LETTER

Multi-Scale Multi-Level Generative Model in Scene Classification

Wenjie XIE[†a)], ***Student Member***, **De XU**[†], ***Nonmember***, **Yingjun TANG**[†], ***Student Member***, ***and*** **Geng CUI**[††], ***Nonmember***

SUMMARY Previous works show that the probabilistic Latent Semantic Analysis (pLSA) model is one of the best generative models for scene categorization and can obtain an acceptable classification accuracy. However, this method uses a certain number of topics to construct the final image representation. In such a way, it restricts the image description to one level of visual detail and cannot generate a higher accuracy rate. In order to solve this problem, we propose a novel generative model, which is referred to as multi-scale multi-level probabilistic Latent Semantic Analysis model (msml-pLSA). This method consists of two parts: multi-scale part, which extracts visual details from the image of diverse resolutions, and multi-level part, which concentrates multiple levels of topic representation to model scene. The msml-pLSA model allows for the description of fine and coarse local image detail in one framework. The proposed method is evaluated on the well-known scene classification dataset with 15 scene categories, and experimental results show that the proposed msml-pLSA model can improve the classification accuracy compared with the typical classification methods.

key words: *scene classification, msml-pLSA, visual granularity*

1. Introduction

Scene classification is one of most challenging problems in computer vision, especially in the presence of intra-class variations, clutters, partial occlusions, pose changes, changes in viewpoint and illumination. Furthermore, as a scene composed of several entities is often organized in an unpredictable layout, scene classification is much more difficult than conventional object classification and has received considerable attention in recent years.

There are two basic strategies to implement scene classification. The first strategy uses low-level global features extracted from the whole image, such as color, texture, edge response, gradient, power spectrum, etc, to classify images into a small number of categories [4]. This may be sufficient for separating scenes with significant differences in the global properties. However, if the images of different categories (e.g. office vs. living room) have the similar low-level global features, the global features may not be discriminative enough. The second strategy employs quantized local invariant features to model image. Sivic et al. [5] originally proposed to cluster the low-level visual features using K-means algorithm to construct codebook, where each centroid corresponds to a visual word. When building a histogram, each feature vector is assigned to its closest centroid. The codebook describes an image as a bag of discrete visual words and the frequency distributions of visual words in an image allow classification. This method is usually referred to as Bag-of-Visterm (BoV) model.

Although the BoV model can generate an acceptable classification accuracy, high dimensional feature vector and visual words without semantics make the classification task quite challenging. Anna Bosch [1] introduced a new classification algorithm based on a combination of unsupervised probabilistic Latent Semantic Analysis (pLSA) followed by a nearest neighbor classifier. The pLSA model is a generative model from the statistical text literature [2]. In text analysis, this is used to discover topics in a document using the bag-of-words document representation. In scene classification, we regard images as documents, and find topics as object categories. In such a way, an image containing instances of several objects will be modeled as a mixture of topics.

When we perform classification using the pLSA model, we should retain as much information as possible from the topic representation. One question should be given close attention, that is, how to choose appropriate number of topics for a certain dataset in the pLSA model. Generally speaking, images from different categories usually contain diverse number of instances of objects, which means different numbers of topics are needed to represent scenes in finer and coarser granularity. However, the pLSA model uses a certain number of topics to create the final image representation, which results in restricting the image description to one level of visual detail and generating lower accuracy rate. In this paper, we present a novel approach called multi-scale multi-level pLSA (msml-pLSA) to solve this problem. In our proposed method, image is firstly decomposed into multiple scales layers to generate the multi-scale histogram, and then diverse numbers of topics are employed successively to form multiple topic representations. Finally, we concentrate these representations corresponding to diverse numbers of topics in sequence to form the final representation. Our method can create a more complete representation of the image due to the inclusion of the fine and coarse visual detail information in a joint approach. In experiments, we evaluate the proposed msml-pLSA method on the 15-category dataset, and experimental results validate the superiority of

Manuscript received July 15, 2010.
Manuscript revised September 17, 2010.

[†]The authors are with Institute of Computer Science and Engineering, Beijing Jiaotong University, Beijing, China.
[††]The author is with Institute of Orthopedics, General Hospital of PLA, Beijing, China.
a) E-mail: xiewenjiebj@126.com

DOI: 10.1587/transinf.E94.D.167

the proposed method.

The rest of the paper is organized as follows. Section 2 presents the proposed image modeling approach. Section 3 describes the experimental setup and provides classification results. Section 4 concludes the paper.

2. Our Approach

2.1 Overall Framework

In this section, we introduce the framework of msml-pLSA model. As shown in Fig. 1, this framework includes two parts: multi-scale part and multi-level part. In multi-scale part, image is decomposed into multiple layers and then the multi-scale histogram is constructed to represent scene in variant detail. In multi-level part, multiple pLSA models are constructed by using diverse numbers of topics and diverse topics representations are generated. Then, these topic representations are linearly concentrated to form the final representation, which refers to as multi-level histogram. The proposed framework for scene classification attempts to retain as much visual information as possible to represent scenes, that is, the image of variant resolutions and diverse numbers of topics are employed to exact visual details of diverse granularity. Next, we introduce the two parts in detail.

2.2 Multi-Scale Part

In this part, we attempt to extract visual details from the image of diverse resolutions. By taking every second pixel in each row and column, the image I is firstly decomposed into multiple layers $L = \{L_1, L_2, \cdots, L_r, \cdots, L_R\}$, where R denotes the number of layers of image I. Dense SIFT [3] is employed as local invariant feature descriptor, and K-means clustering is used to implement cluster.

In the processing of quantizing the local interest point descriptors, there is inherent weakness of vector quantization, that is, the hard assignment of discrete visual words to continuous image features. In order to solve the problem, in our paper, soft-assign method [7] is adopted to produce histogram. For each feature point in an image, instead of only searching for the nearest visual word, the top-N nearest visual words are selected to form histogram with appropriate weight, that is, given a codebook V, which consists of a set of visual words $t = \{t_1, t_2, \cdots, t_k, \cdots, t_K\}$, we use a K-dimensional vector $W = \{w_1, w_2, \cdots, w_k, \cdots, w_K\}$, with each component w_k representing the weight of a visual word k in an image such that

$$w_k = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{1}{2^{i-1}} sim(j,k) \tag{1}$$

where M_i represents the number of feature point whose ith nearest neighbor is visual k. $sim(j,k)$ represents the similarity between feature point j and visual word k. Notice that in Eq. (1) the contribution of a feature point is dependent on its similarity to visual word k weighted by $\frac{1}{2^{i-1}}$, representing the word is its ith nearest neighbor. In our experiments, we empirically find $N = 4$ is a reasonable setting. As a result, for each layer L_r of image I, one histogram is generated according to the codebook V, that is, a set of histograms $H = \{H_1, H_2, \cdots, H_r, \cdots, H_R\}$, are produced corresponding to multiple layers L.

Generally speaking, the layer which contains abundant of visual points will do more contribution to classification. So, according to the resolution of the layer, we concentrate each component of H linearly to generate the multi-scale histogram H_{scale} with a weight $\frac{1}{2^{r-1}}$, as Eq. (2) shown. The multi-scale histogram contains more visual details exacted from diverse resolutions of image and is more descriptive than the classic histogram exacted from the original resolution of image.

$$H_{scale} = \left[H_1, \frac{1}{2} H_2, \cdots, \left(\frac{1}{2}\right)^{r-1} H_r, \cdots, \left(\frac{1}{2}\right)^{R-1} H_R \right] \tag{2}$$

2.3 Multi-Level Part

In general, images from different categories usually contain diverse numbers of instances of objects, which means different numbers of topics are needed to represent scenes in finer and coarser granularity. So, in multi-level part, different numbers of topics are selected to construct the pLSA model, and then we concentrate the representations linearly to model scenes. Let us denote two topic representations with different numbers of topics by T_a and T_b. The new representation associated with these two topic representations T_{a+b} is obtained by concatenating the topic representation associated with T_a and T_b, as shown in Eq. (3).

$$T_{a+b}(I) = [T_a(I), T_b(I)] \tag{3}$$

Moreover, the number of levels is not limited to two, being extensible to as many as may be useful for a given task. In such a way, we can take into account variant visual granularity to form the final multi-scale multi-level representation and generate a significant classification accuracy.

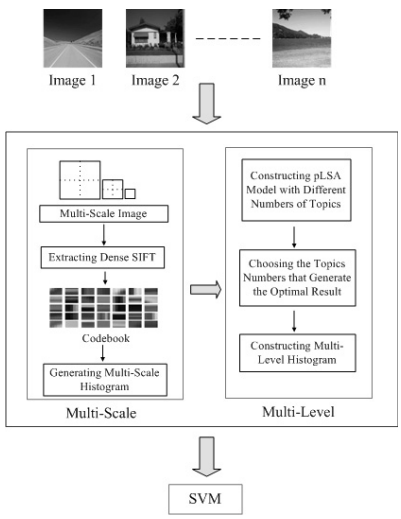

Fig. 1 The framework of msml-pLSA model.

3. Experiments and Discussion

This section reports the experimental setup and the comparing results. The performance of our proposed msml-pLSA method is evaluated on Lazebnik 15 dataset [8], which has been widely used in the previous work. It contains 4485 images from 15 categories: 360 coast, 328 forest, 274 mountain, 410 open country, 260 highway, 308 inside city, 356 tall buildings, 292 streets, 216 bedroom, 210 kitchen, 289 living room, 215 office, 241 suburb, 311 industrial and 315 store. In order to remove the effect of color information of the images, here the gray version of the images is used for our experiments. To the best of our knowledge, this dataset is the current published largest data set for scene categorization.

In our experiments, each scene category is divided randomly into two separate sets of images: 100 images for training and the rest images for testing. Experiments are run with Matlab 7.9 by using computer with Xeon 3.0 GHz processor 16 G RAM. In each image, the dense SIFT feature [3] is computed over a regular grid with the size of 16×16 pixels.

As the size of codebook has influence on the generalization ability and the discriminative power of our method, we first discuss how to decide the size of the codebook to suit the given dataset. Moreover, since the classification accuracy is simultaneously influenced by the number of topics and the size of codebook in the pLSA model, so, experiments are implemented using BoV model, where the classification performance is only affected by the size of codebook, to choose an appropriate size of codebook. The performance variations with different sizes of codebook in BoV model are shown in Fig. 2. We can see that the codebook with the size of 1100 can obtain the highest accuracy and is adopted in our method.

The main contribution of this paper is to investigate how to represent scene image in variant granularity. Our proposed msml-pLSA model consists of two parts: 1) in multi-scale part, image is firstly decomposed into four layers, and for each layer, we construct a classic histogram based on the codebook with the size of 1100. Then we construct the multi-scale histogram by concentrating all the classic histograms with different weight value as Eq. (2) shown; 2) in multi-level part, we investigate the performance variations with different numbers of topics in the pLSA model. In this paper, the pLSA model is random initialization in the EM algorithm. The experimental result is shown in Fig. 3 and topic representations with the number of 40 and 50 generate a better classification accuracy in the pLSA model, so these two topic representations are concentrated linearly to generate the final histogram, and svm classifier with linear kernel is trained to implement the classification task. The classification accuracy can reach 76.26%, and the classification result is shown in Fig. 4.

Additionally, extended experiments are implemented to compare our proposed method with several previous representative scene classification methods, including BoV model, the "gist" feature based method [9], the pLSA model [1], and the Spatial Pyramid Model (SPM) [8]. The implementations of these methods are depicted as follows:

GIST: the "gist" feature is implemented based on the code provided by Oliva and Torralba [10]. Four scale levels ($1:256, 1:128, 1:64, 1:32$) and four orientations (0, 45, 90, 135) are used for the "gist" feature.

pLSA model: as to the setting in paper [1], the size of codebook is set to 1500, the number of topics is set to 25.

SPM: each image is respectively segmented to 1×1, 2×2, 3×3 patches, and histograms based on different segmentation are concatenated to form a high dimension vector, the same setting can be seen in paper [8].

The comparing results are shown in Table 1, and we can get the conclusion that our proposed method respectively outperforms the "gist" feature based method, the BoV model, the pLSA model by 8.418%, 3.76%, 2.16% with SVM classifier and by 13.45%, 4.25%, 2.2% with KNN classifier. Although the SPM method can generate the highest accuracy among the compared methods on the Lazebnik

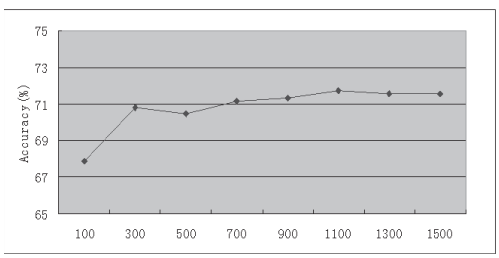

Fig. 2 Performance variation with different size of codebook in BoV model.

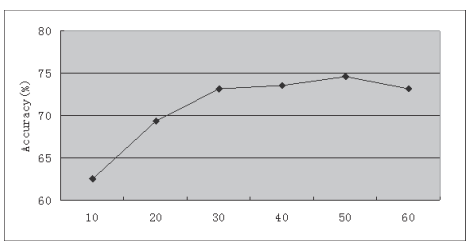

Fig. 3 Performance variation with different number of topics in pLSA model.

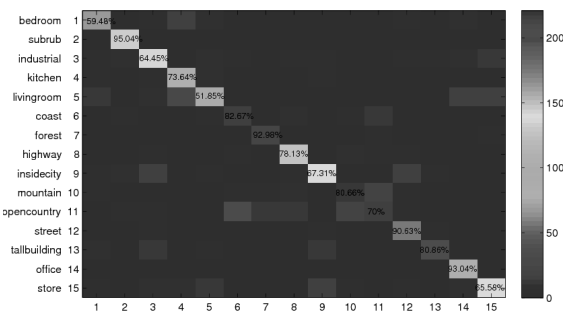

Fig. 4 Classification result with msml-pLSA model.

Table 1 Performance comparison between different methods with two classifiers: SVM with linear kernel and KNN, K is set to 10.

	msml-pLSA	Gist	BoV	pLSA	SPM
SVM	76.26%	67.85%	72.5%	74.1%	83.3%
KNN	74.8%	61.35%	70.55%	72.6%	75.85%

dataset, it has significantly limitations on the image dataset. For instance, if an image dataset contains lots of images of the same category that have the similar content but are only rotated or partial occlusions, the SPM method will generate a lower classification accuracy, since the absolute spatial information will provide the wrong classification information to classify these images to different categories. Additionally, as the SPM method bases on image segmentation, large size of codebook or excessive segmentation may lead to "curse of dimensionality". For example, if the size of the codebook is set to 300 and the image is respectively segmented to 1×1, 2×2, 3×3 patches, we will process a feature vector with 4200-dimension. Generally, spatial information can provide useful cue to scene classification, but how to use spatial information appropriately is still a tough problem. So, the comprehensive study of the use of spatial information is the key point in the future work.

4. Conclusion

In this paper, we propose a novel and practical framework for scene category, where multi-scale multi-level pLSA model is constructed to represent scene in variant visual granularity. Our proposed method consists of two parts: multi-scale part, where the image is decomposed into multiple layers and variant visual details are extracted from the different layers to construct the multi-scale histogram, and multi-level part, where representations corresponding to diverse numbers of topics are linearly concentrated to form the multi-level histogram. The msml-pLSA model can create a more complete representation of the scene due to the inclusion of fine and coarse visual detail information in a joint approach. A comparative study of the proposed method with four state-of-the-art scene classification algorithms shows the superiority of the proposed method.

Acknowledgement

This work is supported by National Nature Science Foundation of China 60803072, and 90820013.

References

[1] A. Bosch, A. Zisserman, and X. Munoz, "Scene classification using a hybrid generative/discriminative approach," IEEE Trans. Pattern Anal. Mach. Intell., vol.30, no.4, pp.712–727, 2008.
[2] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," Mach. Learn., vol.41, pp.177–196, 2001.
[3] D.G. Lowe, "Distinctive image features from scale invariant keypoints," Int. J. Comput. Vis., vol.60, no.2, pp.91–110, 2004.
[4] A. Vailaya, A. Jain, and H. Zhang, "On image classification: City vs landscapes," Pattern Recognit., vol.31, no.12, pp.1921–1935, 1998.
[5] J.S. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," Proc. International Conference on Computer Vision, vol.2, pp.1470–1477, 2003.
[6] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," IEEE Trans. Pattern Anal. Mach. Intell., vol.27, no.10, pp.1615–1630, 2005
[7] Y.G. Jiang, C.W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," Proc. 6th ACM International Conference on Image and Video Retrieval, 2007.
[8] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.2169–2178, 2006.
[9] C. Siagian and L. Itti, "Gist: A mobile robotics application of context-based vision in out-door environment," 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol.3, pp.1063–1069, 2005.
[10] http://people.csail.mit.edu/torralba/code/spatialenvelope/