A SPATIAL CLASS LDA MODEL FOR CLASSIFICATION OF SPORTS SCENE IMAGES

Jin Jeon, Munchurl Kim

Department of Electrical Engineering
Korea Advanced Institute of Science and Technology
291, Daehak-ro, Yuseong-gu, Daejeon, 305-701, Korea
wlsheon@kaist.ac.kr,mkim@ee.kaist.ac.kr

ABSTRACT

Recently, the bag-of-visual words (BoW) models have widely been studied in computer vision area. Owing to the limit of the BoW models that only consider the distributions of visual words in images, the Latent Dirichlet Allocation (LDA) model has drawn an attention to discover the structure of the visual word distributions over latent topics which can represent semantic objects in images. In order to reflect the spatial information of images, the LDA model has been extended to so-called a spatial LDA model for image segmentation, which is not applicable for image classification. Therefore, in this paper, we propose a spatial class LDA (scLDA) model for image classification where the topic distributions over visual words are found per image segments and a class-specific-simplex LDA (cssLDA) model is applied for image classification. From our experimental results, the proposed scLDA model outperforms the previous LDA models in terms of correct classification rates.

Index Terms— LDA, Bag-of-word, Image classification, Spatial information, Sports scene.

1. INTRODUCTION

In computer vision, various methods have been proposed for image classification. One of the popular approaches is to use the BoW models. The BoW models represent an image with a set of visual words where some feature descriptors such as scale-invariant feature transform (SIFT) and histogram of oriented gradients (HoG) are often used to represent visual words for image patches [4, 5]. Then, these feature descriptor values are grouped by a clustering algorithm such as K-means [5]. A set of centroids of each cluster forms a codebook and then each feature descriptor value is mapped to its closest centroid to assign a visual word index. In this way, each image can be represented as a histogram for visual words from the codebook. For classification on image represented by BoW, one of the most popular classifiers is a support vector machine (SVM) and multi-class SVM is generally used due to its powerful performance. However this approach cannot discover the structure of visual word distributions. To overcome this limitation, the LDA model has been used [3, 7, 8, 11]. The LDA model is a generative probabilistic model for collection of text to model a finite mixture over a set of topics [1]. Since the basic LDA model is an unsupervised learning, a supervised LDA (sLDA) model and a class LDA (cLDA) model were proposed [6, 7]. Since these models are not capable of capturing class semantics, a class-specific-simplex LDA (css-LDA) model was proposed which combines the labeling strength of topic supervision with the flexibility of topic-discovery [3]. Although the css-LDA model overcomes the limitation of the previous LDA models, it is still lack of dealing with spatial information which is often important in image processing and computer vision problems.

In this paper, a novel spatial class LDA (scLDA) model is proposed. The main contribution of this paper is to extend the cssLDA model with spatial information for image classification. Our proposed scLDA model can capture the latent topics of visual words by considering spatial locations of image patches. By doing so, each semantic object region can well be represented consistently in terms of the most dominant topic indices of its belonging image patches. This is very essential to improve image classification performance.

This paper is organized as follows: Section 2 briefly reviews an LDA model and BoW model applied for image classification; In Section 3, we describe our proposed scLDA model; We provide experiment results to show the effectiveness of our scLDA model in Section 4 and conclude our work in Section 5.

2. LDA MODEL AND BOW MODEL FOR IMAGE CLASSIFCATION

2.1. LDA model for image classification

Fig. 1 shows the graphical models of LDA for images. In the LDA models, an image is represented as random mixtures over latent topics z where each topic is characterized by a distribution over words, w [1]. For image classification, Fei-Fei *et al.* proposed a class LDA (cLDA) model which introduces class label Y for topic prior α to consider that the images in a same class have a same topic prior distribution as shown in Fig. 1-(a) [7]. Since the LDA model is unsupervised learning, the cLDA model supplements class

variables for image classification. On the other hand, in Fig. 1-(b), Wang *et al*. proposed a multi-class sLDA for image classification based on the supervised LDA which introduces a response label *Y* [11]. The multi-class sLDA treats the class label as a global description of the image, and this model is extended to the multi-class sLDA with annotations. As shown in Fig. 1-(c), Rasiwasia *et al*. proposed a css-LDA model to overcome the limitation of cLDA and sLDA caused by structure problem for classification [3]. The css-LDA model introduced supervision on the visual words directly so that this model can discover the class specific topic simplex.

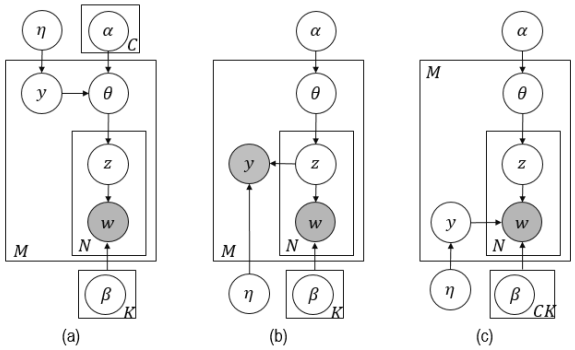

Fig. 1 The graphicl models for (a) cLDA model, (b) sLDA model and (c) css-LDA model.

An extended LDA model, called a spatial LDA model, which includes spatial information was first proposed by Wang *et al*. [13]. The spatial LDA model is applied for images to discover objects. In this model, image patches are regarded as documents, and only one image is considered a corpus. This model is unsupervised learning to cluster the co-occurring and spatially neighboring visual words.

2.2. BoW model for image classification

In our approach, images are represent bag-of-visual words. Fig. 2 shows that the overall process for image representation by the BoW model.

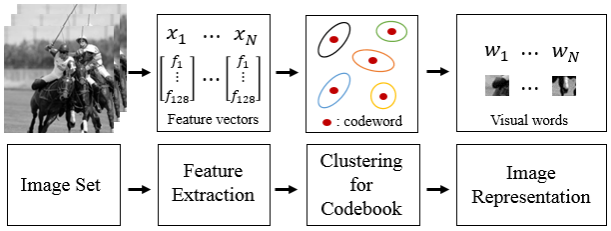

Fig. 2 Codebook generation process. Feature vectors are clustered for generate codebook. This codebook consist of a set of codewords which are the centroid of each cluster.

To represent image as visual words, visual descriptors such as SIFT, HOG and LBP are employed for feature extraction. The Scale Invariant Feature Transform (SIFT) which was proposed by Lowe [10] is widely used the BoW model due to its robustness property. The set of SIFT feature vectors are clustered by *k*-means algorithm where *k* is the number of codeword. Codebook are then defined as a set of centroids of each cluster. Using this codebook, each feature vector is mapped to its closest centroid for generating visual word.

3. PROPOSED SPATIAL CLASS LDA (SCLDA) MODEL FOR IMAGE CLASSIFICATION

3.1. Spatial information

The previous BoW models have a limitation in considering spatial information for image calssification. In images, the spatial locations of image feature vectors play a significant role in image classification where the image feature vectors in image patches which are close each other are in general likely to have similar texture properties. To replect such spatial information into the LDA model for image classification, each image segment region is assumed to be best represented by the visual word at its centroid. Fig. 3 shows an image segment region where two visual words are calculated at different locations.

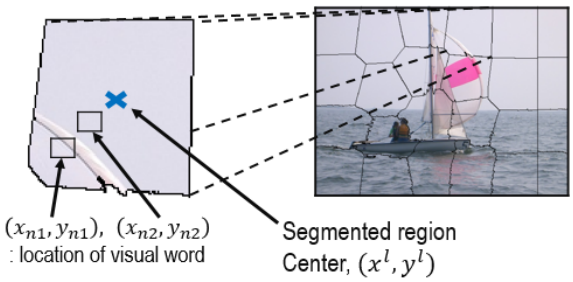

Fig. 3 An example of an image segment with two visual words computed at different locations.

In Fig. 3, the visual word which is close to the center of the segmented region is more proper to represent this segmented region than the other visual word. In order to reflect the spatial information of visual words with weight of topic distributions on the LDA model, we employ a Gaussian kernel which is commonly used [2] as follows:

$$p(s_n \mid p_l, \sigma) \propto \exp\{-\frac{(x^l - x_n)^2 + (y^l - y_n)^2}{2\sigma^2}\} \qquad (1)$$

where $s_n = (x_n, y_n)$ is the location of a current visual word and $p_l = (x^l, y^l)$ is the location of the center of the segmented region. In this way, the closer the location of a viusal word is to the center of the image segment region, the higher the weight is imposed on the topic distribution of the viusal word. This is a reasonable practice because image regions are not perfectly segmented.

Furthemore, we focus on the fact that visual words in the same segmented region have high probability to get the same dominant topic index. For example, if the segmented region is a mountain, visual words in this region are likely to have the same topic to represent the mountain. Fig. 4 shows an example of topic assignment in segmented regions. As shown in Fig. 4, the visual word 5 is likely to get topic index 1 instead of topic index 2. While the earlier LDA models find a topic distribution for each whole image region, we try to subdivide the image for exploring partial topic distribution.

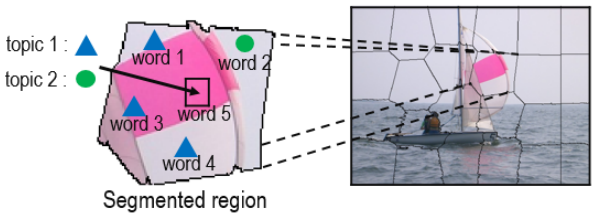

Fig. 4 An example of topic assignment in image segment regions.

3.2. The proposed spatial class LDA model

As mentioned, our spatial class LDA (scLDA) model is designed to utilize not only a location of visual word but also the topic distribution in segmented image regions. Fig. 5 shows the graphical model of the scLDA where there are M images and each image has N visual words and the number of topics is K. In Fig. 5, α, σ, β, η, and η' are the hyperparameters of topic proportion θ, visual word location s, visual word w, class label y and segment region p. z is a latent variable, indicating the topic label of w.

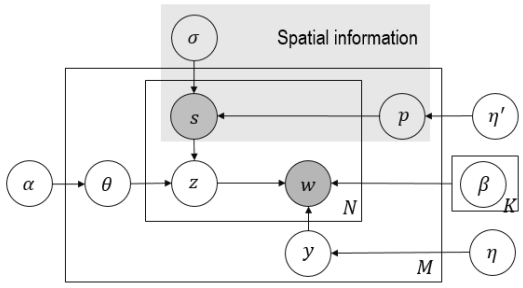

Fig. 5 The graphical model of scLDA.

The generative process of scLDA is described as follows:

i) Draw topic proportions $\theta \mid \alpha \sim Dir(\alpha)$.

ii) Draw class label $y \mid \eta \sim uniform\{1,...,C\}$.

iii) For each visual word w_n :

(a) Draw topic label $z_n \mid \theta, s \sim \text{Mult}(\theta) \cdot p(s \mid \cdot)$; and

(b) Draw word $w_n \mid z_n, y, \beta \sim \text{Mult}(\beta_{z_n}^y)$.

The joint distribution of the scLDA model for a set of K topics (z), a set of N words (w), and a location (s) and a segmented region (p) for each visual word is given by

$$P(\mathbf{w},\mathbf{z},\theta,y,\mathbf{s},p \mid \alpha,\beta,\eta,\eta',\sigma) \propto P(\theta \mid \alpha)\prod_{n=1}^{N} P(y \mid \eta)P(z_n \mid \theta, s_n, p, \sigma, \eta')P(w_n \mid z_n, y, \beta) \quad (2)$$

The distribution of a latent topic z with the spatial terms is given by

$$\log P(\mathbf{z} \mid \theta, \mathbf{s}, p, \sigma, \eta') \propto z_{ni} \log \theta_i^p \exp\{-\frac{(x^l - x_n)^2 + (y^l - y_n)^2}{2\sigma^2}\} \quad (3)$$

where z_{ni} is the i-th topic of visual word n, θ_i^p is the i-th topic distribution of a segmented region p, and the exponent term, which is spatial information, is defined in Eq. (1).

3.3. Parameter Learning

Learning involves estimating the parameters by maximizing the log likelihood of Eq. (2) for a training image dataset D. Since Eq. (2) is intractable for inference as the case of the LDA model, we use an approximate inference algorithm [1] where variational E-step approximates the posterior by a variational distribution $q(\gamma,\phi)$ via the following update rules:

$$\phi_{ni} \propto \beta_{ij}^y \exp\{S'(\Psi(\gamma_{ip}) - \Psi(\textstyle\sum_{i,p}\gamma_{ip}))\} \quad (4)$$

$$\gamma_{ip} = \alpha_i + \textstyle\sum_l \sum_n \delta(p_l, p)\phi_{ni} \quad (5)$$

where ϕ_{ni} is the probability of the i-th topic for the n-th word w_n in an image, β_{ij}^y is the multinomial parameter of the i-th topic of the j-th word in the vocabulary for class y, S' is a spatial distance weight defined as the right-hand side term of Eq. (1), and $\Psi(x)$ is the first derivative of the $\log\Gamma(x)$ function where $\Gamma(x)$ is the Gamma function. γ_{ip} in Eq. (5) is the variational Dirichlet topic parameter of the i-th topic for the p-th segmented region. Variational M-step computes the value of parameters. Since the scLDA model in Fig. 5 is not sensitive to α, we set α to a fixed value. β_{ij}^y is computed as in [3] by

$$\beta_{ij}^y \propto \textstyle\sum_d \sum_n \delta(y^d, y)\phi_{ni} w_{dn}^j \quad (6)$$

where $\delta(\cdot)$ is a selection operator such that $\delta(y^d, y) = 1$ if the class of the d-th image is equal to y, that is, $y^d = y$.

3.4. Image Classification

For classification, the category of an image I is decided by the maximum posterior probability such as

$$y^* = \arg\max_y P(y \mid I) \quad (7)$$

where given an image I, the posterior probability of class y is computed as in [1] by maximizing the evidence lower bound $L(\gamma,\phi;\alpha,\beta,\eta,\eta',\sigma)$ which is equivalent to minimizing KL divergence between the actual distribution and the variational distribution according to Eq. (4) - Eq. (6).

4. EXPERIMENTAL RESULTS

4.1. Datasets

We test our model on UIUC Sports Dataset [8] which contains 8 categories of sports events: *badminton*, *bocce*, *croquet*, *polo*, *rock climbing*, *rowing*, *sailing*, and *snowboarding*. The UIUC Sports Dataset dataset contains 1,579 images in total, and each category has 137 to 250 images. For experiments, we convert the color images to gray scale and resize images to a maximum of 256 along the larger border. As is the previous studies, we selected 70 images per category for training and 60 images for test.

4.2. Codebook generation

As reported in [7], we performed feature extraction in 8×8 regular grids. For feature extraction, we used SIFT

descriptor [10] over 4×4, 6×6, 8×8, and 10×10 pixel patches with overlap allowed. For each image patch, 128-dimensional feature vectors are obtained and total 2,400 SIFT features were computed per image in average. The codebook of visual words were obtained by K-means clustering where K ranges between 1,024 and 4,096. For each dataset, the codebook was generated from 30 randomly selected training images per category.

4.3. Classification results

We first perform segmentaion on images by using the superpixel algorithm [9], and then compute the locations of patch centers in each segmented images. The classification accuracy for images in UIUC Sports Dataset is 81.6% for the codebook of 4,096 size.

Fig. 6 shows that the confusion table for the 8-class sports event classification experiment. The vertical axis of the confusion table is predicted categories and the horizontal axis is the ground truth of the categories. The classification precision of each category is the diagonal line of the confusion table. The experiment result shows that the precision of the category of '*badminton*' is high and the category of '*croquet*' is low. Some images in the '*croquet*' class are misclassified, instead being classified to the '*bocce*' class due to their background similarity.

polo	0.85	0.02	0.02	0.07	0.02	0.02	0	0.02
badminton	0	0.93	0	0	0	0.05	0	0.02
bocce	0.02	0.05	0.73	0.03	0.05	0.07	0	0.05
croquet	0.08	0	0.23	0.65	0.02	0.02	0	0
rock	0.02	0	0.08	0	0.87	0	0	0.03
rowing	0	0	0.07	0.03	0.02	0.8	0.03	0.05
sailing	0.02	0	0.05	0.05	0.02	0.03	0.83	0
snow	0.02	0	0.02	0.02	0.03	0.03	0.02	0.87

Fig. 6 The confusion table for the 8-class event classificaion experiment.

Fig. 7 shows the classification accuracy of our scLDA model in comparison with other LDA-based models. The proposed scLDA model outperforms the css-LDA model and sLDA model in terms of classification accuracy. The scLDA model effectively works with the spatial information so that it can better infer topic distributions of visual words.

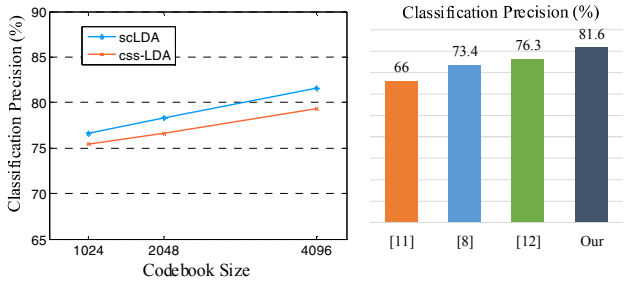

Fig. 7 The classification accuracy according to the codebook size and the comparison with other models.

Fig. 8 illustrates the topic distribution of images in details. In Fig. 8, small square regions in the image grids indicates image patches, each of which is indicated with different shaded colors. The same shaded colors indicate the same dominant topics for image patches. As shown in Fig. 8, each segment region tends to be represented by a same shaded color. This implies that the image patches in a same image segment have almost the same dominant topic index. On the other hand, it is very hard to see some semantic topic distribution over an image by the css-LDA model for which the dominant topic indices of image patches appear almost ramdomly over the image.

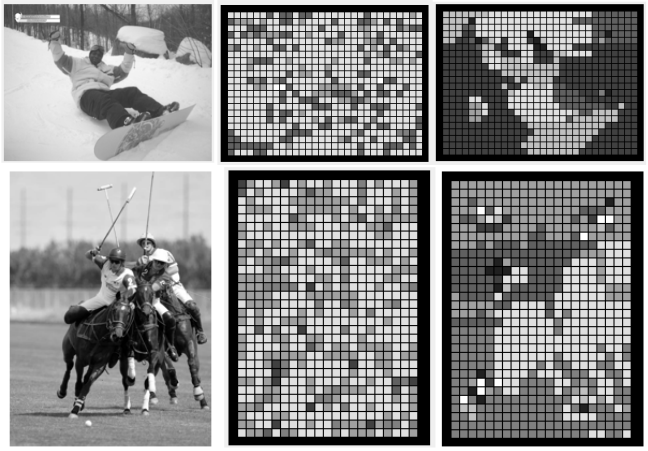

Fig. 8 Topic distribution results. (left) original image. (middle) the topic distribution by css-LDA model. (right) the topic distribution by scLDA model.

For our scLDA, image segmentation is required *a priori* which increases computational complexity. The complexity of parameter learning for the scLDA model is similar to the cssLDA model. However, since the scLDA performs the parameter learning for each segmented image region, the parameter learning can be processed in parallel.

5. CONCLUSION

In this paper, we propose a spatial class LDA model for image classification. In order to incorporate spatial information into the LDA model, we supplement the location of visual words and explore partial topic distribution in segmented image regions. From the experiment results, the proposed scLDA model outperforms other LDA based models with higher precision performance for image classification and shows semantic topic distributions over segment regions in images. As our future work, we will consider Dirichlet parameter update to improve image classification performance.

ACKNOWLEDGEMENT

This work (Project ID: 2014R1A2A2A01006642) was supported by Mid-career Researcher Program through NRF grant funded by the MEST.

REFERENCES

[1] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," *The J. Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[2] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[3] N. Rasiwasia, N. Vasconcelos, "Latent Dirichlet Allocation Models for Image Classification," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 1, no. 9, pp. 946–949, 2013.

[4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 2169-2178, New York, NY, USA, 2006.

[5] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 29-44, 2001.

[6] D. M. Blei and J. McAuliffe. "Supervised topic models," *Proc. Advances in Neural Infor. Process. Sys.*, 2007.

[7] L. Fei-Fei and P. Perona, "A Bayesian Hierarchical Model for Learning Natural Scene Categories," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 524-531, San Diego, CA, USA, 2005.

[8] L.J. Li and L. Fei-Fei, "What, where and who? Classifying events by scene and object recognition," *Proc. IEEE Int. Conf. Compute Vision,* pp. 1-8, Rio de Janeiro, Brazil, 2007.

[9] R. Achanta, A. Shaji, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Analysis and Machine Intell.*, vol. 34, no.11, pp. 2274-2281, 2012.

[10] D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no.2, pp.91-110, 2004.

[11] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," *IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1903–1910, Miami, FL, USA, 2009.

[12] L.J. Li, H. Su, Xing and L. Fei-Fei, "Object bank: a high-level image representation for scene classification and semantic feature sparsification," *Proc. Advances in Neural Infor. Process. Syst.,* vol. 24, 2010.

[13] X. Wang and E. Grimson, "Spatial Latent Dirichlet Allocation," *Proc. Advances in Neural Infor. Process. Syst.,* 2007.