Scene Categorization Using Topic Model Based Hierarchical Conditional Random Fields

Vikram Garg, Ehtesham Hassan, Santanu Chaudhury, and M. Gopal

Department of Electrical Engineering, IIT Delhi, India
gargvikram07@gmail.com, hassan.ehtesham@gmail.com,
santanuc@ee.iitd.ernet.in, mgopal@ee.iitd.ac.in

Abstract. We propose a novel hierarchical framework for scene categorization. The scene representation is defined by latent topics extracted by Latent Dirichlet Allocation. The interaction of these topics across scene categories is learned by probabilistic graphical modelling. We use Conditional Random Fields in a hierarchical setting for discovering the global context of these topics. The learned random fields are further used for categorization of a new scene. The experimental results of the proposed framework is presented on standard datasets and on image collection obtained from the internet.

Keywords: Scene categorization, Latent dirichlet allocation, Conditional random fields.

1 Introduction

In this paper we propose a novel hierarchical framework for scene categorization. The framework utilizes the latent concept information for scene categorization. The concept extraction is performed by latent topic modelling at the local patch level. For similar scenes, the organization of these topics represents similar graphical structure. We learn the graphical structure of these topics across various scene categories through a probabilistic graphical model. The framework recognizes a new scene based on the learned graphical structures.

The natural scene categorization is a widely researched problem in content based image retrieval. The fundamental problem here lies in identifying different objects and understanding their interaction for scene definition. However, the object extraction in a natural scene requires robust segmentation algorithm. Topic modelling based methods provide efficient solution to this problem. These methods represent the image through mixture of latent topics. The topic assignment effectively performs the semantic image segmentation in feature space. In the related works, Yamaguchi and Maruyama have used topic distribution based image representation for image classification [4]. The topic modelling is performed by Probabilistic Latent Semantic Analysis using the SIFT key points based bag-of-words model. In another work, Ergul and Arica perform the pLSA based topic analysis of the image at multiple scales [5]. The topic modelling by

S.O. Kuznetsov et al. (Eds.): PReMI 2011, LNCS 6744, pp. 206–212, 2011.

PLSA shows biased topic assignment towards training images. The LDA incorporates the document distribution for topic learning and provides robust topic modelling framework. [7] have learned a LDA based hierarchical model for image categorization while [11] have employed a generative hierarchical LDA model for unsupervised discovery of topic hierarchies in visual objects. The recent work Wang et al. have applied supervised LDA for discovering the latent topics [13]. The contextual relationship between the topics is important semantic information for scene categorization. [12] has used a generative Bayesian network to utilize the position and expected appearance of object parts while in the present work, we propose a novel scene categorization framework by exploiting the topic level contextual relationship in a natural image. The framework extracts the local patch level topics in scene using LDA. The probabilistic graphical models exhibit excellent capability to learn the sequential data. The recent works have extensively applied Conditional Random Fields (CRFs) based graphical model for Image classification [9][10][13]. However discovering the contextual relationship between the topics by CRFs is not yet explored. Our work presents a novel approach in this direction by applying hierarchical CRF to discover the topic based global context for scene recognition.

The organization of the paper is as follows: The section 2 presents scene representation patch level topic modelling of the image. The section 3 presents the proposed framework for scene categorization. The section 4 presents the experimental results. Finally we conclude and present perspective of our work.

2 Feature Based Image Representation Using LDA

The visual content representation forms important step for a scene understanding. The bag-of-words model has been the preferred approach for natural image representation. Using the model, image representation is defined as the distribution of local features properties. The SIFT based descriptor provide an efficient approach to extract the local image features [1]. The descriptor is computed as set of local orientation histograms centred at key-points obtained by multi-scale analysis of the image. The SIFT descriptor extracts dense points at the boundaries, and image segments having sharp changes. Whereas the smooth segments are represented by sparse key points. The bag-of-words model in this case represents the image by a biased representation leaving significant amount of visual information unutilized. For example, a landscape scene having sky tree and building will get biased distribution of key-points near tree and building. Therefore we follow grid based approach for extracting the local image features. The approach covers all the segment of the image and assigns uniform weight to all the segments. Using the image patches in the grid, bag-of-words model is generated. The bag-of-words representation is used to generate the image representation by discovering the latent topics in a scene. The topics here represent the semantic entity which defines the distribution of various image patches in the latent space. The Latent Dirichlet allocation is applied to extract the topic distribution over the scene image collection. The LDA is defined as a generative probabilistic model for collections of documents [2]. In the context of topic

modelling, the documents could be a text corpus or image collection. The local image properties are defines as words of the image document. The documents are represented as as random mixtures over latent topics, where each topic is characterized by a distribution over words. The generative process for each document $\mathbf{w}$ in the collection D is defined as follows:

- Select $N = Poisson(\zeta)$
- Choose $\theta = Dirichlet(\alpha)$
- For each word w_n of the document $\mathbf{w}$: select a topic $z_n = Multinomial(\theta)$, select w_n from multinomial probability $p(w_n|z_n;\beta)$

The inference step includes the computation of posterior distribution of the hidden variables for a given document.

$$p(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)} \tag{1}$$

The intractable form of the above distribution is because of complex prior distributions. For the practical applications approximate inference based methods. The LDA assigns a topic distribution to each image patch (figure (1 shows topics on face image using color based bag-of-words model). The scene representation is defined as the vector of topic distributions which is subsequently used for categorization.

Fig. 1. Topic assignment over image patches

3 Hierarchical Framework for Scene Categorization Using CRFs

The figure (2) shows the proposed hierarchical framework for scene categorization. The initial step performs the topic analysis of the image at local patch level. The step performs robust topic assignment to all the image patches. The topic assignment essentially performs the coarse semantic segmentation of scene for the given number of topics [8]. The spatial context between these topics can be

efficiently learned by a probabilistic graphical model. The CRFs are an efficient tool for such problems. However, the long range contextual dependencies in the semantic segmentation cannot be captured by a simple graphical model. Therefore, we introduce category specific graphical model (CRFs). We train category specific CRFs using the patch-ground-truth, and assign semantic labels over the latent topics. The exact notion of the category specific CRFs is the smoothening of latent topics by supervised learning. The top level CRFs learns the structure of the contextual graph belonging to each category of images.

The CRFs is a discriminative modelling tool for segmenting and labelling the sequential data [3]. The CRFs work under the maximum entropy principle and observe the features as sequence data. Consider X the random variable over data sequences to be labelled and Y the random variable for the corresponding label sequences. Using the fundamental theorem of random fields, joint distribution over the label sequence Y given X is defined as

$$p_\theta(y|x) \propto exp(\sum_{e \in E,k} \lambda_k f_k(e, y|_v, x) + \sum_{v \in V,k} \mu_k g_k(v, y|_e, x)) \tag{2}$$

V and E represent the vertices and edges of graph G, such that label sequence Y is indexed by the vertices of G. Here x represents a data sequence, y represents a label sequence, and $y|_S$ is set of components of y associated with the vertices in sub-graph S. The function f_k defines the input dependent evidences and g_k g_k represents the pair-wise coupling labels of sequence data. The parameter estimation problem determines the parameters $\theta = (\lambda_1, \mu_1, \lambda_2, \mu_2, ...)$ by the maximization of log-likelihood objective as

$$O(\theta) = \sum_{i=1}^{N} log\ p_\theta(y^i|x^i) \propto \sum_{x,y} p(x, y) log\ p_\theta(y|x) \tag{3}$$

For category specific CRFs, graph G is represented by function f_k as the likelihood of image patch having i^{th} topic and function g_k represents smoothness prior to encourage the neighbouring image patches to obtain same label. Function f_k has the form as $f_k(e, y|_v, x) = x_k \delta(y, l_k)$, δ is the Kronecker delta and k represents the indices of parameter set from θ and l_k is the label of the patch for kth parameter set. The smoothness prior g_k is defined as $\delta(y_1, l_k)\delta(y_2, l_k)$. It is clear that the smoothness prior is independent of patch features. The top level CRF learns the graphical structure of latent topics for image collection using the sequence of semantic labels assigned from first level CRFs.

A new image is categorized by topic assignment using patch level image features. The topic level representation of the image is supplied to the entire category specific CRFs. For N scene categories, the top level CRF receives N semantic graphs for the new image. For each semantic graph, the top level CRF assigns marginal probability distribution. The final category assignment for the new scene is done as the label corresponding to maximum marginal probability for all the semantic graphs.

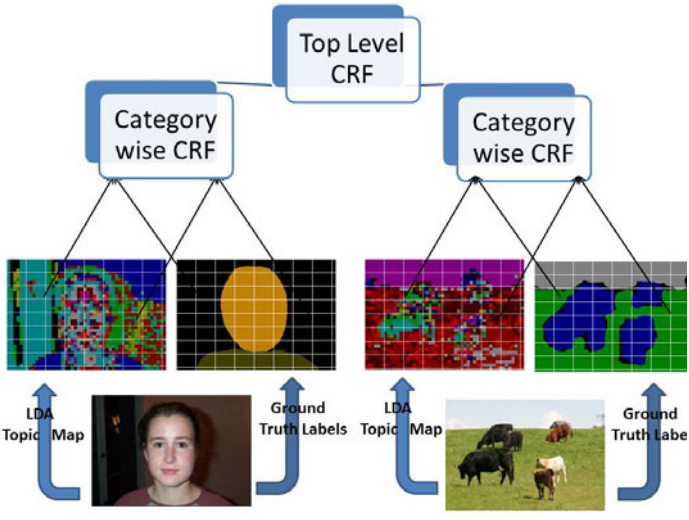

Fig. 2. Hierarchical framework for scene categorization

4 Experimental Results and Discussion

We have evaluated the proposed framework on standard MSRC dataset[1], Caltech 101 dataset[2] and image collection from internet. The MSRC dataset is applied for training our system as the dataset contains pixel level ground-truth. The pixel-level ground truth is converted to patch-level by majority voting. The training set consists of 30 randomly selected corresponding to 9 super categories {Bicycle, Building, Car, Cattle, Cow, Face, Sheep, Plane, and Tree}. The image set (270 images) have been normalized to 105X154. The images patches are detected by overlaying a grid of 15×22, therefore giving 330 patches of 7X7. The bag-of-words model is generated by clustering the image patches from 5 randomly selected training images corresponding to each category. The topic assignment over complete set of image patches is done by learning a LDA for 10 semantic concepts. We use topic modelling toolbox provided by Steyers and Griffiths [14] for topic modelling. The probabilistic graphical modelling is performed by CRF implementation provided by Sarawagi [15]. Initially the framework has been tested for the same set of training images. We have compared our results with the best case accuracy preseted in [8] in table 1. The second evaluation the proposed framework is performed using the Caltech 101 dataset images. The experiment evaluates the proposed framework for unseen images belonging to two categories. The experiment evaluates the robustness of the proposed framework as the dataset is primarily prepared for object recognition problem. The experiment is performed by randomly selecting 30 images corresponding to *Plane*

[1] http://research.microsoft.com/en-us/projects/objectclassrecognition/
[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101/

Table 1. Recognition accuracy for training image set

Category	Bicycle	Building	Car	Cattle	Cow	Face	Sheep	Plane	Tree	Average
Presented Acc.	90	96	93	90	93	96	96	96	96	94
Best Average Acc. [8]	94	77	73	N.A.	86	99	N.A.	73	71	82

*categories with N.A. were not availabe in [8].

Table 2. Recognition accuracy for image set from Caltech dataset

Category	Plane	Face
Accuracy	67	45

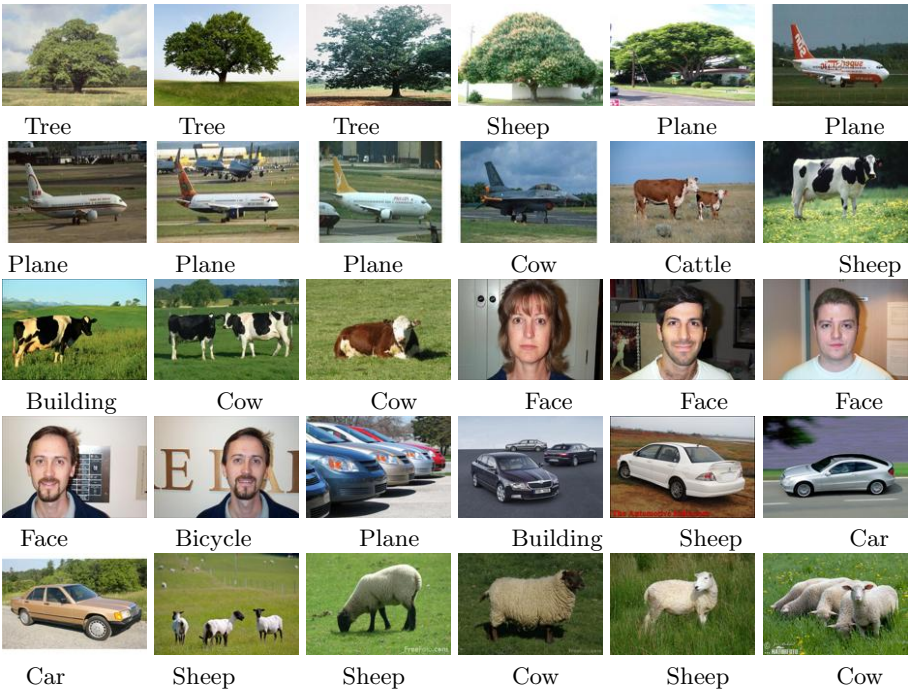

Fig. 3. Evaluation Images and assigned category

and 30 images corresponding to *Face* category. The results are presented in the table 2. The third evaluation of the proposed framework is performed on image collection obtained from internet. The collection contains 30 images belonging to 6 training categories. The average accuracy of 60% is achieved. The image with their final category assignment are shown in the figure 3. The results (figure 3) establish the robustness of the proposed framework as testing images belong to variety of domains. The category assignment for these images can be easily

verified by close observation. The results show that our framework efficiently utilizes semantic information of the scene for categorization.

5 Conclusion

A novel framework for scene categorization is presented. The patch level latent topics extracted by LDA have been used for scene representation. A novel application of hierarchical CRFs is demonstrated for scene recognition. The two-layer CRFs performs topic level smoothening at first layer and subsequently perform the scene recognition by learning the spatial context of different topics. The experimental evaluation on standard dataset and images obtained from the internet validate the efficiency and robustness of the proposed framework.

References

1. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the International Conference on Computer Vision, vol. 2, pp. 1150–1157 (2000)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research, 993–1022 (2003)
3. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labelling sequence data. In: Proceedings of the International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann, San Francisco (2001)
4. Yamaguchi, T., Maruyama, M.: Image categorization by a classifier based on probabilistic topic model. Pattern Recognition (2008)
5. Ergul, E., Arica, N.: Scene Classification Using Spatial Pyramid of Latent Topics. In: International Conference on Pattern Recognition (2010)
6. Fei, L.F., Persona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
7. Wang, C., Blei, D., Fei, L.F.: Simultaneous Image Classification and Annotation. In: IEEE Conference on Computer Vision and Pattern Recognition (2009)
8. Passino, G., Patras, I., Izquierdo, E.: Latent Semantic Local Distribution for CRF-based Image Semantic Segmentation. In: British Machine Vision Conference (2009)
9. Zhong, P., Wang, R.: Learning Conditional Random Fields for Classification of Hyperspectral Images. In: IEEE Transactions on Image Processing (2010)
10. Wang, X., Liu, X., Shi, Z., Shi, Z., Sui, H.: Voting Conditional Random Fields for Multi-label Image Classification. In: International Congress on Image and Signal Processing (2010)
11. Sivic, J., Russell, B.C., Zisserman, A., Freeman, W.T., Efros, A.A.: Unsupervised discovery of visual object class hierarchies. In: Computer Vision and Pattern Recognition (2008)
12. Sudderth, E.B., Torralba, A., Freeman, W.T., Willsky, A.S.: Learning Hierarchical Models of Scenes, Objects, and Parts. In: IEEE International Conference on Computer Vision (2005)
13. Wang, Y., Gong, S.: Conditional Random Field for Natural Scene Categorization. In: British Machine Vision Conference (2007)
14. Steyvers, M., Griffiths, T.: Matlab Topic Modelling Toolbox Version 1.3.2
15. Sarawagi, S.: CRF package, `http://crf.sourceforge.net/`