataset, and from 79% to 81% on LabelMe dataset; jointly modeling both the global and the local context can further improve the performance to 78% on UIUC-Sports dataset, and to 87% on LabelMe dataset. The confusion matrix of DiscLDA+GC and CA-TM on these two datasets are shown in Figure 5. Moreover, by comparing S-DiscLDA [18] and DiscLDA+GC, it is obvious that encoding global context in a non-parametric fashion is better than Gaussian distribution.

5.3. Influence of topic number

When modeling scene with topic model, we assume that each topic corresponds a scene element. So it is expected that we can discover and represent more scene elements with more topics. However, the performance of recognition may not be improved monotonically when the number of topics is increased [6]. Indeed, the performance may be relatively independent of the number of topics for some dataset [19]. Thus, there is a upper bound for recognition performance, which cannot be improved by simply increasing the number of topics.

In this experiment, we will demonstrate that the bound can be significantly improved by encoding global and local contextual information. In Figure 10, we demonstrate how the performance will be improved for the three models, i.e., *DiscLDA*, '*DiscLDA+GC*' and *CA-TM*. We can see that encoding global context can improve the performance; furthermore, jointly encoding global and local context can significantly and consistently improve the performance. It is noted that contextual information can improve the performance as the number of topics increasing, but there is

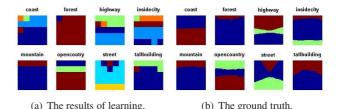


Figure 9. Learning global context, i.e., spatial layout of scene elements, on LabelMe dataset.

still a saturation point in recognition performance. As the red line shows in Figure 10, the performance saturates when the number of topics is larger than certain value.

5.4. Learning global context

In the CA-TM model, the global and local contextual information are jointly encoded into a discriminative topic model. However, the global context, i.e., spatial layout of scene elements, is usually not available as a prior. So we will need to learn it from the training data in the training phase.

We illustrate the results of learning spatial layout on the LabelMe dataset in Figure 9. Since the scene element in LabelMe dataset has already been labeled manually, as shown in Figure 11, we can easily obtain the exact spatial layout of the different scene elments for the LabelMe dataset, which serves as the ground truth for evaluating the estimation from our learning algorithm, as shown in Figure 9(b). Comparing Figure 9(a) and Figure 9(b), we can see that the proposed algorithm largely and correctly estimates the spatial layout of the different scene elements.

6. Conclusion and future work

In this paper, we present a context aware discriminative latent topic model for scene recognition. Global and local spatial context were jointly modeled in a discriminative Latent Dirichlet Allocation framework. Our extensive experiments on several scene recognition benchmarks clearly validate our hypothesis that both global and local spatial con-

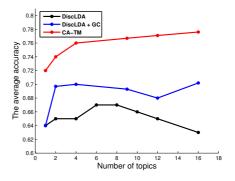


Figure 10. Accuracy as a function of the number of topics on the UIUC-Sports dataset for *DiscLDA*, '*DiscLDA+GC*' and *CA-TM*.

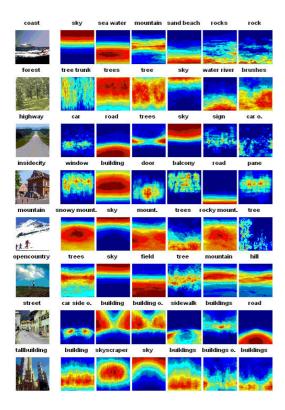


Figure 11. Spatial layout of scene elements for 8 scene categories in LabelMe dataset. The per-pixel frequency counts of the scene elements are shown by the color of pixels. It is the global contextual information for LabelMe dataset. Compared with prior in [16], this prior is category specific spatial layout of scene elements.

text are useful in improving the performance scene recognition, and they are complementary to each other. Our future work will explore means of extending the proposed CA-TM model for joint scene recognition and decomposition. We will also try to extend the model temporally for video scene decomposition.

7. Acknowledgement

This work is supported by the NSFC (Grant Nos. 61125204, 61172146, 60832005), the Fundamental Research Funds for the Central Universities, and the Ph.D. Programs Foundation of Ministry of Education of China (Grant No. 20090203110002). This work was supported in part to Dr. Qi Tian by ARO grant W911BF-12-1-0057, NSF IIS 1052851, Faculty Research Awards by Google, FXPAL, and NEC Laboratories of America, respectively. This work was supported in part to Dr. Gang Hua by start-up funds from Stevens Institute of Technology.

References

[1] D. Blei and M. Jordan. Modeling annotated data. In *SIGIR*, 2003

- [2] D. M. Blei and J. D. McAuliffe. Supervised topic models. In NIPS, 2007.
- [3] D. M. Blei, A. Ng, and M. I. Jordan. Latent dirichlet allocation. In *JMLR*, 2003.
- [4] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *ICCV*, 2007.
- [5] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang. Spatial-bag-of-features. In *CVPR*, 2010.
- [6] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In CVPR, 2005.
- [7] P. Felzenszwalb and D. Huttenlocher. Efficient graph-based image segmentation. In *IJCV*, 2004.
- [8] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In CVPR, 2003.
- [9] T. Harada, Y. Ushiku, Y. Yamashita, and Y. Kuniyoshi. Discriminative spatial pyramid. In CVPR, 2011.
- [10] T. Hofmann. Probabilistic latent semantic indexing. In SI-GIR, 1999.
- [11] S. L. Julien, F. Sha, and M. I. Jordan. DiscLDA: Discriminative learning for dimensionality reduction and classification. In NIPS, 2008.
- [12] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In CVPR, 2006.
- [13] L. Li and L. Fei-Fei. What, where and who? classifying event by scene and object recognition. In *ICCV*, 2007.
- [14] L. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In CVPR, 2009.
- [15] L. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature. In NIPS, 2010.
- [16] C. Liu, J. Yuen, and A. Torralba. Nonparametric scene parsing: label transfer via dense scene alignment. In CVPR, 2009
- [17] D. Liu, G. Hua, P. A. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *CVPR*, 2008.
- [18] Z. Niu, G. Hua, X. Gao, and Q.Tian. Spatial-DiscLDA for visual recognition. In CVPR, 2011.
- [19] P. Quelhas, F. Monay, J. M. Odobez, D. Gatica-Perez, T. Tuyelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *ICCV*, 2005.
- [20] Y. Su and F. Jurie. Visual word disambiguation by semantic contexts. In *ICCV*, 2011.
- [21] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005.
- [22] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In CVPR, 2009.
- [23] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In NIPS, 2007.
- [24] J. Zhu, L. Li, L. Fei-Fei, and E. P. Xing. Large margin learning of upstream scene understanding models. In NIPS, 2010.