Image categorization by a classifier based on probabilistic topic model

Takuma Yamaguchi, and Minoru Maruyama
Department of Information Engineering
Shinshu University, Nagano, 380–8553, Japan
s07t213@shinshu-u.ac.jp, maruyama@cs.shinshu-u.ac.jp

Abstract

With rapid increase of number of accessible images and videos, ability to recognize visual information is getting more and more important for content-based information retrieval. Recently, probabilistic topic models, which were originally developed for text analysis, have been used for image categorization successfully. Usually, "topics" which represent contents of an image is detected based on the underlying probabilistic model, then image categorization is carried out using topic distribution as the input feature. Typical method is to use k-nearest neighbor classifier based on L2-distance after topic discovery. In the method, topic distribution is just treated as a feature point. In this paper, we propose a categorization method based on more natural use of the topic distribution, which is derived by using pLSA model. Categorization is carried out by estimating conditional probability $p(category|data)$. We present two types of image categorization tasks, scene classification and document image segmentation, and show the proposed method performs very well. In addition, we also examine the performance of the proposed method under the situation where only the limited number of labeled examples are available. We show our method can perform quite well even in the circumstances.

1. Introduction

With increase of digital images and videos we can access, ability to recognize visual information is getting more and more important for content-based information retrieval. Although, we, humans, can easily analyze and classify images based on their contents, it is still very hard for machine to do such tasks, and much research has been done [3, 7].

Recently, probabilistic topic models, which were originally developed for text analysis, have been used for image analysis successfully [2, 5, 6, 13]. Among the probabilistic topic models, both LDA (Latent Dirichlet Allocation) [1] and pLSA (probabilistic latent semantic analysis) [9] are often used for image analysis. Usually, "topics" which represent contents of an image is detected based on the underlying probabilistic model, then image categorization is carried out using topic distribution as the input feature. Bosch *et al.* [2] have proposed pLSA-based image classification method, which is combination of pLSA and k-nearest neighbor (kNN) classification. They have compared several methods including LDA-based method [5], and showed that the performance of the pLSA-based method is superior to the others'.

In this paper, we propose an image categorization method based on the probabilistic topic model. Like [2], we also use a method based on pLSA model. In our method, categorization is carried out by evaluating $p(category|data)$, which is approximated by using topic distribution. Unlike the kNN-based method, which relies on Euclid distance in the topic space, our method is based on more natural (straightforward) use of topic distribution. We present two types of image categorization tasks, scene classification and document image segmentation, and show the proposed method can outperform pLSA+kNN method. In this paper we also examine the performance of the proposed method under the situation where only the small number of labeled training samples are available. We show our method can perform quite well even in the circumstances.

2. PLSA model for image analysis

2.1. Probabilistic latent semantic analysis

Probabilistic Latent Semantic Analysis (pLSA) is a generative statistical model for text analysis [9]. The model is used to discover topics in a document with the bag-of-words document representation, where spatial relationships between words are ignored.

Let D be a collection of N documents $D = \{d_1, \ldots, d_N\}$. Each document d is a set of words. A word w is an element of the vocabulary $w \in W = \{w_1, \ldots, w_V\}$. Additionally, there is a hidden (latent) topic variable $z \in Z = \{z_1, \ldots, z_T\}$ associated with each occurrence of a word w in a document d. The pLSA model is parameterized by $p(w|z)$ and $p(z|d)$. The document is generated as follows:

1. A document d is selected with probability $p(d)$.

2. For each word in the document, a topic z is selected with $p(z|d)$.
3. A word w is generated with probability $p(w|z)$.

It is assumed that the distribution of words given a latent topic z, $p(w|z)$ is conditionally independent of the document. Marginalizing over topics z, following joint probability is obtained.

$$p(w,d) = p(d) \sum_{z \in Z} p(w|z)p(z|d), \tag{1}$$

2.2. Model learning

The model parameters $p(w|z)$ and $p(z|d)$ are estimated by maximizing the data log-likelihood using an Expectation Maximization (EM) algorithm [4]. The log-likelihood is given by

$$\mathcal{L} = \sum_{d \in D} \sum_{w \in W} n(w,d) \log p(w,d) \tag{2}$$

where $n(w,d)$ is the number of occurrences of a word w in document d. The EM algorithm for estimating parameters of pLSA is as follows :

E-Step:

$$p(z|w,d) = \frac{p(w|z)p(z|d)}{\sum_{z \in Z} p(w|z)p(z|d)} \tag{3}$$

M-Step:

$$p(w|z) = \frac{\sum_{d \in D} n(w,d)p(z|w,d)}{\sum_{w \in W} \sum_{d \in D} n(w,d)p(z|w,d)} \tag{4}$$

$$p(z|d) = \frac{\sum_{w \in W} n(w,d)p(z|w,d)}{\sum_{z \in Z} \sum_{w \in W} n(w,d)p(z|w,d)} \tag{5}$$

With the training procedure described above, parameters $p(w|z)$ and $p(z|d)$ are estimated. When a novel document d_{new} is given, assuming $p(w|z)$s are unchanged, the remaining unknown set of parameters $p(z|d_{new})$ are obtained by the following "folding-in" method :

E-Step:

$$p(z|w,d_{new}) = \frac{p(w|z)p(z|d_{new})}{\sum_{z \in Z} p(w|z)p(z|d_{new})} \tag{6}$$

M-Step:

$$p(z|d_{new}) = \frac{\sum_{w \in W} n(w,d_{new})p(z|w,d_{new})}{\sum_{z \in Z} \sum_{w \in W} n(w,d_{new})p(z|w,d_{new})} \tag{7}$$

where $p(w|z)$ is kept fixed.

2.3. Image representation

To apply the pLSA model to images, visual words should be detected based on the image feature extraction. The image representation, which consists of a set of visual words, is derived through extracting feature points in an image, and then describing the appearance around the feature points.

In this research, Harris-affine interest point detector [12] and SIFT (Scale Invariant Feature Transform) descriptor [11] are used for feature extraction. The Harris-affine detector relies on the combination of corner points detected thorough Harris corner detection [8], multi-scale analysis through Gaussian scale-space and affine normalization using an iterative affine shape adaptation algorithm [10]. The SIFT descriptor is derived from windowed histograms of gradient magnitudes at varying locations and orientations, normalized to correct for contrast and saturation effects. This approach provides some invariance to lighting and poses changes. We use 128 dimensional SIFT descriptor[1].

To define pLSA model on images, visual analogues of a word is needed. The visual vocabulary is obtained by vector quantization of image features. We use k-means algorithm for vector quantization. Each cluster is treated as a visual word.

3. Image categorization via pLSA model

3.1. Image categorization by estimating conditional probability

Applying the probabilistic topic model, we can extract topic distribution from "documents". Typical method for categorization is to use k-nearest neighbor (kNN) classifier. When a novel "document" is given, kNN selects k nearest neighbors of the document based on the Euclidean distance. In this method, topic distribution is treated as just a T-dimensional feature vector. In this paper, we propose another categorization method based on more natural use of the topic distribution. Our goal is to categorize a given (novel) document d_{new}. For that purpose, what we need is the conditional probability $p(category|d_{new})$. In our method, we approximate the conditional probability. Let c be an element of a set of possible categories $\{1, 2, \cdots, C\}$. $p(c|d_{new})$ is given as :

$$p(c|d_{new}) = \sum_{z \in Z} p(z|d_{new})p(c|z) \tag{8}$$

where we assume $p(c,z|d) = p(c|z)$ holds for any document d. For novel document d_{new}, $p(z|d_{new})$ is obtained by fold-in procedure described in 2.2. $p(c|z)$ is estimated from labeled examples as follows :

$$p(c|z) \propto p(z|c)p(c) \tag{9}$$

[1] In the experiments of this paper, we used the software which can be obtained from Visual Geometry Group of Oxford University.

We approximate $p(z|c), p(c)$ as follows:

$$p(c) \approx \frac{N_c}{N},\ p(z|c) \approx \frac{1}{N_c} \sum_{\{i|category(d_i)=c\}} p(z|d_i) \tag{10}$$

where N_c is the number of documents of category c, and N is the total number of examples.

3.2. Experimental results

To examine the effectiveness of the proposed method, two kinds of image categorization tasks are carried out; one is scene image classification using Caltech101 [6] data set (Caltech image) and the other is the content-based document image segmentation (Document). In our experiments, the numbers of topics (T) for Caltech image and Document are 10 and 20, respectively.

Experiment 1 (Caltech images): Experiments of scene image classification were carried out using Caltech 101 data set. The categories we used are airplanes, faces, leopards and motorbikes. We treat each image as a “document”. Our goal with this dataset is classifying “document” to its correct category. The number of training and testing “documents” are 200 and 400, respectively. In the experiment, we examine the following methods: [pLSA+CP] the proposed method, [pLSA+kNN] kNN-based classification after topic discovery based on pLSA, [pLSA+SVM] SVM classifier is trained after topic discovery, [kNN] kNN classifier is used for vocabulary histogram, [SVM] SVM is used for vocabulary histogram. The vocabulary histogram is normalized so that the total number of words is 1. The results are shown in Fig.1-(a). As the figure shows, the proposed method outperforms the others.

Experiment 2 (Document image): Another kind of categorization task is examined. A document can be divided into “parts”, which have same kind of contents. In the experiment, we collected scanned images of mathematical formulas, printed Japanese, printed English and hand written texts from scientific papers. In addition these categories, Caltech101 data set is used as a picture category. After image feature extraction by SIFT detector and descriptor, grouping of the feature points is carried out based on their spatial proximity. In this experiment, each group, which is represented as a set of visual words, is treated as a “document”. For grouping of feature points k-means clustering method is used. Throughout our experiments with this dataset, the “document” categories we consider are $\{$printed Japanese, printed English, math formula, handwritten Japanese, pictures $\}$. The number of training and testing “documents” are 500 and 2,500, respectively. The results are shown in Fig.1-(b). Several classification methods performed equally well for this segmentation task. As the figure shows, the proposed method is one of the winners. In Fig.1-(e)–(g) show segmentation results by pLSA+CP when the document line is treated as a “document”.

4. Image categorization by learning from small number of labeled examples

Usually, one of the most difficult tasks for developing accurate image classifier using machine learning techniques is to gather sufficient amount of labeled examples. Practically, classification method that can be learnt from small number of labeled examples is desirable. While collecting many labeled examples is very hard, it is relatively easy to collect many unlabeled examples. As for the pLSA-based classification methods, as described in 2.2, topic distribution for each document is given by (unsupervised) learning from unlabeled examples. Even if the number of the labeled examples is limited, discovering topic distribution could be made possible without problems. In this section, we examine the performance of the proposed method under the limited number of labeled examples. In the experiments, scene image classification (Caltech image) and the document image segmentation (Document image) were carried out as before. The following two learning procedures from limited number of labeled examples were examined:

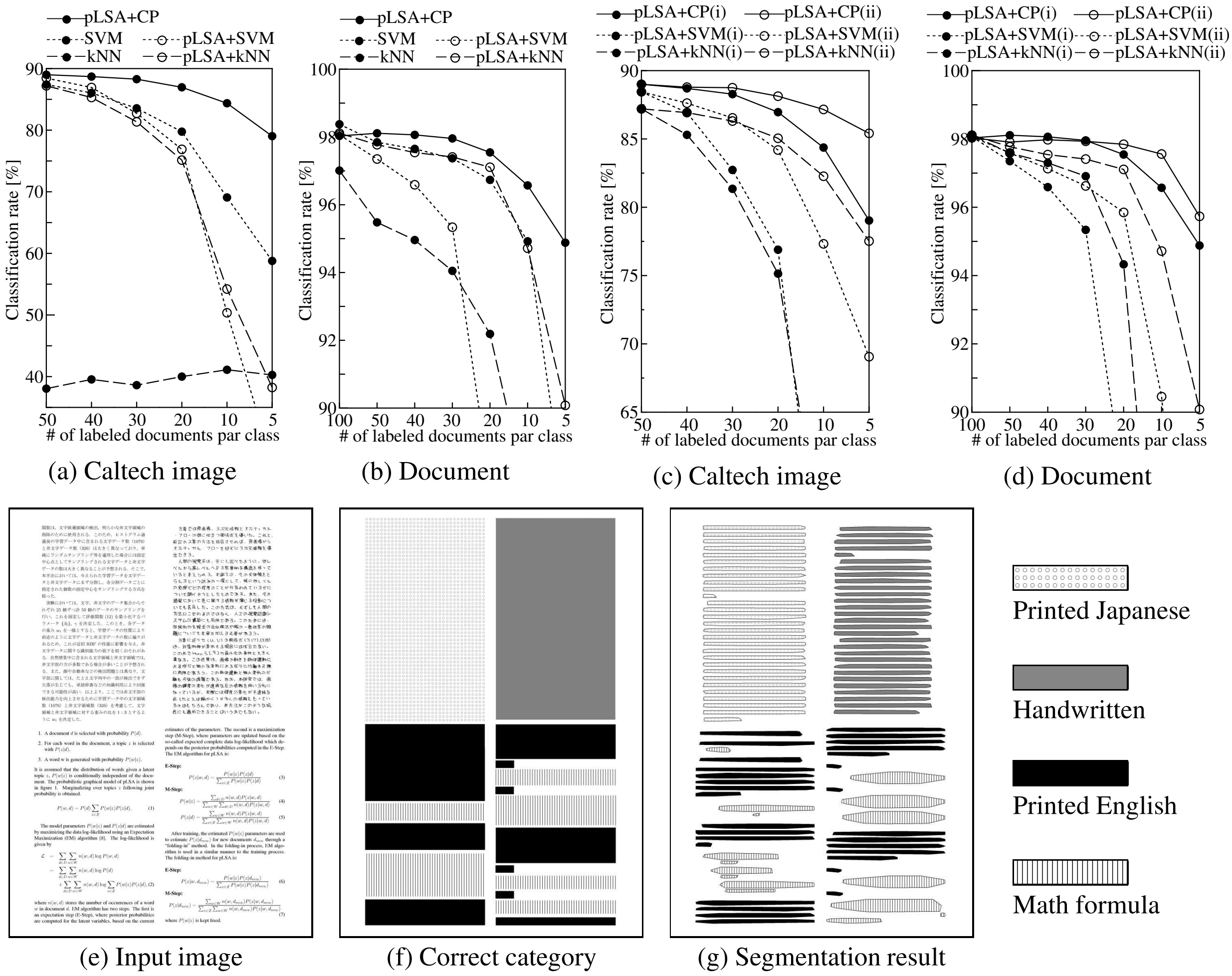
(i) From the whole data, small number of examples are sampled and labeled. In the experiments, for each category, same number of documents are sampled. Both topic discovery and categorization are tried using the labeled samples.

(ii) Unlike (i), topic distribution is calculated using both labeled and unlabeled examples. Classifiers are obtained based on the labeled examples.

We have investigated the performance of the several categorization methods, which are obtained by the above two types of learning procedures. We have tried pLSA+CP, pLSA+kNN and pLSA+SVM. In Fig.1-(c)–(d), classification results are shown under varying number of labeled examples provided. As the figure shows, the proposed method (pLSA+CP) performs well enough even if very limited number of labeled examples are available.

5. Conclusions

In this paper an image categorization method based on topic discovery is proposed. Our method is based on more natural use of the topic distribution, which is derived by using pLSA model. Categorization is carried out by estimating conditional probability $p(category|data)$. Through two types of image categorization experiments (scene classification and document image segmentation), we show the proposed method performs very well. In addition, we also examine the performance of the proposed method under the situation where only the limited number of labeled examples



(a) Caltech image (b) Document (c) Caltech image (d) Document

(e) Input image (f) Correct category (g) Segmentation result

Figure 1. Classification results.

are available. We show our method can perform quite well even in the circumstances.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[2] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via plsa. *Proc of ECCV*, 2006.

[3] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Region-based image querying. *Proc. of International Workshop on Content-Based Access of Image and Video Libraries*, 1997.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society B*, 39(1):1–38, 1977.

[5] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. *Proc. of CVPR*, pages 524–531, 2005.

[6] L. Fei-Fei, R.Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. *Proc. of CVPRW*, 12, 2004.

[7] V. Gudivada and V. Raghavan. Content-based image retrieval-systems. *IEEE Computer*, 28(9):18–22, 1995.

[8] C. Harris and M. Stephens. A combined corner and edge detector. *In Alvey Vision Conference*, pages 147–151, 1988.

[9] T. Hofmann. Probabilistic latent semantic analysis. *Proc. of UAI*, pages 289–296, 1999.

[10] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. *Image and Vision Computing*, 15(6):415–434, 1997.

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, 2004.

[12] K. Mikolajczykm and C. Schmid. Scale & affine invariant interest point detector. *International Journal on Computer Vision (IJCV)*, 60(1):63–86, 2004.

[13] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering object categories in image collections. *MIT Computer Science and Artificial Intelligence Laboratory Technical Report*, 2005.