Scene Recognition by Jointly Modeling Latent Topics

Shaohua Wan J.K. Aggarwal
The University of Texas at Austin
shaohuawan@utexas.edu aggarwaljk@mail.utexas.edu

Abstract

We present a new topic model, named supervised Mixed Membership Stochastic Block Model, to recognize scene categories. In contrast to previous topic model based scene recognition, its key advantage originates from the joint modeling of the latent topics of adjacent visual words to promote the visual coherency of the latent topics. To ensure that an image is only a sparse mixture of latent topics, we use a Gini impurity based regularizer to control the freedom of a visual word taking different latent topics. We further show that the proposed model can be easily extended to incorporate the global spatial layout of the latent topics. Combined together, latent topic coherency and sparsity can rule out unlikely combinations of latent topics and guide classifier to produce more semantically meaningful interpretation of the scene. The model parameters are learned using Gibbs sampling algorithm, and the model is evaluated on three datasets, i.e. Scene-15, LabelMe, and UIUC-Sports. Experimental results demonstrate the superiority of our method over other related methods.

1. Introduction

Automatic image scene categorization has become more and more important with the ever increasing amount of images that are stored and processed digitally. The *Bag-of-visual-words* model (BoW), which represents an image as an orderless collection of local features (e.g. SIFT [14]), has demonstrated impressve recognition performance ([6, 19, 22]) because they can be reliably detected and matched across objects or scenes under different viewpoints or lighting conditions.

More recently, the success of topic models, such as probabilistic Latent Semantic Analysis (pLSA) [7] and Latent Dirichlet Allocation (LDA) [3], which originate from statistical natural language processing, has motivated researchers to apply them to visual recognition tasks. These methods typically represent an image in a three-layer hierarchy, i.e., the bottom level corresponds to a scene (e.g., street, coast, forest), the middle level corresponds to a set of scene topics (e.g., buildings, ground, sea), and the top level corresponds to a set of image features such as SIFT. The advantage of modeling an image with a three-layer hierarchy is the automatic discovery of semantic scene elements without the requirement of manual labeling them.

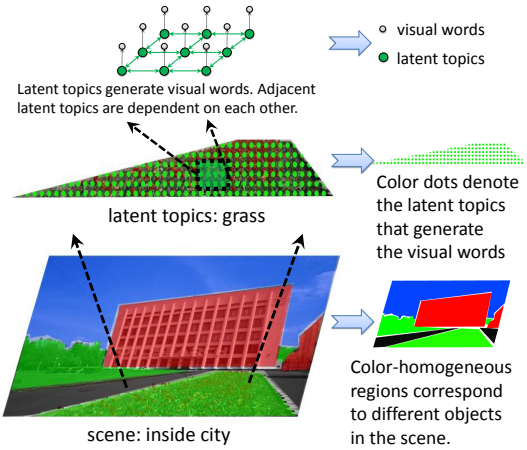


Figure 1: The three-layer hierarchy of our supervised Mixed Membership Stochastic Block Model for scene recognition. Different from previous topic models, we form a network of latent topics in which each latent topic is correlated with its neighbors, making our representation context aware.

Though effective, these methods ignore the semantic coherency of the latent topics and take a simplified assumption that each latent topic is independently generated by the scene. This largely motivates recent work on context-aware topic models for scene recognition [4, 16, 21].

In this paper, we propose a new topic model for scene recognition based on Mixed Membership Stochastic Model (MMSB) [1]. The key advantage of our method is the achievement of topic coherency and sparsity by jointly generating the latent topics of adja-

cent visual words while imposing an impurity regularizer on the latent topics. The scene category is recognized by fitting a softmax classifier to the discovered latent topics. Combined together, topic coherency and sparsity helps rule out unlikely combinations of latent variables and guide the classifier to produce accurate scene recognition results.

Figure 1 illustrates our approach, which uses a three-layer hierarchy to model an image. However, different from previous topic models, the latent topics of adjacent visual words are considered to be interdependent on each other. Later we will show how to exploit this dependency to encourage the semantic coherency and sparsity of the latent topics.

The rest of this paper is structured as follows. Section 2 briefly reviews related work on scene recognition and motivates our method. Section 3 formulates the proposed model in detail and develops an efficient inference algorithm for parameter estimation. Experimental setup and results are given in Section 4 and 5 respectively. Section 6 concludes this paper.

2. Related Work & Motivation

Scene recognition using visual words has been widely studied. Built upon a BoW representation of scene images, [6, 19, 22] directly train classifiers to recognize the scene category. [10] incorporates spatial information into the BoW representation of images for better scene recognition performance. [23, 24] take into account the collocation patterns of visual words and learn "visual phrases" for high-level scene/object recognition. [15] builds a compact codebook for pairs of spatially close SIFT descriptors for visual recognition tasks.

In contrast to the above-mentioned methods which infer scene categories directly from the low-level features, topic models take a hierarchical Bayesian view of the generation of images [5, 16, 17, 20]. In the following, we mainly review related work on topic model based scene recognition since they are more relevant to our work.

Topic models assume the existence of intermediate-level topics in between low-level features and scene categories. The low-level features are considered as generated from the scene topics. And the scene topics are considered as generated from the scene. The scene category is determined as the one most likely capturing its correlation with the latent topics. Therefore, to derive semantically meaningful topics for visual words becomes crucial to recognize scenes.

For example, to promote *local spatial homogeneity*, which requires spatially close visual words to have identical topic, Cao et al. [4] propose to oversegment an image into regions of homogeneous appearances. Only one single topic is assigned to the visual words within each region. Wang et al. [21] propose a spatial LDA model in which visual words that correspond to the same object are clustered into the same topic.

Based on the idea that scene elements are located within an image according to some *global spatial layout*, e.g., "sky" or "cloud" is more likely to be found in the top part of an image than in the bottom, Niu et al. [16] learn a global layout map from manually labeled training data for more accurate sampling of the latent topics. Similarly, Niu et al. [17] incorporate location information into DiscLDA [9] for visual recognition by modeling the spatial arrangement of scene regions.

Although simple and generally performing well, these methods rely on either the underlying image segmentation algorithm for local spatial homogeneity or manual scene elements labeling for building global layout map, making them brittle and sensitive to the visual content of images. Furthermore, while previous topic models capture ambiguous visual word senses by permitting a visual word to be generated from different latent topics, it is often seen that an image is only a sparse mixture of scene elements. For example, in a "coast" scene, one may never find scene elements such as "car" or "tall building". Therefore, without restriction of the freedom of visual words taking excessive, noisy latent topics, one may end up with a overfitted topic model and have the recognition performance negatively affected.

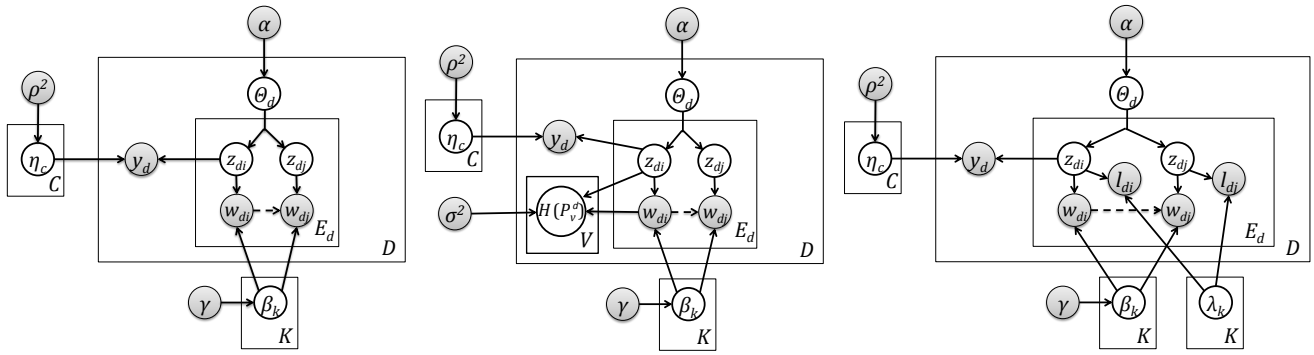
By addressing the above difficulties, the advantages of our method are three-folds:

- Local spatial homogeneity of latent topics is inherently built into our model by jointly generating adjacent latent topics.
- Our model can be easily extended to learn the global spatial layout of the latent topics by adding additional functional nodes without the requirement of manually localizing them in an image.
- We propose a measure of sparsity of the latent topics using Gini impurity. Based on that, a regularizer is imposed to encourage sparse latent topics.

3. Model Description

Supervised MMSB

The basic form of our model, termed as supervised Mixed Membership Stochastic Model (sMMSB), is shown in Figure 2a. For image d from the image corpus $\{1, \cdots, D\}$, we first extract a set of visual words W_d from a dense regular grid with a spacing of 10 pixels to represent the visual appearance of the image. The codebook V of all possible visual words is obtained by vector quantization of SIFT features computed over



(a) sMMSB (b) rsMMSB (c) sMMSB with global spatial layout

Figure 2: The plate diagram of three topic models proposed in this paper. Nodes represent random variables and edges indicate dependencies. The variable at the right lower corner of each plate denotes the number of replications. Shaded nodes indicate observed variables.

the dense grid[1]. We further assume the existence of K latent topics which generate the visual words with different probability. We assume an interdependency between the adjacent latent topics in the grid (as shown in Figure 1). In the following, we use E_d, Z_d and $|V|$ to denote the set of all pairs of adjacent visual words in image d, the set of all pairs of adjacent latent topics in image d, and the total number of visual words in the codebook V, respectively.

Next, we describe how to infer the topics of the visual words in an image, and how to classify the scene category based on these topics. It is easier to understand this model by going through the generative process of an image.

For image $d \in \{1, \cdots, D\}$, we begin by randomly choosing a matrix $\Theta_d \in R^{K \times K}$, which satisfies

$$\sum_{k,l=1}^{K} \Theta_{d\langle k,l\rangle} = 1$$

where the $\langle k, l\rangle^{th}$ entry of Θ_d, $\Theta_{d\langle k,l\rangle}$, is the probability of generating a pair of adjacent topics $\langle k, l\rangle$. Also, for each topic $k \in \{1, \cdots, K\}$, we randomly choosing a vector $\beta_k \in R^{|V|}$, which satisfies

$$\sum_{i=1}^{|V|} \beta_{ki} = 1$$

where the i^{th} entry of β_k, β_{ki}, is the probability of generating the i^{th} visual word w_i in V from topic k.

Having chosen Θ_d and β_k, the probability of observing a pair of adjacent visual words $\langle w_{di}, w_{dj}\rangle$ can be calculated by first generating a pair of adjacent topics $\langle z_{di}, z_{dj}\rangle$ with probability $\Theta_{d\langle z_{di}, z_{dj}\rangle}$, then generating w_{di} and w_{dj} with probability $\beta_{z_{di} w_{di}}$ and $\beta_{z_{dj} w_{dj}}$ respectively.

By maximizing the likelihood of the visual words in the image, we can find the most probable latent topics for generating these visual words. The scene category is recognized using a softmax classifier with the frequency vector of the latent topics as input.

Formally, the generative process of the images in the corpus under supervised MMSB can be stated as follows

1. For each topic $k \in \{1, \ldots, K\}$, draw $\beta_k \sim$ Dir(γ), where Dir(γ) is a Dirichlet distribution with parameter γ.

2. For each scene class $c \in \{1, \ldots, C\}$, draw $\eta_c \sim$ N$(0, \rho^2)$, where η_c are the softmax regression coefficients of class c, and N$(0, \rho^2)$ is a normal distribution with mean 0 and variance ρ^2.

3. For each image $d \in \{1, \ldots, D\}$, draw $\Theta_d \sim$ Dir(α), where Dir(α) is a Dirichlet distribution with parameter α.

4. For each pair of adjacent visual words $\langle w_{di}, w_{dj}\rangle \in E_d$

   (a) Draw topic pair $\langle z_{di}, z_{dj}\rangle \sim$ Multi (Θ_d).

   (b) Draw $w_{di} \sim$ Multi$(\beta_{z_{di}})$ and $w_{dj} \sim$ Multi$(\beta_{z_{dj}})$, respectively.

   where Multi$(\cdot)$ denotes a multinomial distribution.

5. For each image $d \in \{1, \ldots, D\}$, draw its scene label $y_d \sim$ softmax $(\overline{z_d}, \eta)$, where $\overline{z_d} = \frac{1}{2|Z_d|} \sum_{\langle z_{di}, z_{dj}\rangle \in Z_d} (z_{di} + z_{dj})$ is the normalized

[1]SIFT descriptors extracted from a dense grid instead of interest points is able to capture uniform regions such as sky, calm water, or grass.

topic frequency histogram [2] of image d, and the softmax function provides the following distribution over the class variable $p(c|\overline{z_d}, \eta) = \frac{\exp(\overline{z_d} \cdot \eta_c)}{\sum_{c'} \exp(\overline{z_d} \cdot \eta_{c'})}$.

Supervised MMSB under Regularization (rsMMSB)

sMMSB allows the visual words of an image to take on different roles. That is, multiple instances of a visual word $v \in V$ may be found at different locations in image d, each of which may have been generated by different topics. While this freedom is essential to a flexible model, it is often necessary to restrict this freedom so that the image is only a sparse mixture of the latent topics. To this end, we first calculate the normalized topic frequency for visual word v in image d as

$$P_v^d(k) = n_{kv}^d / (\sum_{k'=1}^{K} n_{k'v}^d), k = \{1, \ldots, K\} \quad (1)$$

where n_{kv}^d is the count of the number of times v is generated from topic k in image d. To express our preference for a low degree of mixed topic membership, we enforce a Gini impurity based regularizer on supervised MMSB, which is calculated as

$$H(P_v^d) = \sum_{k=1}^{K} P_v^d(k)(1 - P_v^d(k)) = 1 - \sum_{k=1}^{K} (P_v^d(k))^2 \quad (2)$$

$H(P_v^d)$ tends to 0 as the topics a visual word v in image d can assume become sparse.

The plate diagram of supervised MMSB under regularization is shown in Figure 2b. In this model, $H(P_v^d)$ follows a Gaussian distribution with mean 0 and variance σ^2. This amounts to penalizing large Gini impurity in the latent topic distribution, with σ^2 dictating the strictness of the penalty. As the variance tends to infinity, the model reduces to a fully unregularized mixed membership model. We point out that other sparsity measure such as the entropy of the topic distribution can also be used to regularize the model [2].

Learning the global spatial layout

It is a common observation that an image is comprised of elements at different global locations. The idea of learning the global spatial layout can be realized by statistically localizing the visual words of an image according to their latent topics. Assume the location of the visual words generated by topic k is dictated by a Gaussian distribution with parameter $\lambda_k = (\nu_k, \xi_k^2)$, where ν_k is the mean, and ξ_k^2 is the variance. Then, the global spatial layout can be incorporated into the basic supervised MMSB model as shown in Figure 2c. As can be seen, l_{di}, the location of w_{di}, is drawn from $P(l_{di}|z_{di}, \lambda)$, i.e. $l_{di} \sim N(\lambda_{z_{di}})$. In the following, we stick to supervised MMSB without modeling global spatial layout for ease of presentation.

[2] In computation of the normalized topic frequency histogram, the topic is represented as a K-dimensional binary vector, with the index of 1 denoting the topic. For example, $[1, 0, \ldots, 0]$ indicates the first topic.

3.1. Parameter Learning

Given the hyperparameters $\alpha, \gamma, \sigma^2, \rho^2$ (predetermined by linear search), the joint distribution of the topic pairs $Z = \{Z_d\}_{d=1}^D$, visual word pairs $E = \{E_d\}_{d=1}^D$, topic pair distribution parameter $\Theta = \{\Theta_d\}_{d=1}^D$, word distribution parameter $\beta = \{\beta_k\}_{k=1}^K$, scene labels $y = \{y_d\}_{d=1}^D$, softmax regression coefficients $c = \{c_k\}_{k=1}^K$, and Gini impurity of visual words $H(P_V)$ under the rsMMSB model can be decomposed as

$$\begin{aligned} &P(Z, E, \Theta, \beta, y, \eta, H(P_V)|\alpha, \gamma, \sigma^2, \rho^2) \\ &= P(\Theta|\alpha)P(Z|\Theta) \times P(\beta|\gamma)P(E|Z, \beta) \\ &\quad \times P(y|\overline{z}, \eta)P(\eta|\rho^2) \times P(H(P_V)|\sigma^2) \end{aligned} \quad (3)$$

Since the exact inference in the model is intractable, we use a collapsed Gibbs sampler for approximate inference, where Θ and β are integrated out. Assuming that Dir(α) and Dir(γ) are symmetric, i.e. $\alpha_{k,l} = \alpha_0, \forall k, l \in \{1, \ldots, K\}$ and $\gamma_v = \gamma_0, \forall v \in V$, the probability of sampling a topic pair for a node pair given all other topic pairs can be written as

$$\begin{aligned} &P(\langle z_{di}, z_{dj}\rangle = \langle k, l\rangle | \langle w_{di}, w_{dj}\rangle, Z^{\neg\langle z_{di}, z_{dj}\rangle}, E^{\neg\langle w_{di}, w_{dj}\rangle}, \\ &\qquad y^{\neg\langle w_{di}, w_{dj}\rangle}, \eta, H(P_V)^{\neg\langle w_{di}, w_{dj}\rangle}, \alpha, \gamma, \rho^2, \sigma^2) \\ &= \frac{\alpha_0 + n_{d\langle k,l\rangle}^{\neg\langle w_{di}, w_{dj}\rangle}}{\sum_{k',l'}(\alpha_0 + n_{d\langle k',l'\rangle}^{\neg\langle w_{di}, w_{dj}\rangle})} \cdot \frac{(\gamma_0 + n_{kw_{di}}^{\neg\langle w_{di}, w_{dj}\rangle})}{\sum_{v \in V}(\gamma_0 + n_{kv}^{\neg\langle w_{di}, w_{dj}\rangle})} \cdot \\ &\frac{(\gamma_0 + n_{lw_{dj}}^{\neg\langle w_{di}, w_{dj}\rangle})^{(1-\delta(w_{di}, w_{dj}))}}{(\sum_{v \in V}(\gamma_0 + n_{lv}^{\neg\langle w_{di}, w_{dj}\rangle})) - \delta(z_{di}, z_{dj})} \cdot \\ &\frac{\exp(\overline{z_d}^{\neg\langle z_{di}, z_{dj}\rangle} \cdot \eta_{y_d})}{\sum_c \exp(\overline{z_d}^{\neg\langle z_{di}, z_{dj}\rangle} \cdot \eta_c)} \cdot \exp\left(\frac{-H((P_{w_{di}}^d)^{\neg\langle w_{di}, w_{dj}\rangle})^2}{2\sigma^2}\right) \cdot \\ &\exp\left(\frac{-H((P_{w_{dj}}^d)^{\neg\langle w_{di}, w_{dj}\rangle})^2}{2\sigma^2} \cdot (1 - \delta(w_{di}, w_{dj}))\right) \end{aligned} \quad (4)$$

where $\delta(\cdot, \cdot)$ is an indicator function that evaluates to 1 when the two inputs are equal, and 0 otherwise.
$n_{d\langle k,l\rangle}$ is the count of node pairs in document d with topic membership $\langle k, l\rangle$. n_{kv} is the count of the number of times a word v is observed under topic k in all documents.The negation $\neg\langle w_{di}, w_{dj}\rangle$ denotes $\langle w_{di}, w_{dj}\rangle$ is

ignored when calculating the corresponding quantity. See the supplementary file for the derivation of Eq 4.

The softmax regression parameters η are then obtained by training a softmax regression model that uses $\overline{z_d}$ as input features and y_d as the response. The inference procedure therefore alternates between sampling $\langle z_{di}, z_{dj} \rangle$ for the sender/receiver pairs in all the graphs and training the softmax regression model to obtain estimates for the η. The topic-specific multinomial distribution over nodes and the topic pair distribution are estimated using the count of observations

$$\beta_{kv} = \frac{n_{kv} + \gamma_0}{\sum_{v'} n_{kv'} + \gamma_0} \tag{5}$$

$$\Theta_{d\langle k,l \rangle} = \frac{n_{d\langle k,l \rangle} + \alpha_0}{\sum_{k',l'} n_{d\langle k',l' \rangle} + \alpha_0} \tag{6}$$

The algorithm for Gibbs sampling based parameter learning is summarized in Algorithm 1.

Algorithm 1 Learning model parameters

Inputs:
A set of graphs $G_d = (W_d, E_d), d \in \{1, \cdots, D\}$, representing the image corpus.
Outputs:
$Z_d, d \in \{1, \cdots, D\}, \Theta, \beta, \eta$.
Begin:

$\forall d \in \{1, \cdots, D\}$, randomly initialize Z_d and increment counters.
while not reaching maximum iterations **do**
  for $d = 1 \rightarrow D$ **do**
    for $\langle w_{di}, w_{dj} \rangle \in E_d$ **do**
      sample $\langle z_{di}, z_{dj} \rangle$ according to Eq. 4
    end for
  end for
  re-estimate Θ, β, using Eq. 6 and Eq. 5, respectively.
  re-estimate $\eta = \arg_{\eta'} \max \prod_d \text{softmax}(P(y_d | \overline{z_d}, \eta'))$.
end while
return $Z_d, d \in \{1, \cdots, D\}, \Theta, \beta, \eta$.

3.2. Classification

To recognize the scene category of an input image m, we first represent it using a grid of visual words $G_m = (W_m, E_m)$. With the parameters Θ, β, we are able to infer the topic membership pair of each node pair $\langle z_{mi}, z_{mj} \rangle$. The scene category is predicted as the one that maximizes the likelihood probability, i.e.

$$y_m = \arg\max_c P(c | \overline{z_m}, \eta) = \arg\max_c \exp(\overline{z_m} \cdot \eta_c) \tag{7}$$

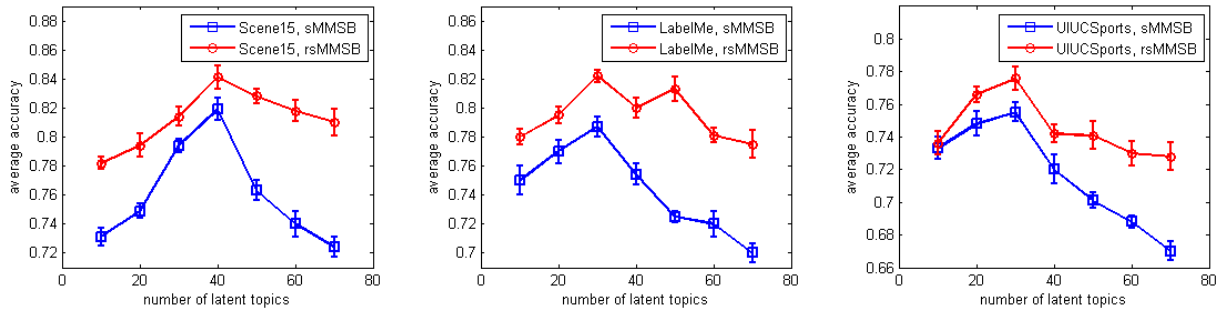
4. Experimental Setup

The proposed model is evaluated on three scene datasets, ranging from generic natural scene images (Scene-15 and LabelMe), to event and activity images (UIUC-Sports):

- Scene-15 [10]. This is a dataset of 15 natural scene classes, and each category has 200 to 400 images. Following [10, 11, 16], we use 100 images in each class for training and rest for testing.
- LabelMe [18]. Following [16, 17, 20], 8 classes are used: highway, inside city, tall building, street, forest, coast, mountain, and open country. 100 images randomly drawn from each scene class are used for training and 100 for testing.
- UIUC-Sports [12]. This is a dataset of 8 event classes, and each has 137 to 250 images. 70 randomly drawn images from each class are used for training and 60 for testing following [12, 17, 20]

All images are converted to grayscale and resized to be no larger than 600×450 pixels with preserved aspect ratio. All experiments are repeated ten times. The final performance metric is reported as the mean of the results from the individual runs. We compared 8 methods. One is a BoW based model with explicit incorporation of the spatial information. Two are BoW based model which exploit visual word correlations for building optimized codebook. Three are LDA based topic models. The last two are MMSB based models proposed in this paper. For a comprehensive comparison of BoW based methods for visual recognition, see [15].

- SPM [10]. SPM partitions the image into increasingly fine sub-regions, and performs histogram matching on local features inside each sub-region.
- Morioka 10 [15]. A codebook for pairs of spatially close SIFT descriptors is built which encodes local spatial information.
- Huang 11 [8]. A codebook graph is constructed where the edges correspond to related visual features.
- Fei-Fei 05 [11]. A LDA based model in which a class-specific Dirichlet prior generates the topic distribution for each document.
- Chong 09 [20]. A LDA based model in which the class is generated from the topic distribution of a document via softmax regression.
- S-DiscLDA [17]. A LDA based model that encodes the appearance and spatial arrangement of scene elements simultaneously.
- sMMSB. This is the supervised MMSB model described in this paper.
- rsMMSB. This is the regularized and supervised MMSB model described in this paper.



(a) Scene-15 dataset. (b) LabelMe dataset. (c) UIUC-Sports dataset.

Figure 3: The recognition accuracy vs. the number of latent topics.

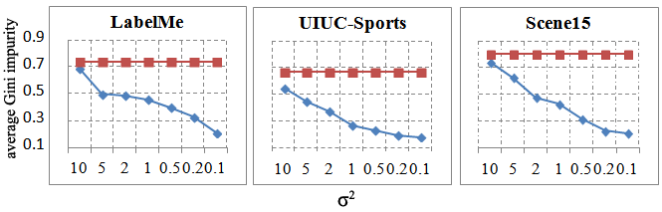


Figure 4: The average Gini impurity with respect to σ^2. Horizontal red line indicates no-regularization baseline.

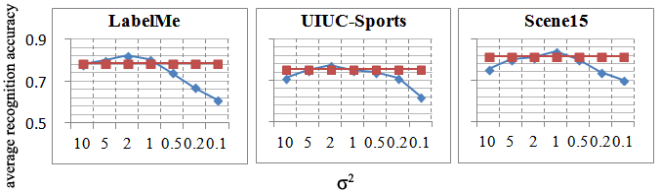


Figure 5: Average recognition accuracy with respect to σ^2. Horizontal red line indicates no-regularization baseline.

5. Experimental Results

5.1. The effect of the number of latent topics

We first investigate how the recognition accuracy varies with respect to the number of latent topics. The results are plotted Figure 3. The first observation is that the recognition performance of rsMMSB is much better than that of its unregularized version, sMMSB, on all three datasets. rsMMSB outperforms sMMSB by as high as 8.6% at 70 latent topics for the Scene-15 dataset. A second observation is that as the number of latent topics increases, both rsMMSB and sMMSB experience performance downgrade, but to different degrees. rsMMSB maintains a satisfactory recognition accuracy even after optimal number of latent topics, and sMMSB begins to overfit severely after the optimal topic number. This can be attributed to the fact that the sparsity regularization enforced by rsMMSB increases the robustness of topic membership inference and scene recognition when the number of latent topics is relatively large compared to the size of the training corpus.

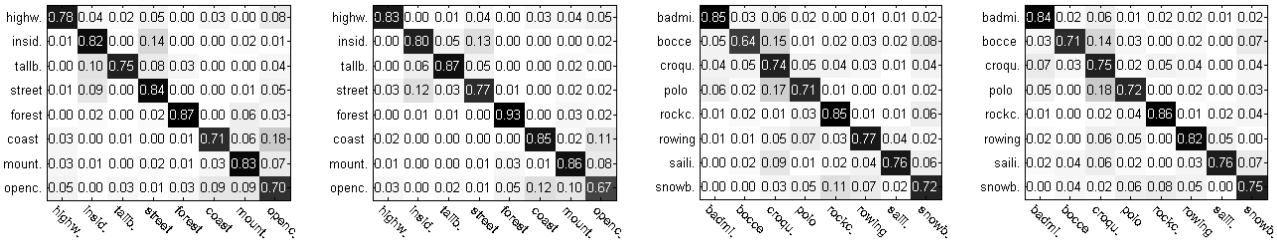
5.2. The effect of the sparsity regularization

In this experiment, we fix the number of latent topics and vary the variance σ^2, the sparsity regularization strictness. Figure 4 shows the average Gini impurity with respect to σ^2. Figure 5 shows the average recognition accuracy with respect to σ^2. As can be seen, a small variance value leads to low Gini impurity due to strict regularization. The recognition accuracy first increases with the variance and then falls as it increases further. Another observation is that three datasets differ in the value of the variance at which the highest recognition accuracy is achieved, with Scene-15 having the most strict regularization (lowest variance value) and UIUC-Sports having the least strict regularization (highest variance value). This can be explained by the fact that Scene-15 has more scene classes and are intrinsicly more complicated in its visual elements. A much lower variance value is thus needed to eliminate false topic inference.

5.3. Comparison to other methods

	SPM	Morioka 10	Huang 11	Fei-Fei 05
Scene-15	0.722 [10]	0.834 [15]	0.829 [8]	0.652 [10]
LabelMe	0.600 [10]	-	-	0.715 [20]
UIUC-Sports	0.720 [13]	-	-	0.360 [20]
	Chong 09	S-DiscLDA	sMMSB	rsMMSB
Scene-15	-	-	0.819	**0.841**
LabelMe	0.760 [20]	0.800 [17]	0.787	**0.812**
UIUC-Sports	0.657 [20]	0.680 [17]	0.755	**0.776**

Table 1: The average accuracy of various methods.



(a) The confusion table of (LabelMe, sMMSB).

(b) The confusion table of (LabelMe, rsMMSB).

(c) The confusion table of (UIUC-Sports, sMMSB).

(d) The confusion table of (UIUC-Sports, rsMMSB).

Figure 6: The confusion table of sMMSB and rsMMSB.

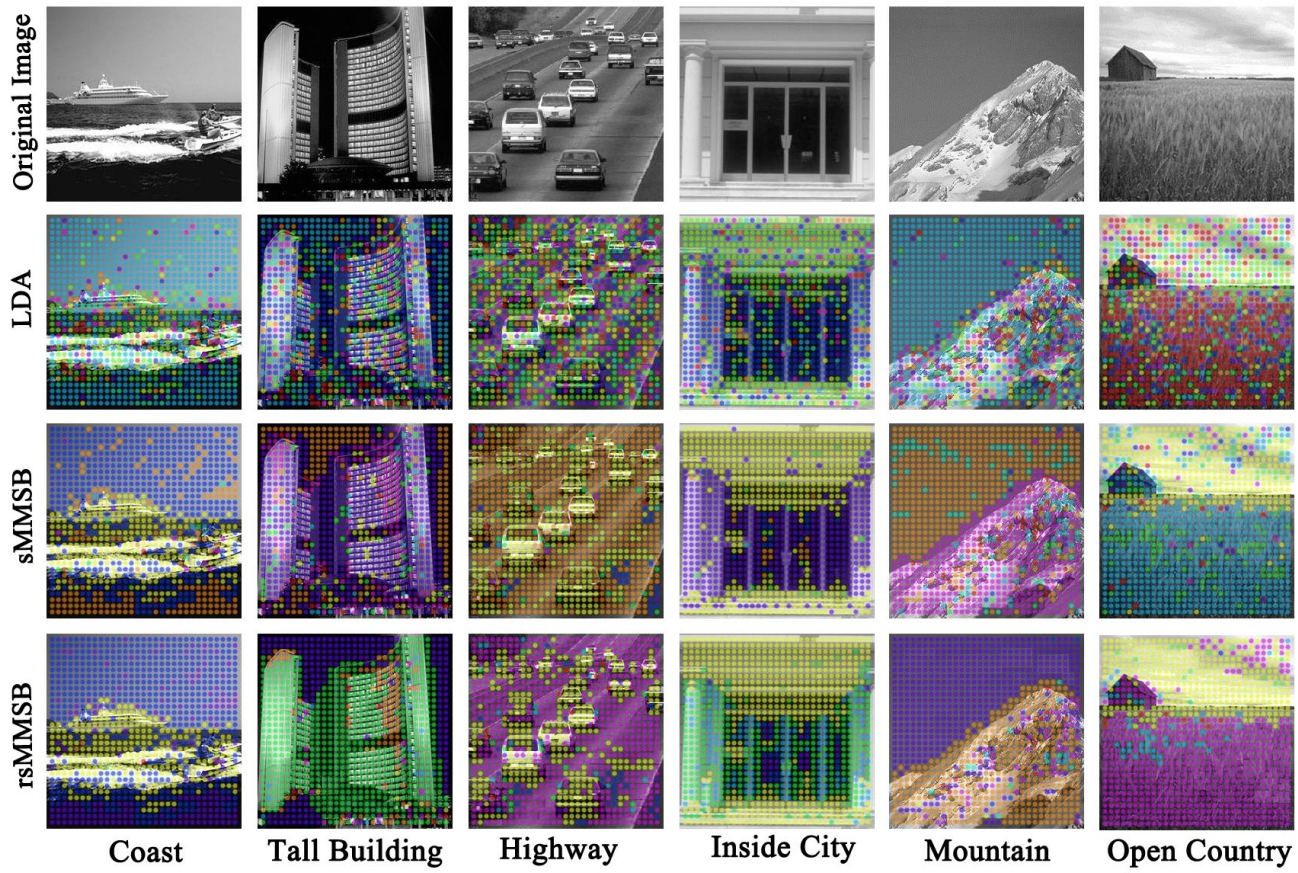


Figure 7: The latent topics inferred for testing images from the Scene 15 dataset. The first row shows the original images, corresponding to coast, tall building, highway, inside city, respectively. Row 2 $\sim$ 4 show the latent topics inferred by LDA, sMMSB, and rsMMSB, respectively. Best viewed in color.

In this subsection, we give a detailed comparison of the methods listed in Section 4. The recognition accuracy of those methods are given in Table 1. The confusion table of sMMSB and rsMMSB at their respective optimal topic numbers are given in Figure 6 (Due to limited space, the confusion table for the Scene-15 dataset is included as a the supplementary file).

As can be seen, rsMMSB performs better than other methods on all three datasets. rsMMSB reduces the error of Chong 09 by at least 6% on both LabelMe and UIUC-Sports, and even more for Fei-Fei 05. For the experiment on LabelMe dataset, S-DiscLDA performs almost as well as rsMMSB, and is better than sMMSB. We conjecture that it is because S-DiscLDA enhances its performance by exploiting the spatial information of scene elements. To verify this, we perform supervised

classification on LabelMe dataset using DiscLDA [9], which makes no use of spatial information. The best average accuracy of DiscLDA is 75.7%, approximately the same as Chong 09 and lower than both sMMSB and rsMMSB. Therefore, it is fair to say that the S-DiscLDA's performance gains from the spatial information. And one naturally expects that, with the incorporation of global spatial layout, sMMSB and rsMMSB will have their performance boosted even further.

5.4. Visualization of the latent topics

Figure 7 visualizes the latent topics of the visual words of the testing images from the UIUC-Sports dataset. The color of the dot denotes the topic membership of the visual word. Three methods are compared, including LDA, sMMSB, rsMMSB. For sMMSB and rsMMSB model, the topic of a visual word is determined as its most frequent one assumed by itself in all its interactions with 4 neighbors (a topic is randomly chosen if the frequency counts are equal). As can be seen, the topics estimated by LDA is rather noisy, i.e. adjacent topics can be quite different from each other, although the image patch appearance is visually similar and semantically correlated. In the results from sMMSB, we obtain much smoother topic estimation. The rsMMSB model gives the best results of all three, as exhibited by the local spatial homogeneity and semantic coherency of the color dots in the image.

6. Conclusion

In this paper, we present a new scene recognition method based on MMSB that achieves coherent and sparse interpretation of the latent topics and produces accurate recognition result. The image is represented in a three-layer hierarchy where a layer of interdependent latent topics generate the visual words. Local spatial coherency of latent topics is naturally built into this model by modeling the joint generation of adjacent topics. The sparsity of the topics is achieved via a Gini impurity based regularizer. Moreover, this framework can be easily extended to incorporate a global spatial layout of latent topics. Experiments demonstrate that our method outperforms traditional topic models and BoW models for scene recognition.

References

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing. Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014, June 2008.

[2] C. W. Balasubramanyan, R. Regularization of latent variable models to obtain sparsity. In *SIAM Intern. Conf. on Data Mining*, pages 414–422, 2013.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, Mar. 2003.

[4] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *ICCV*, pages 1–8, 2007.

[5] I. González-Díaz, D. García-García, and F. Díaz-de María. A spatially aware generative model for image classification, topic discovery and segmentation. ICIP'09, pages 781–784.

[6] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–1465, 2005.

[7] T. Hofmann. Probabilistic latent semantic indexing. SIGIR '99.

[8] Y. Huang, K. Huang, C. Wang, and T. Tan. Exploring relations of visual codes for image classification. In *CVPR*, pages 1649–1656, 2011.

[9] S. Lacoste-Julien, F. Sha, and M. I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *NIPS*, pages 897–904, 2008.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. CVPR '06, pages 2169–2178.

[11] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. CVPR '05.

[12] L.-J. Li and F.-F. Li. What where and who? classifying events by scene and object recognition. ICCV 07.

[13] E. P. X. Li-Jia Li, Hao Su and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *NIPS*, 2010.

[14] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004.

[15] N. Morioka and S. Satoh. Building compact local pairwise codebook with joint feature space clustering. In *ECCV*, 2010.

[16] Z. Niu, G. Hua, X. Gao, and Q. Tian. Context aware topic model for scene recognition. In *CVPR '12*.

[17] Z. Niu, G. Hua, X. Gao, and Q. Tian. Spatial-disclda for visual recognition. CVPR '11.

[18] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. *Int. J. Comput. Vision*, 77(1-3).

[19] C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *ICCV '03*.

[20] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classication and annotation. In *CVPR '09*.

[21] X. Wang and E. Grimson. Spatial latent dirichlet allocation. In *NIPS*, volume 20, pages 1577 –1584, 2007.

[22] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *ICPR*, 2004.

[23] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *CVPR*, 2007.

[24] Y.-T. Zheng, M. Zhao, S.-Y. Neo, T.-S. Chua, and Q. Tian. Visual synset: Towards a higher-level visual representation. In *CVPR*, pages 1–8, 2008.