Supervised LDA for Image Annotation

Guo Qiaojin, Li Ning, Yang Yubin and Wu Gangshan
National Key Laboratory for Novel Software Technology
Nanjing University
Nanjing 210093, China
Email: guoqiaojin@gmail.com, ln@nju.edu.cn, yyb@nju.edu.cn and gswu@nju.edu.cn

***Abstract*—Region-based Image Annotation has received increasing attention in recent years. Topic models such as probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA) have shown great success in object recognition and localization. In this paper, we introduce a supervised topic model for region-based image annotation. Images are segmented into superpixels, and visual features are extracted from each superpixel region. Boosted classifiers are then trained for each class, and the output of boosted classifiers are quantized as boosted visual words. The proposed model builds a generative model on both visual words and corresponding class labels. We tested the model on the 21-class MSRC dataset. Experimental results show that our model improves the annotation performance comparing with boosted classifiers.**

***Index Terms*—Image Annotation; latent Dirichlet Allocation; Variational Inference**

I. Introduction

Automatic image annotation assigns metadata, usually keywords, to images automatically, makes it easier for indexing and maintaining large collections of images, plays an important role in image retrieval systems. It has been studied a lot in the last decades. Region-based image annotation, also known as region-naming, region-labeling, and multi-class image segmentation, is an important part of image annotation. Various machine learning techniques were employed for learning the correspondence between image regions and keywords [1], [2], [3], [4], [5], [6], [7], [8], [9] .

For region-based image annotation, each image is annotated with a set of keywords associated with its location. Fig. 1 shows one sample image from 21-class MSRC dataset [10], each pixel of the images is associated with one of the 21 classes, with additional void class. Region-based image annotation can be divided into two procedures, images are first segmented into several regions and visual features are extracted from each region, then each region is annotated by utilizing different machine learning technologies. There are various methods for image segmentation, Barnard et al. [6] evaluated some image segmentation algorithms for region-based image annotation. The two most frequently used strategies are dense block [1], [3], [4] and over-segmentation [8], [2], [5]. Dividing images with dense blocks is much faster than over-segmentation, while different objects may share the same region. Over-segmentation is usually computationally expensive, while the different objects are segmented into different regions with more accuracy. Fig. 1(c) shows the over-segmented superpixels of one sample image from MSRC, there

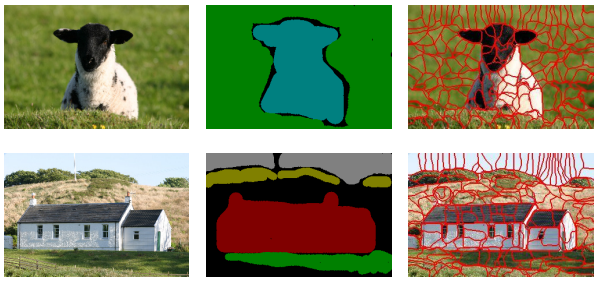

Fig. 1. Sample images from MSRC (the first column) with ground truth annotation (the second column) and over-segmented superpixels (the right column).

are approximately 200 regions in this images, we can see that most of the regions contains only one object class.

After segmentation and feature extraction, a statistical model is build to learn the correspondence between labels and features of each region from training images. Various models have been used for annotating the segmented images. Richard et al. [1] incorporated region and global features with Conditional Random Field for labeling regions of images. Shotton et al. [2] trained a discriminative model for automatic recognition and segmentation by incorporating appearance, shape and context information.

Topic models such as Probabilistic Latent Semantic Analysis (pLSA) [11] and Latent Dirichlet Allocation (LDA) [12] has received increasing attention in recent years for multi-class image segmentation and annotation. Sivic et al. [7] utilized pLSA and LDA for discovering object categories in image collections. Verbeek [3] and Mackey [4] extended the topic models with Markov Random Field over the latent topic for capturing the spatial relations of image regions. Barnard et al. [5] proposed a multi-modal extension to mixture of latent Dirichlet allocation (MoM-LDA) for image segmentation with associated text. Cao and Li [8] presented a generative model for object recognition and segmentation by incorporating the spatial coherence of images and scenes. Images are represented by over-segmented regions and image patches within one region shares the same topic.

Most of the topic models are unsupervised, the latent topics are used to capture the class probabilities in most previous works. Blei and McAuliffe [13] introduced supervised latent Dirichlet allocation to predict response values for new documents. In this paper, we proposed a modified supervised

978-1-4577-0653-0/11/$26.00 ©2011 IEEE

topic model for region-based annotation where each region has its own class label. We build a generative model on both the visual words and labels, all the regions from one image have latent topics drawn from one multi-nominal distribution, classification performance is improved with the help of latent topics.

The rest of this paper is organized as follows. In Section 2, we first briefly reviewed latent Dirichlet Allocation, and then describes our supervised latent Dirichlet Allocation for region-based image annotation. Experiments and results are detailed in Section 3.

II. The proposed method for Image Annotation

A. Latent Dirichlet allocation (LDA)

Latent Dirichlet allocation (LDA) [12] captures the semantic theme by building distributions over a set of words, called vocabulary. The topics of each document are drawn from a Dirichlet distribution. The graphical model representation of LDA is shown in Fig. 2. LDA is one kind of bag of words model and is used for text analysis originally. By representing images with bag-of-words model, the visual features need to be quantized. For image annotation, the images are first segmented into superpixels, as shown in Fig. 1(c). In this paper, we use the segmentation algorithm proposed by Greg Mori [14], each image is segmented into approximately 200 regions with Normalized Cuts algorithm. Each superpixel region is associated with one single class label. For each region, we extract color, texture, geometry and location features. As described in [15], performance can be greatly improved by utlizing boosted classifiers. A one-vs-all boosted classifier is trained for each class, the boosted features are generated by utlizing the output of the learned boosted classifier. The visual vocabulary is built by clustering on the raw features or boosted features, and each region is represented by the quantized visual features.

The generative process of LDA for each image is :

1) Choose θ from $\mathrm{Dir}(\alpha)$.
2) For each of the images of N regions with visual words w_n:
   a) Choose a topic z_n from $\mathrm{Multinomial}(\theta)$.
   b) Choose a visual word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.

B. Class-Specified Latent Dirichlet allocation (csLDA)

LDA is a unsupervised generative model and class labels of each region is ignored during the training procedure. In order to handle the labels, we proposed a modified supervised latent Dirichlet allocation, which builds a generative model with both visual words and class labels. Different with sLDA proposed in [13], each region of the images has different labels, we called it Class-Specified LDA (csLDA). The graphical model representation of csLDA is shown in Fig. 3.

Topic proportions θ are drawn from Dirichlet distribution $\mathrm{Dir}(\alpha)$. Topics of each image are drawn from a multi-nominal distribution $\mathrm{Mult}(\theta)$, and the words are drawn from a multi-nominal distribution $\mathrm{Mult}(\beta, z)$, where z is topic and β is describes the probabilities of each word w with corresponding topic z. The class of each region c is also drawn from a multi-nominal distribution $\mathrm{Mult}(\eta, z)$.

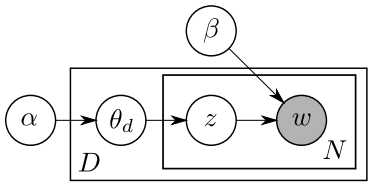

Fig. 2. The graphical model representation of LDA.

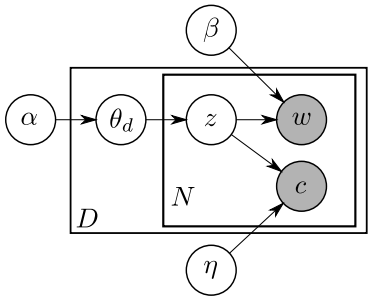

Fig. 3. The graphical model representation of csLDA.

The generative process of csLDA for each image is :

1) Choose θ from $\mathrm{Dir}(\alpha)$.
2) For each of the words N regions with visual words w_n and class label c_n :
   a) Choose a topic z_n from $\mathrm{Multinomial}(\theta)$.
   b) Choose a visual word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic z_n.
   c) Choose a class label c_n from $p(c_n|z_n,\eta)$, a multinomial probability conditioned on the topic z_n.

The parameters α, β, η are corpus-level parameters, θ_d are document-level variables, and the topics z_{dn} are word-level variables.

For a given image with visual words $w_i, i = 1, ..., N$ and labels $c_i, i = 1, ..., N$, with N the number of regions in current image. The probability conditional on parameters α, β, η is :

$$P(w_{1:N}, c_{1:N}|\alpha,\beta,\eta) = \int P(\theta|\alpha) \{\prod_{n=1}^{N} P(z_n|\theta)P(w_n|z_n,\beta)P(c_n|z_n,\eta)\}d\theta \quad (1)$$

The probability of the training corpus is :

$$P(D|\alpha,\beta,\eta) = \prod_{d=1}^{D} \int P(\theta_d|\alpha) \{\prod_{n=1}^{N} P(z_{dn}|\theta)P(w_{dn}|z_{dn},\beta)P(c_{dn}|z_{dn},\eta)\}d\theta_d \quad (2)$$

C. Variational Inference

Directly maximizing the likelihood is intractable, thus, variational method [12] is employed to maximize the lower bound of log likelihood.

$$\begin{aligned}\log P(w_{1:N}, c_{1:N}|\alpha,\beta,\eta) \geq\ & E_q[\log p(\theta|\alpha)] \\ & + E_q[\log p(z_{1:N}|\theta)] \\ & + E_q[\log p(w_{1:N}|z_{1:N},\beta)] \\ & + E_q[\log p(c_{1:N}|z_{1:N},\eta)] \\ & - E_q[\log q(\theta)] \\ & - E_q[\log q(z_{1:N})]\end{aligned} \quad (3)$$

where $E_q[\cdot]$ is the expectation and q is the variational distribution:

$$q(\theta, z_{1:N}|\gamma,\phi) = q(\theta|\gamma)\prod_{n=1}^{N} q(z_n|\phi_n) \quad (4)$$

As shown in (3), the only difference between csLDA and LDA is $E_q[\log p(c_{1:N}|z_{1:N},\eta)]$, thus the coordinate update of α,β,γ is the same as LDA. For maximizing the lower bound of log likelihood described in (3). The update of η and ϕ are :

$$\eta_{ij} \propto \sum_{d=1}^{D}\sum_{n=1}^{N} \phi_{dni} c_{dnj} \quad (5)$$

$$\phi_{ni} \propto \beta_{iv}\eta_{iu} exp(\psi(\gamma_i) - \psi(\sum_{j=1}^{k}\gamma_j)) \quad (6)$$

where D is the number of training images and v, u are the indexes of word and label of region n, $c_{dnj} = 1$ if and only if the label of the nth region in image d has index j, otherwise, $c_{dnj} = 0$.

The procedure of training csLDA is shown as fllowing:

initialize $\tau = 0$
initialize $\alpha^\tau, \beta^\tau, \eta^\tau$
repeat
  for each image d in training set
    initialize $t = 0$
    initialize $\phi^t_{dni} = \frac{1}{k}, i = 1, ..., k$ for all n
    initialize $\gamma_{di} = \alpha_i + \frac{N}{k}, i = 1, ..., k$
    repeat
      for $n = 1$ **to** N
        for $i = 1$ **to** k
          update parameter ϕ^{t+1}_{dni}
      update parameter γ^{t+1}_d
    until convergence
  $\tau = \tau + 1$,update parameter $\alpha^\tau, \beta^\tau, \eta^\tau$
until convergence

D. Prediction

In the procedure of prediction, the labels of regions are unknown. Thus, the label node is removed from the graphical model, and same as LDA. The topic probabilities of each region $\phi_k, k = 1, ..., K$ is computed, and the labels of each region is computed by:

$$\hat{c} = \text{argmax}_c \sum_{k=1}^{K} \eta_{ck}\phi_k \quad (7)$$

The prediction procedure of each test image is :

initialize $t = 0$
initialize $\phi^t_{dni} = \frac{1}{k}, i = 1, ..., k$ for all n
initialize $\gamma_{di} = \alpha_i + \frac{N}{k}, i = 1, ..., k$
repeat
  for $n = 1$ **to** N
    for $i = 1$ **to** k
      update parameter ϕ^{t+1}_{dni}
  update parameter γ^{t+1}_d
until convergence
predict the class label of each region with (7)

E. csLDA-MRF

Verbeek [3] and Mackey [4] extended the topic models with Markov Random Fields over the latent topic for capturing the spatial relations of image regions. In this paper, we utilize Markov Random Field to capture the spatial relations of the labels, not topics.

$$P(\boldsymbol{c}) \propto \exp(\sum_{i\in\mathcal{N}} \log(\sum_{k=1}^{K}\eta_{ck}\phi_{ik}) + \sum_{ij\in\mathcal{E}} f(c_i, c_j)) \quad (8)$$

where $\mathcal{N}, \mathcal{E}$ are the set of nodes and edges of markov random fields, and $f(c_i, c_j) = \rho\delta(c_i = c_j)$. δ is the indicator function and $\delta(c_i = c_j) = 1$ if and only if $c_i = c_j$. ρ is set empirically, and in this paper we set $\rho = 0.7$.

Different with [3], [4], the inference procedure is only executed during prediction.

III. Experiments

We evaluated csLDA on the 21-class MSRC dataset [10], which consists of 591 pixel-wise annotated images. The sample images are shown in Fig. 1. We split the images into 296 training and 295 test. First, each image is over-segmented into superpixels. We use the code provided by Greg Mori [14], and each image is segmented into approximately 200 regions. For each region, we extract color, texture, geometry and location features. Visual vocabulary is constructed by using kmeans, and the features are quantized into 500 visual words. Experimental results shows that training csLDA directly with raw appearance features poorly on MSRC dataset. Thus according to [15], we trained boosted classifiers for each class and using the output of boosted classifiers as features instead of using the raw appearance features, and built new vocabulary on the boosted features. The feature extraction and training boosted classifiers are performed by utilizing STAIR Vision

TABLE I
THE ACCURACY OF BASELINE METHOD, CSLDA AND CSLDA-MRF WITH 100 TOPICS.

Accuracy	building	grass	tree	cow	sheep	sky	airplane	water	face	car	
baseline	**63**	93	**80**	37	36	**92**	35	48	59	26	
BFS	55	93	74	59	**60**	80	76	51	77	52	
BFS-MRF	61	**94**	78	**63**	**60**	91	**81**	**52**	**87**	**62**	
bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	Average
58	58	29	8	58	10	79	24	22	30	9	64
64	72	28	6	77	21	79	38	45	45	23	69
73	**74**	**35**	**13**	**85**	**24**	**82**	**40**	**50**	**52**	**24**	**73**

Library [15] provided by Stephen Gould. The Markov Random Fields is inferenced with Loopy Belief Propagation provided by libDAI [16]. Void regions are ignored for both training and testing.

The training and testing process of csLDA with boosted visual words are :

1) Segment each image in the training set into superpixels.
2) Extract features from each region.
3) Train boosted classifiers on the raw features.
4) The output of boosted classifiers are used as boosted features.
5) Quantize the boosted visual features and build boosted vocabulary.
6) Represent each image with $\{w_n, c_n\}, n = 1, ..., N$.
7) Train csLDA with training images for parameters α, β, η.
8) For each test image, get parameters ϕ with general LDA and predict each region with (7).
9) The final class labels of regions are smoothed with MRF.

The resulting mean accuracy of boosted classifiers is 64.40%, and the confusion matrix is shown in Fig. 5(a). The confusion matrix of csLDA with different number of topics are also shown in Fig. 5. The mean accuracy of csLDA on raw features and boosted features are shown in Fig. 4. As the number of topic increases, the mean accuracy of csLDA with BFS increases and finally converged at approximately 69% and csLDA-MRF converged at approximately 72%, which performs better or equally well with previous works [2], [3], [9]. Table I shows the accuracy of each semantic class of baseline method and csLDA with 100 topics, and csLDA-MRF performs better than baseline method on 18 classes. The average accuracy of csLDA with 100 topics on all 21 classes is 69.1%, and csLDA-MRF achieved 72.6%, and the accuracy of boosted classifiers is 64.4%. Training csLDA on quantized boosted features (BFS) performs better than raw features (RFS). Fig. 6 shows some sample images from MSRC dataset and the annotation results performed by boosted classifiers and csLDA. We observed that by utilizing csLDA, the annotation accuracy performance are improved.

IV. CONCLUSION

This paper presented a new supervised topic model for region-based image annotation. LDA is unsupervised topic model, and cannot handle the labels of segmented regions.

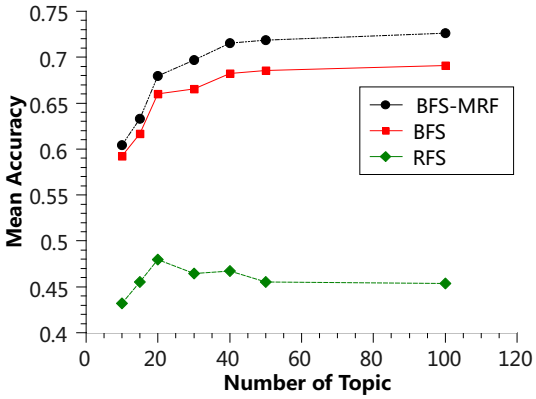

Fig. 4. Accuracy of csLDA with quantized raw features (RFS) and quantized boosted features (BFS), and BFS with MRF, the x-axis refers to different number of topics, and the y-axis refers to the average accuracy of 21-classes.

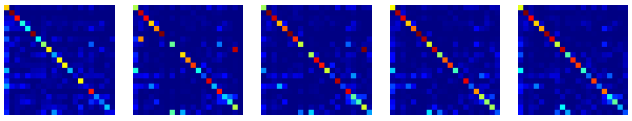

Fig. 5. The confusion matrix of baseline method and csLDA with different number of topics. (a): the confusion matrix of boosted classifier on test images; (b): the confusion matrix of csLDA based on boosted words with 30 topics; (c): the confusion matrix of csLDA based on boosted words with 40 topics; (d): the confusion matrix of csLDA based on boosted words with 50 topics; (e): the confusion matrix of csLDA based on boosted words with 100 topics.

In this paper, We build a generative model on both the visual words and labels, all the regions from one image have latent topics drawn from one multi-nominal distribution, the visual words and labels are both drawn from multi-nominal distribution with specific topic, and classification performance is improved with the help of latent topics. Images are first over-segmented into superpixels and features are extracted from each region. Boosted classifiers are trained with the raw features and the output of classifiers are quantized to generate boosted words and vocabulary. The proposed method built a generative model on the boosted words and labels of over-segmented regions. By utilizing csLDA-MRF, the annotation accuracy increases from 64.4% to 72.6% with 100 topics.

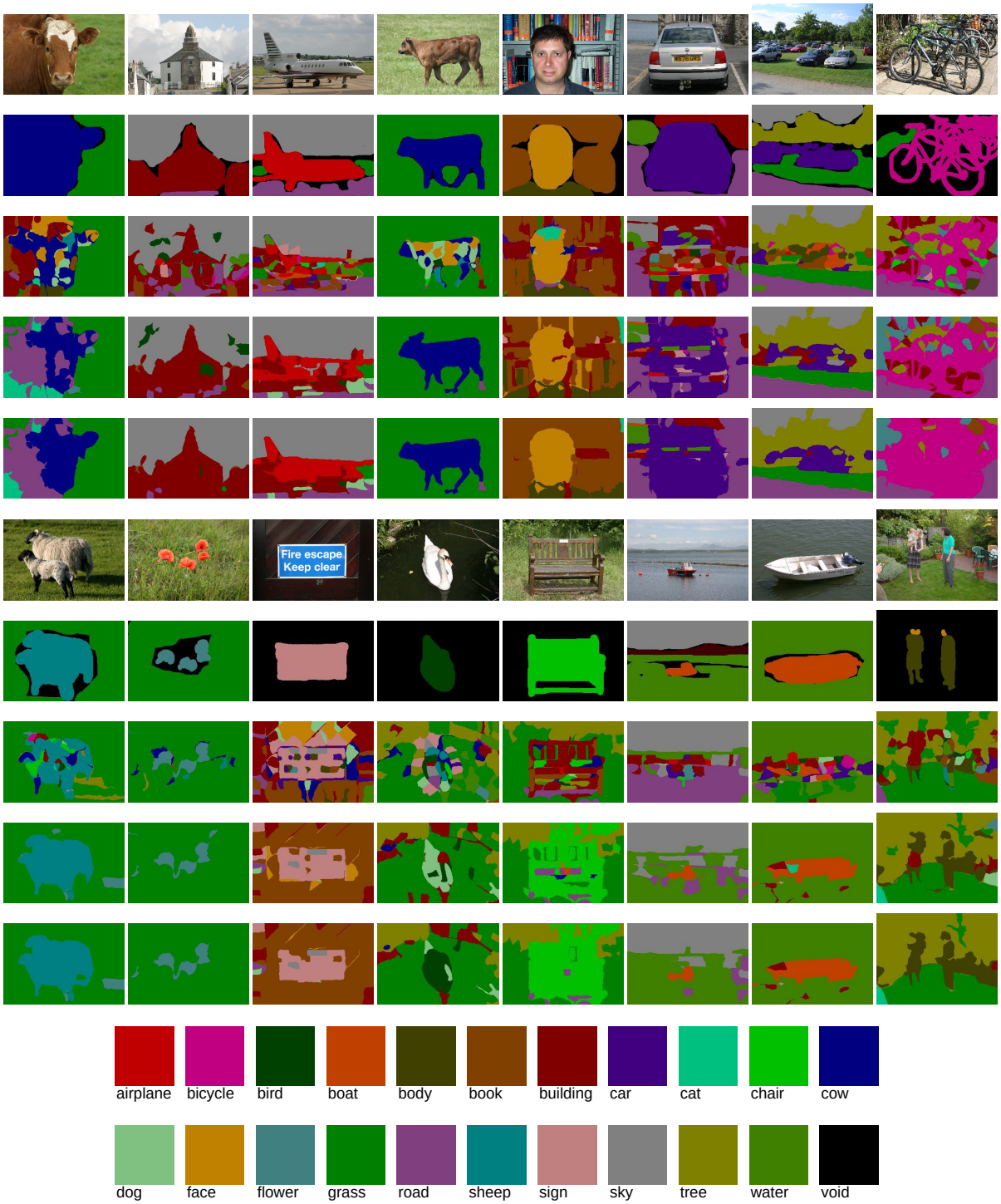

Fig. 6. Example test images from the 21-class MSRC database (the first row) ,the groudtruth annotation (the second row), and the annotations produced by the baseline method (the third row) and csLDA with 100 topics (the fourth row), the firth row refers to the annotation produced by csLDA-MRF with 100 topics. The void class is ignored for both training and testing.

ACKNOWLEDGMENT

This work is supported by the "973" Program of China (Grant No. 2010CB327903), the National Natural Science Foundation of China (Grant Nos. 60875011, 60723003), and the Key Program of Natural Science Foundation of Jiangsu Province, China (Grant BK2010054).

REFERENCES

[1] X. H. Richard, R. S. Zemel, and Miguel, "Multiscale Conditional Random Fields for Image Labeling," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 695–702, 2004.

[2] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation," in *Proceedings of European Conference on Computer Vision*, pp. 1–15, 2006.

[3] J. Verbeek and B. Triggs, "Region classification with markov field aspect models," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2007.

[4] L. Mackey, "Latent Dirichlet Markov Random Fields for Semi-supervised Image Segmentation and Object Recognition," *Technical Report, Computer Science, University of California, Berkeley*, 2007.

[5] K. Barnard, P. Duygulu, D. Forsyth, N. De Freitas, D. Blei, and M. Jordan, "Matching words and pictures," *The Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003.

[6] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. Forsyth, "The effects of segmentation and feature choice in a translation model of object recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 675–82, 2003.

[7] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering Object Categories in Image Collections," in *Proceedings of IEEE International Conference on Computer Vision*, 2005.

[8] L. L. Cao and F. F. Li, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *Proceedings of IEEE International Conference on Computer Vision*, pp. 1–8, 2007.

[9] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[10] A. Criminisi, "Microsoft research cambridge object recognition image database." http://research.microsoft.com/vision/cambridge/recognition/, 2004.

[11] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57, 1999.

[12] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[13] D. Blei and J. McAuliffe, "Supervised topic models," in *Advances in Neural Information Processing Systems*, vol. 20, pp. 121–128, 2008.

[14] G. Mori, X. Ren, A. Efros, and J. Malik, "Recovering human body configurations: Combining segmentation and recognition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 326–333, 2004.

[15] S. Gould, O. Russakovsky, I. Goodfellow, P. Baumstarck, A. Ng, and D. Koller, "The stair vision library (v2.4)." http://ai.stanford.edu/ sgould/svl, 2010.

[16] J. M. Mooij, "libDAI: A free and open source C++ library for discrete approximate inference in graphical models," *Journal of Machine Learning Research*, vol. 11, pp. 2169–2173, 2010.