A Novel Context-Aware Topic Model for Category Discovery in Natural Scenes

Zehuan Yuan and Tong Lu*

National Key Laboratory of Software Novel Technology, Nanjing University, China

Abstract. Automatic category discovery from images is a challenging problem in computer vision community especially from natural scene images due to the great variability in them. This paper proposes a novel context-aware topic model for category discovery in complex natural scenes. The proposed model constructs a generative probabilistic procedure from three-level features consisting of patch, region and the entire image by introducing latent topic variables to every patch and every region. Additionally, a new kind of scene context prior, namely, the spatial preference of categories, is also modeled using only a few parameters to reduce the ambiguity of categories in scene images. By regarding "topics" as "categories", category discovery is thus converted to the inference of the proposed probabilistic model, which will further be addressed under a Gibbs-EM framework effectively. Experimental results on two benchmark datasets comprising MSRC-v2 and SIFT Flow show its effectiveness and the advantages comparing with other methods.

1 Introduction

Unsupervised visual category discovery has been a research hot spot in computer vision community in the past decades due to its potential uses in automated visual content summarization, scene structure mining and automatic image labeling. Its ultimate target is to recognize visually similar categories and segment out their various instances by directly mining an unlabeled image set. Indeed, many efforts [1–14] have been made to achieve this goal. Roughly, for visual category discovery, most of them use either probabilistic graphical models or any clustering method to group image patterns such as patches and regions that have similar appearance and simultaneously co-occur in images. Although these methods obtain good results in particular datasets like MSRC-v2 [15], they still face a lot of challenges especially for complex natural scene images which more likely have much variability in their appearances.

Topic models including Latent Dirichlet Allocation (LDA) [16] are a kind of generative probabilistic graphical models. They are popular in unsupervised category discovery due to the strength of *Bag of Words* representation for an image when regarding topics as categories. As known, topic models are appearance-based and ignore any extra priors like the spatial compactness of objects. Unfortunately, the main challenges of category discovery for the images captured from one natural scene generally include diversified shotting environments and complex image configurations due to occlusion, viewpoint variations and so on. Thus on one hand, scene context priors like the spatial preference of any category or category concurrence are necessary to be included to mitigate negative effects of photometry like weak illumination, shade or reflectance, and the large intra-class variability. For example, [17] introduces a context-aware topic model (CA-TM) to facilitate category discovery in natural scenes by including the s-patial preference of each category. However, the learning of the prior of spatial preferences is separated from category discovery itself in their model. On the other hand, the features of different levels theoretically need to be integrated to extend a single level representation (sparse patches) in the traditional topic model since sparse patches are essentially insufficient to discriminate different topics. For example, besides image patches, [3] first introduces region features to reduce the ambiguity and enforce the spatial coherence of topic assignments.

In this paper, after integrating the image-level GIST feature [18] into category discovery, we bring forward a novel context-aware topic model named NCA-TM, which not only makes full use of multi-level representation of images from small patches to the entire image, but also succeeds in integrating the s-patial preference of categories based on the conclusion that the GIST feature of an image can predict the location and the scale of instances of any category effectively [19]. Note that we model the prior explicitly in our graphical model by a few parameters instead of learning many global maps via complex steps as in [17]. NCA-TM assigns every region or patch a latent topic label and then derive each observation (e.g., features of patches, regions and images). In this way, category discovery is converted to the inference of NCA-TM and Gibbs-EM [20] is adopted to address it.

Our main contributions of this paper include:

1. We put forward a novel context-aware topic model by integrating multi-level image features, and
2. Spatial preference of categories is characterized in a more flexible way for assisting category discovery from complex natural scene images. The experimental results on two benchmark datasets consisting of MSRC-v2 and SIFT Flow [21] show the effectiveness of the proposed model.

The rest of the paper is organized as follows. Section 2 discusses the related work. In Section 3 we give our multi-level image representation. Section 4 shows the details of the proposed generative model for category discovery, and its inference is discussed in Section 5. Experimental results and discussions are given in Section 6, and finally Section 7 concludes the method.

2 Related work

Currently, many techniques have been exploited in unsupervised category discovery. Roughly, these methods can be categorized into two classes: generative probabilistic models and clustering-based methods. The former searches for repeated patterns from a large number of unlabeled images using the variants of topic models [1–8], while the latter groups features or image regions with similar appearance through clustering methods [9–13]. [14] gives a systematic introduction and comparative study to the earlier methods.

Generative probabilistic model. Most generative methods are extensions to a topic model. Among these methods, the most typical one is Latent Dirichlet Allocation (LDA) [1], which regards image segments as documents and categories in images as topics for object discovery. Since the traditional LDA ignores the spatial compactness of words in images, [2, 3, 7] extend it by including spatial compactness priors. Besides, [6] uses extra information of correspondences between features to improve the results. Recently, [5] adds the mutual correlation between topics and scene spatial context to facilitate visual modeling. However, scene context priors have to be learned in advance in their methods.

Clustering-based Methods. These methods are different from each other mainly on the strategies they considered to construct the similarity measurement between features or regions. For example, [11] uses link analysis on an appearance-similarity network between features and then constructs a structure-similarity matrix between features. As a result, the problem is reduced to spectral clustering to classify features belonging to the same object into the same group. Recently, [12] puts forward object-graphs to model regions, in which the regions of similar appearance and surrounding context are clustered together to form a object. They extend it to make the system automatic by searching for easy objects first and then hard objects gradually [13]. However, these methods are relatively limited to the ability to either image segmentation or features integration of different levels to compare regions. More details can be found in [14].

Additionally, there are methods [22–24] that unify category discovery and other applications simultaneously. For example, [17] considers scene labels to category discovery in order to perform scene classification. [25] integrates unsupervised object discovery and image-cosegmentation under an MRF framework to co-segment co-occurring foreground objects. However, it only segments out instances of a single foreground object. Besides, [26] models unsupervised object class discovery under an multi-instance learning framework based on saliency detection. However, they only search for object classes rather than any categories in our model. Although [27] models any categories, they focus more on co-segmentation rather than constructing appearance model of each category explicitly.

3 Image description

In our method, we represent each natural scene image I_d through three levels (See Fig. 1), namely, patch, region and image level features. Specifically, on the first level, we sample image patches P_d over the image densely and describe their appearances using SIFT. Then visual word w_{dp} is adopted to approximate the appearance of p_d by assigning it to the nearest word in the visual vocabulary pre-obtained by vector quantization of all the patch features. Simultaneously, we oversegment I_d into plenty of regions R_d to form the second level with each region r_d corresponding to a homogeneous area. Alike, we assign r_d with a visual word v_{dr} from an off-line region appearance codebook according to its feature. In addition, l_{dr} is used to represent the location of r_d corresponding to its center. There is only one image-level GIST g_d feature on the third level. To further remove redundancy, we perform PCA on traditional GIST features. Note that as the level goes up, the features pay more attention to entirety and by the contrary, more details are included into the features at the bottom level. The goal to construct the multi-level representation is to model the spatial compactness of each topic and its consistency to the spatial preference of categories in the scene.

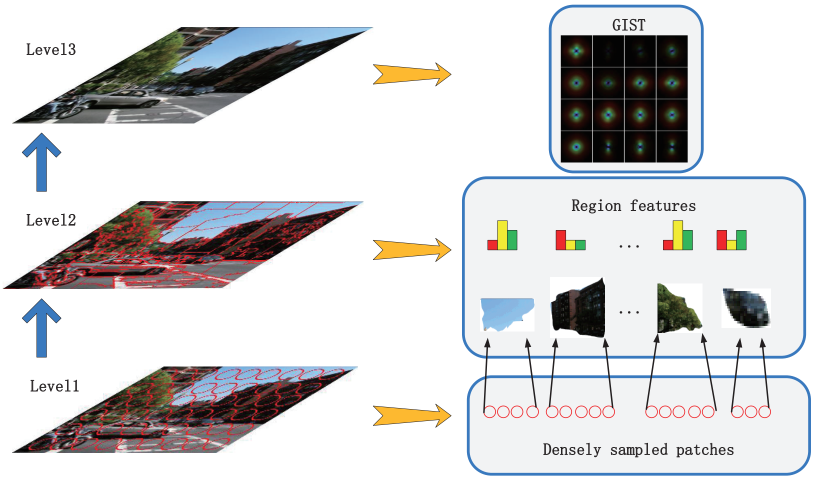


Fig. 1. Three-level image representation for an image. The first level are the features of dense image patches. The middle level corresponds to regions features and the top level has only one GIST feature of this image.

4 The proposed generative model

After the three-level representation of any image is generated, we further construct a generative probabilistic model (NCA-TM) to derive these observations. Like LDA, we regard each scene image as a document and categories in images as ”topics”. Topic proportions of images are represented by θ and for an image I_d, we need firstly sample its topic proportion $\theta_d \sim \mathrm{Dir}(\alpha)$ which governs the likelihood of each topic appearing in I_d. Thus topic labels of patches and regions are generated based on θ_d and finally all the observations of its three-level representation. To make all the parameters and the variables clear for understanding, we list all the notations in Table 1. The overall generative process is summarized in Table 2.

Table 1. Important notations in our model

Notations	Descriptions
$d = \{1, \cdots, \lvert I\rvert\}$	the index of all images I
$dr = \{1, \cdots, \lvert R_d\rvert\}$	the index of regions in R_d
$dp = \{1, \cdots, \lvert P_d\rvert\}$	the index of patches in P_d
$dt = \{1, \cdots, T\}$	the index of topics in $\boldsymbol{t_d}$
$v_{dr} = \{1, \cdots, V\}$	the visual word of dr
$w_{dp} = \{1, \cdots, W\}$	the visual word of dp
$F_d = \mu_{dt}, s_{dt} \mid \forall dt \in \boldsymbol{t_d}$	all attributes in d
μ_{dt}	the center of the topic t in d
s_{dt}	the scale of the topic t in d
l_{dr}	the region center of dr
g_d	the GIST features reduced by PCA

Latent variables	Description
$t_{dr} = \{1, \cdots, T\}$	the topic label of dr
$t_{dp} = \{1, \cdots, T\}$	the topic label of dp
$\theta_d \in [0, 1]^T$	the topic proportion of d
$\mathit{\Psi} \in \mathbf{R}^{T\times\lvert V\rvert}, \mathit{\Phi} \in \mathbf{R}^{T\times\lvert W\rvert}$	the probability of each word in each topic

Parameters	Description
$\mathit{\Omega}$	the parameters of $P(g_d \mid F_d)$

Hyperparameters	Descriptions
α, β, γ	control the prior of $P(\theta\mid\alpha), P(\mathit{\Phi}\mid\beta)$ and $P(\mathit{\Psi}\mid\gamma)$ respectively

Table 2. The generative process of our generative model

(1) For each topic t, sample $\mathit{\Phi}_t \sim \mathrm{Dir}(\beta)$ and $\mathit{\Psi}_t \sim \mathrm{Dir}(\gamma)$;
(2) For each image I_d, sample its topic proportion $\theta_d \sim \mathrm{Dir}(\alpha)$ firstly;
(3) For each region $R_r \in I_d$, sample $t_{dr} \sim \mathrm{Multi}(\theta_d)$ and then sample $v_{dr} \sim \mathrm{Multi}(\mathit{\Psi}_{t_{dr}})$;
(4) For each patch $P_p \in R_r$, sample its visual word $w_{dp} \sim \mathrm{Multi}(\mathit{\Phi}_{t_{dr}})$;
(5) Given all sampled $\boldsymbol{t_d}$, sample $g_d \sim P(g_d \mid F_d(\boldsymbol{t_d}, \boldsymbol{l_d}))$.

Generating the first two-level features. Visual words of patches and regions can be treated as words in topic models. In another word, each visual word is derived from one unique latent topic and thus they are conditionally independent. However, since any region in the second level manifests a consistent appearance within it, we enforce the patches in one region to share the same topic as the region similar to [3]. Note that the proportion of different words in any topic is particular and stable, we model them as $\mathit{\Phi}$ and $\mathit{\Psi}$ for patch words and region words, respectively. For each topic t, $\mathit{\Phi}_t$ is generated from a prior Dirichlet distribution $\mathit{\Phi}_t \sim Dir(\beta)$. Likewise, $\mathit{\Psi}_t \sim Dir(\gamma)$. Thus given a region r, we first select its topic t_{dr} via a multinomial distribution $t_{dr} \sim \mathrm{Multi}(\theta_d)$. Then its visual word v_r is sampled from $v_r \sim \mathrm{Multi}(\mathit{\Phi}_{t_{dr}})$. Simultaneously, the visual words of all the image patches in r are drawn from $\mathrm{Multi}(\mathit{\Psi}_{t_{dr}})$ one by one. Since we do not know any location priors of each topic and thus we assume $P(l_{dr} \mid t_{dr})$ is uniform.

Generating the third-level features. As a kind of image level feature, GIST is representative for image characteristics such as the occurrence, location and scale of a single topic and spatial layout among topics. Note that these abstract attributes of topics in one image is also fixed after topic labels are sampled for all regions on the second layer. Thereby, we define $P(g_d|\boldsymbol{l_d}, \boldsymbol{t_d}) \equiv P(g_d|F_d)$ for any image I_d. F_d is a series of attribute functions $\{f^c(T_d, R_d, P_d), c = 1, \cdots, |F|\}$ where T_d is a collection of all t_{dr}. In order to simplify the modeling, we use only independent attributes for each topic: location μ and scale s. When we concatenate the two attributes of all topics together, $F_d = \{\mu_{dt}, s_{dt}|\forall t \in T_d\}$. Thus we sample g_d from $P(g_d|F_d)$. If we assume $P(F_d)$ is uniform, $P(g_d|F_d) \propto P(F_d, g_d)$. Assume that μ and s are conditionally independent, the generative process can be defined by

$$P(g_d|F_d) \propto P(F_d, g_d) = \prod_{t \in t_d} P(\mu_{dt}, g_d) P(s_{dt}, g_d) \tag{1}$$

We adopt the simple generalized linear model to formulate $P(s, \boldsymbol{g})$ and $P(\mu, \boldsymbol{g})$ rather than more complex mixtures of Gaussians in [19]:

$$P(\mu_{dt}, g_d) = G(\mu_{dt}; b_t^0 + \boldsymbol{b_t}^T g_d, \sigma 1_t^2) \tag{2}$$

$$P(s_{dt}, g_d) = G(s_{dt}; q_t^0 + \boldsymbol{q_t}^T g_d, \sigma 2_t^2) \tag{3}$$

where $\mathit{\Omega} = \{b_t^0, q_t^0, \sigma 1_t^2, \sigma 2_t^2, \boldsymbol{b_t}, \boldsymbol{q_t}|\forall t \in T\}$ are model parameters to learn.

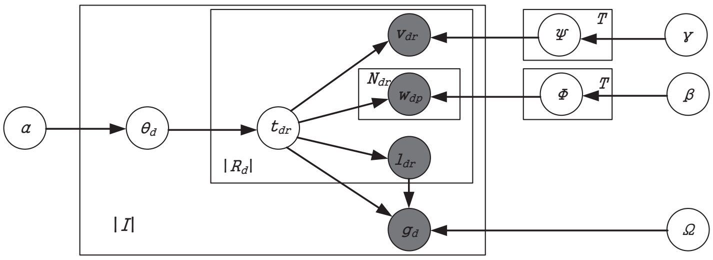


Fig. 2. The proposed graphical model.

Overall, the graphical model of our context-aware generative model is shown in Fig. 2. Given corresponding parameters, the joint distribution of all the variables can be obtained by

$$\begin{aligned} P(\boldsymbol{w}, \boldsymbol{v}, \boldsymbol{l}, \boldsymbol{g}, \boldsymbol{t}) = \int_\theta \int_{\Phi, \Psi} \prod_k P(\Psi_k|\gamma) P(\Phi_k|\beta) \prod_d^{|I|} P(\theta_d|\alpha) \prod_r^{|R|} P(t_{dr}|\theta_d) P(v_{dr}|\Psi_{t_{dr}}) \\ P(l_{dr}|t_{dr}) \prod_{p \in r} P(w_{dp}|\Phi_{t_{dr}}) P(g_d|F(\boldsymbol{t_d}, \boldsymbol{l_d})) d\theta d\Psi d\Phi \end{aligned} \tag{4}$$

5 Model learning and inference

The goal of category discovery corresponds to the inference of the graphical model, namely, maximizing the posterior distribution of latent variables given all observations $P(\boldsymbol{t}|\boldsymbol{w}, \boldsymbol{l}, \boldsymbol{g}, \boldsymbol{v}; \Omega, \alpha, \beta, \gamma)$. Different from the traditional topic model, we also need to estimate the parameters Ω during the inference. Thereby, we adopt a Gibbs EM algorithm [20] to the model inference and parameter learning. The main difference between our method and the typical EM is that Gibbs EM uses Gibbs sampling to estimate the posterior distribution in the E-step. Likewise, the E-step and M-step are interleaved and process iteratively into convergence.

E-step: In the E-step, by integrating out θ, Φ and Ψ, t_{dr} can be sampled from a Gibbs sampling procedure. The distribution of t_{dr} conditioned on $\boldsymbol{t_{-dr}}$ is

$$P(t_{dr} = k|\boldsymbol{t_{-dr}}, \boldsymbol{w}, \boldsymbol{v}, \boldsymbol{l}, \boldsymbol{g}; \Omega, \alpha, \beta, \gamma) \propto (n_{d,(\cdot)}^{k,-dr} + \alpha_k) \frac{n_{(\cdot),v_{dr}}^{k,-dr} + \beta_v}{\sum_{c=1}^{V} n_{(\cdot),c}^{k,-dr} + \beta_c} \frac{\prod_{w_{dp}, p \in R_{dr}}^{W} A_{m_{w_{dp},k}^{(\cdot)} + \gamma_{w_{dp}} - 1}^{m_{w_{dp},(\cdot)}^{dr}}}{A_{m_{(\cdot),k}^{(\cdot)} + \gamma - 1}^{m_{(\cdot),(\cdot)}^{dr}}} P(g_d|\boldsymbol{t_d}, \boldsymbol{l_d}; \Omega) \tag{5}$$

where $m_{w,k}^{r}$ denotes the number of patches in the region r with the visual word w and the topic label k. $n_{d,v}^{k,-r}$ represents the number of regions in the image d with the visual word v and their topic labels equal to k except the region r. If any dimension is not limited to some specific value, we use to $(\cdot)$ to replace it. A is the P-permutation operator.

M-step: In the M-step, we need to estimate Ω based on sampled $\boldsymbol{t}$ in the E-step. Firstly, for each topic t appearing in any image I_d, μ_{dt} are calculated following $\mu_{dt} = \frac{1}{N_{dt}} \sum_{t_{dr}=t} l_{dr}$ and alike, $s_{dt} = \frac{1}{N_{dt}} \sum_{t_{dr}=t} (l_{dr} - \mu_{dt})^2$. N_{dt} is the number of regions in I_d with their topic labels equal to t. As [19] validated, GIST feature is not sensitive to horizontal locations of topics. Thus we only calculate the two attributes along the vertical direction. When all u_{dt} and s_{dt} are calculated, the corresponding parameters Ω can be obtained by

$$\begin{aligned} &B = (U^T U) U^T G \qquad Q = (S^T S) S^T G \\ &\Sigma 1 = (U - GB^T)^T (U - GB^T) \\ &\Sigma 2 = (S - GQ^T)^T (S - GQ^T) \end{aligned} \tag{6}$$

where U and S are two $|D| \times T$ matrices with each element corresponding to u_{dt} and s_{dt}, respectively. G is a matrix with each row corresponding to g_d. $B, Q, \Sigma 1 \text{and} \Sigma 2$ are model parameters with each row representing $\boldsymbol{b}, \boldsymbol{q}, \sigma 1$ and $\sigma 2^2$ of each t, respectively. Experimental results show the simple generalized linear model also functions well to model the spatial preference of categories.

Simultaneously, after the Gibbs EM framework are converged, we can also get the characteristic visual word distributions Ψ and Φ. Since Ψ and Φ are conditionally independent on the samples of $\boldsymbol{t}$, the evaluation of Ψ and Φ are not correlated. Thus they can be estimated as in the traditional LDA

$$\Phi_k^w = \frac{m_{w,k}^{(\cdot)} + \beta}{m_{(\cdot),k}^{(\cdot)} + W\beta} \quad \Psi_k^v = \frac{n_{(\cdot),v}^{k,(\cdot)} + \gamma}{n_{(\cdot),(\cdot)}^{k,(\cdot)} + V\gamma} \tag{7}$$

6 Experiments

6.1 Datasets

We evaluate our methods on two datasets: MSRC-v2 and SIFT Flow. MSRC-v2 dataset has altogether 21-categories (591 images). SIFT Flow datasets including images from 8 natural scene classes with 2688 images (256 × 256) and 33 categories in all the images totally. SIFT Flow is selected following two considerations: 1) all the images are captured in natural scenes with relatively stable context in each scene, and 2) all the images have pixel-level ground-truth labels. Note that all the images in MSRC-v2 are also resized into 256 × 256.

6.2 Experimental settings

In order to generate three-level representation for an image, we first sample 12 × 12 patches densely in the image with step 3 pixels and then extract their SIFT features. These features are then vector quantized to form a codebook of size 500 using K-means. SLIC [28] is used to generate homogeneous regions for each image with the initial region-size 30 and then for each region, we extract its texture feature with 40 dimensions using the same filter bank as [12] and its 3-dimension color. Likewise, a region color codebook of size 20 and a texture codebook of size 200 are obtained by clustering all region color features and texture features, respectively. Additionally, for the third level GIST feature, we apply PCA to reduce the typical GIST feature of 512 dimensions to 64 to prevent overfitting of our selected generalized linear model. As to hyperparameters, we set $\alpha = 50/T, \beta = 200/W$ and$\gamma = 200/V$.

6.3 Evaluation Metrics

As [12] also states, it is difficult to ensure what each topic corresponds to. In another word, it may represent either an semantic category or a part of any category such as the window of any building. Thus without semantic information, it is difficult to evaluate it in the same way as image segmentation or labeling with supervised assistance, especially for the practical case that we don't know the topic number in advance. Thereby, we adopt the *purity* score to measure the coherence of topic assignments to pixels. Note that a higher *purity* score indicates that topic assignments are more consistent with ground-truth labels.

6.4 Evaluation and results

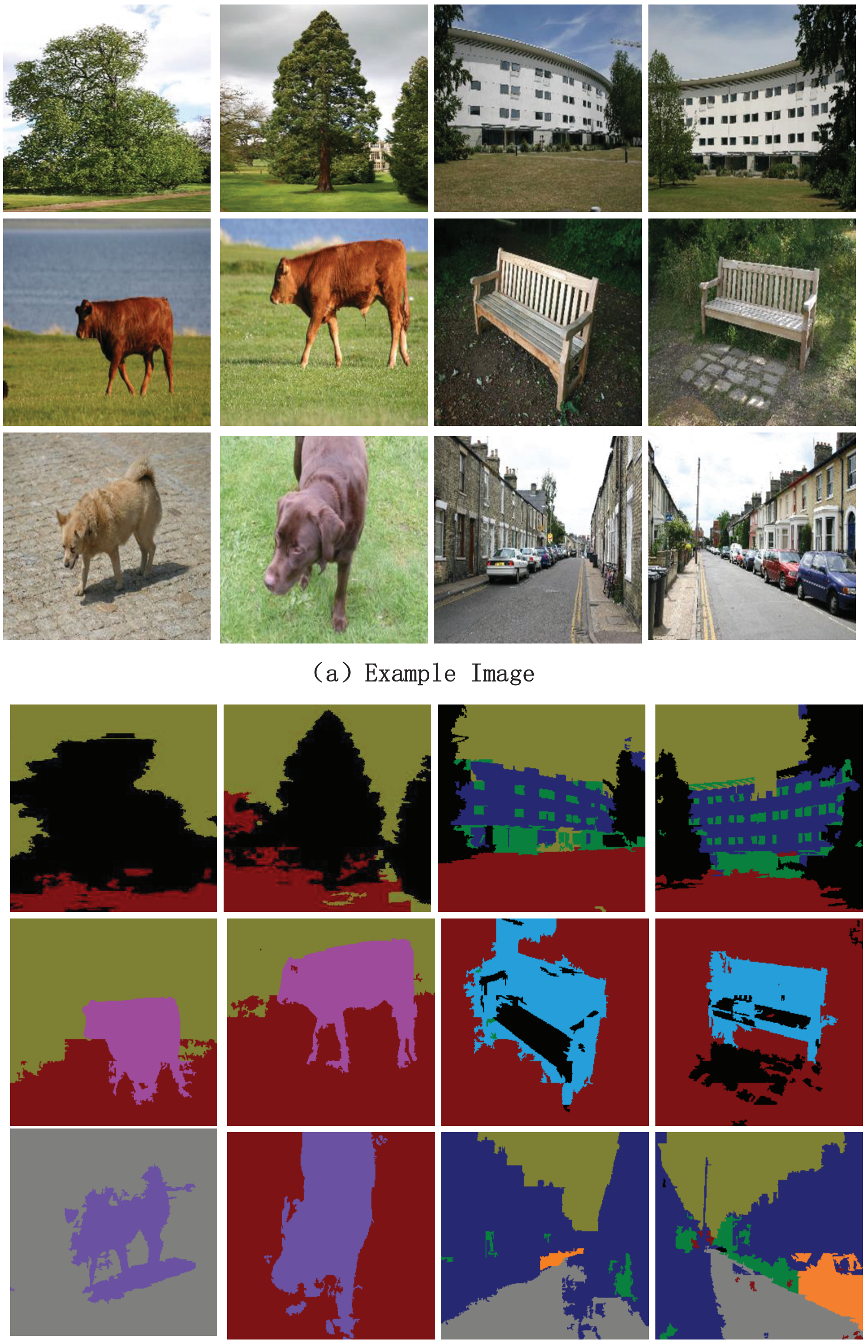
Without smoothing topic assignments, several example raw results for MSRC-v2 and SIFT Flow are shown in Fig. 3 and Fig.4, respectively. We find that most of the categories in the images are distinguished from each other despite some noise, which can be removed in any practical application. Note that windows are labelled as a different topic from that of buildings since there indeed exists semantic gap. Since the best topic number is not known to us and thus we report *purity* scores with different topic numbers. We find in Fig. 5 that for the two datasets, as the topic number arise, the performance become better and get the best performance within the interval $[1, 1.5] \times N$ where N is the ground-truth category number. Besides, it is not beneficial to our model that the topic number is too large.

In order to validate the effectiveness of modeling GIST and its implying scene context in NCA-TM, we conduct comparative study with our modified version (a), and the spatial LDA method [2] (b) on MSRC-v2 and SIFT Flow. In (a), we delete the variable g and related edges. Essentially, (a) is unsupervised spatial-LTM [3] because $P(l|z)$ is uniform in our model. Spatial LDA (b) only adds the prior that the patches of the same topic should be close and no any scene context prior is included. From Fig. 5, we can see our modeling of GIST has two different impacts for MSRC-v2 and SIFT Flow. Our model is inferior compared to (a) and (b) in MSRC-v2 and gets the best performance in SIFT Flow by the contrary. It is observed that for most of images in MSRC-v2, instances of the foreground categories are likely to locate in the center. Thus there are less stable context reflected by GIST since GIST indicates locations and scales of categories by the environment around them [19]. However, the images in SIFT Flow are from natural scenes with stable scene context (See Fig. 4 for image examples). Thereby, including GIST in the dataset like MSRC-v2 to category discovery has little improvement or even inferior performance. According to the comparison, we conclude that our modeling of GIST in NCA-TM is effective and necessary for natural scenes.

Simultaneously, we further compare the method (c) in [17] to show our modeling of GIST is more flexible and effective than their global location maps. Note that the method is intended to scene classification and it is an extension of DISC-LDA [24] to include global contexts and semantic labels. Thus we only simplify it by replacing DISC-LDA by spatial-LTM and the modeling of their global context does not change. The comparative results are shown in Fig. 5 and we find in SIFT Flow, our performance is superior to (c) obviously. The results further validate the intuition that GIST is effective and flexible to model scene contexts and category discovery in natural scenes can benefit from our model.

The modeling of the category spatial preference prior into category discovery is indeed explored earlier. However, our model moves forward by considering them as specific cases of our NCA-TM. Thereby, the proposed model is essentially a more generalized framework:

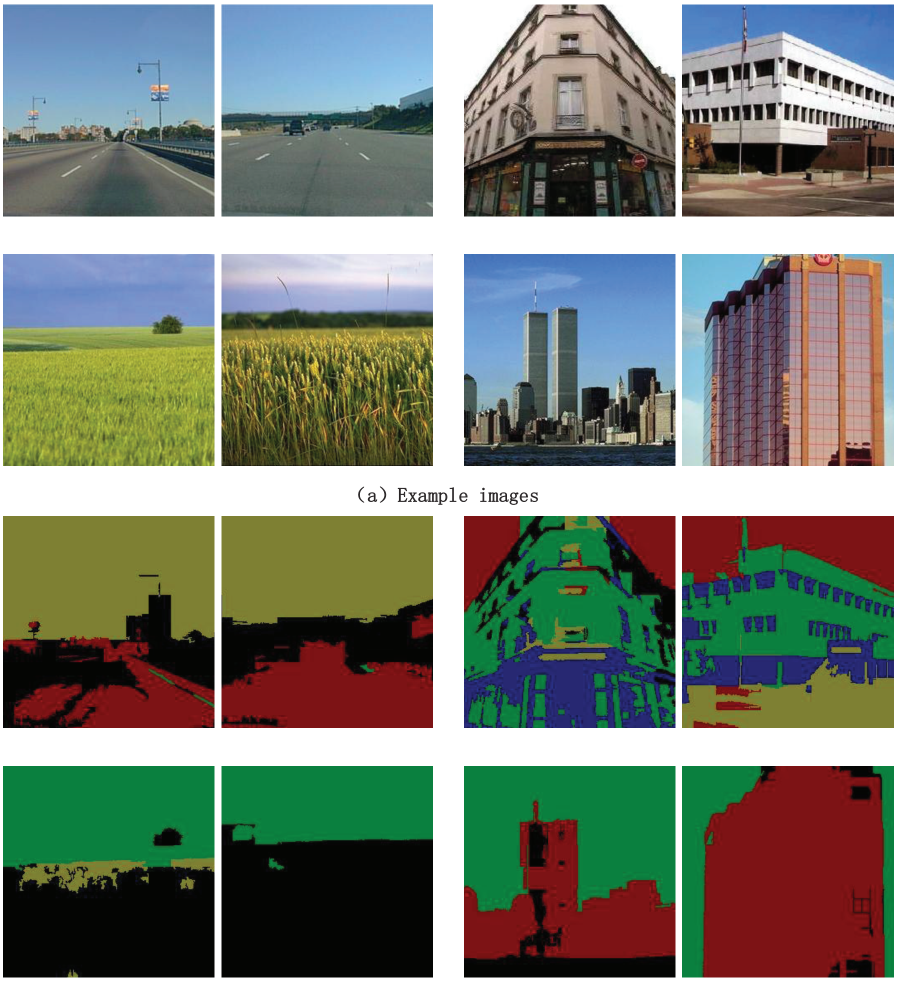
1. If we remove GIST features, namely the node of g and the related nodes of parameters in the graphical model, the model will be reduced to [7] where



(a) Example Image

(b) category discovery results

Fig. 3. Example category results of MSRC-v2. Different colors in category discovery results represent different topics.



(a) Example images

(b) Category discovery results

Fig. 4. Category results of 4 scenes in SIFT Flow with two examples for each scene. Different from the experiments on MSRC-v2, we perform NCA-TM on 8 scenes respectively rather than all scene images together. Different colors also represent different topics. However, colors of different scene examples are independent.

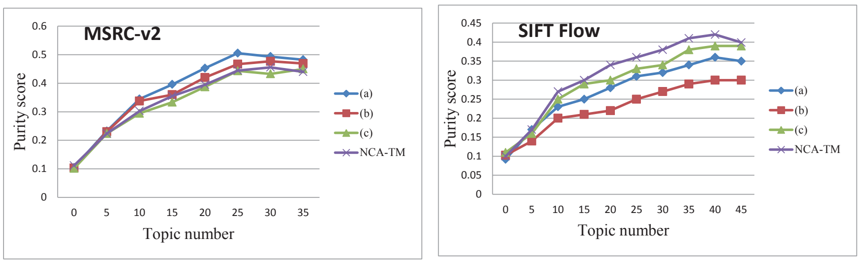


Fig. 5. Purity scores of MSRC-v2 and SIFT-Flow at different topic numbers.

they use Gaussian to model $P(\boldsymbol{l_d}|\boldsymbol{t_d})$ rather than our uniform assumption. Thereby, their model can only enforce the same topic to the patches that are spatially compact.

2. If we replace GIST features with many global location maps of topics, namely cutting the image-related relation between $\boldsymbol{t_d}$ and $\boldsymbol{g_d}$, the model will be similar to [17] for scene recognition. As known, global location map of categories is unstable and does not make use of valuable image-specific information.

Therefore, our model is an generalization to the existing models to use scene context prior and the spatial preference of categories for category discovery in natural scene images. Experimental results show our model outperforms these models especially in natural scene images.

To conclude, our model succeeds in modeling the scene context prior, the spatial preference of categories, and integrate multi-level features in a flexible and effective way and it is better for category discovery in natural scene images.

7 Conclusion

In this paper, we propose a novel context-aware topic model to make use of multi-level features and scene context priors to facilitate category discovery in natural scenes by a flexible and effective way. The model constructs a generative probabilistic procedure for all the three-level features by regarding ”topics” as ”categories”. Category discovery corresponds to the inference of the model, which is addressed under the Gibss-EM framework. Experimental results show its effectiveness and the advantage in natural scenes. For future work, we will focus on modeling more complex and accessible scene contexts into category discovery in natural scenes.

Acknowledgement.

References

1. Russell, B.C., Freeman, W.T., Efros, A.A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR (2). (2006) 1605–1614
2. Wang, X., Grimson, E.: Spatial latent dirichlet allocation. In: NIPS. (2007)
3. Cao, L., Li, F.F.: Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In: ICCV. (2007) 1–8
4. Zhao, B., Li, F.F., Xing, E.P.: Image segmentation with topic random field. In: ECCV (5). (2010) 785–798
5. Lin, D., Xiao, J.: Characterizing layouts of outdoor scenes using spatial topic processes. In: ICCV. (2013) 841–848
6. Liu, D., Chen, T.: Unsupervised image categorization and object localization using topic models and correspondences between images. In: ICCV. (2007) 1–7
7. Liu, D., Chen, T.: Semantic-shift for unsupervised object detection. In: CVPR. (2006)
8. Fergus, R., Li, F.F., Perona, P., Zisserman, A.: Learning object categories from internet image searches. Proceedings of the IEEE **98** (2010) 1453–1466
9. Lee, Y.J., Grauman, K.: Shape discovery from unlabeled image collections. In: CVPR. (2009) 2254–2261
10. Lee, Y.J., Grauman, K.: Foreground focus: Unsupervised learning from partially matching images. International Journal of Computer Vision **85** (2009) 143–166
11. Kim, G., Faloutsos, C., Hebert, M.: Unsupervised modeling of object categories using link analysis techniques. In: CVPR. (2008)
12. Lee, Y.J., Grauman, K.: Object-graphs for context-aware visual category discovery. IEEE Trans. Pattern Anal. Mach. Intell. **34** (2012) 346–358
13. Lee, Y.J., Grauman, K.: Learning the easy things first: Self-paced visual category discovery. In: CVPR. (2011) 1721–1728
14. Tuytelaars, T., Lampert, C.H., Blaschko, M.B., Buntine, W.L.: Unsupervised object discovery: A comparison. International Journal of Computer Vision **88** (2010) 284–302
15. Shotton, J., Winn, J.M., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International Journal of Computer Vision **81** (2009) 2–23
16. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
17. Niu, Z., Hua, G., Gao, X., Tian, Q.: Context aware topic model for scene recognition. In: CVPR. (2012) 2743–2750
18. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision **42** (2001) 145–175
19. Torralba, A.: Contextual priming for object detection. International Journal of Computer Vision **53** (2003) 169–191
20. Andrieu, C., de Freitas, N., Doucet, A., Jordan, M.I.: An introduction to mcmc for machine learning. Machine Learning **50** (2003) 5–43
21. Tighe, J., Lazebnik, S.: Superparsing - scalable nonparametric image parsing with superpixels. International Journal of Computer Vision **101** (2013) 329–349
22. Su, H., Sun, M., Li, F.F., Savarese, S.: Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. In: ICCV. (2009) 213–220

23. Li, L.J., Socher, R., Li, F.F.: Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In: CVPR. (2009) 2036–2043
24. Niu, Z., Hua, G., Gao, X., Tian, Q.: Spatial-disclda for visual recognition. In: CVPR. (2011) 1769–1776
25. Rubinstein, M., Joulin, A., Kopf, J., Liu, C.: Unsupervised joint object discovery and segmentation in internet images. In: CVPR. (2013) 1939–1946
26. Zhu, J.Y., Wu, J., Wei, Y., Chang, E.I.C., Tu, Z.: Unsupervised object class discovery via saliency-guided multiple class learning. In: CVPR. (2012) 3218–3225
27. Yuan, Z., Lu, T., Shivakumara, P.: A novel topic-level random walk framework for scene image co-segmentation. In: Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I. (2014) 695–709
28. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34** (2012) 2274–2282