doi:10.3772/j.issn.1006-6748.2017.04.004

Semi-supervised learning based probabilistic latent semantic analysis for automatic image annotation[1]

Tian Dongping (田东平)[2] * **
(* Institute of Computer Software, Baoji University of Arts and Sciences, Baoji 721007, P. R. China)
(** Institute of Computational Information Science, Baoji University of Arts and Sciences, Baoji 721007, P. R. China)

Abstract

In recent years, multimedia annotation problem has been attracting significant research attention in multimedia and computer vision areas, especially for automatic image annotation, whose purpose is to provide an efficient and effective searching environment for users to query their images more easily. In this paper, a semi-supervised learning based probabilistic latent semantic analysis (PLSA) model for automatic image annotation is presenred. Since it's often hard to obtain or create labeled images in large quantities while unlabeled ones are easier to collect, a transductive support vector machine (TSVM) is exploited to enhance the quality of the training image data. Then, different image features with different magnitudes will result in different performance for automatic image annotation. To this end, a Gaussian normalization method is utilized to normalize different features extracted from effective image regions segmented by the normalized cuts algorithm so as to reserve the intrinsic content of images as complete as possible. Finally, a PLSA model with asymmetric modalities is constructed based on the expectation maximization(EM) algorithm to predict a candidate set of annotations with confidence scores. Extensive experiments on the general-purpose Corel5k dataset demonstrate that the proposed model can significantly improve performance of traditional PLSA for the task of automatic image annotation.

Key words: automatic image annotation, semi-supervised learning, probabilistic latent semantic analysis (PLSA), transductive support vector machine (TSVM), image segmentation, image retrieval

0 Introduction

Probabilistic models with hidden topic variables, originally developed for statistical text modeling of large document collections such as latent semantic analysis (LSA), probabilistic latent semantic analysis (PLSA)[1], latent Dirichlet allocation (LDA)[2] and correlated topic model[3], have recently become an active topic of research in computer vision and pattern recognition. Probabilistic topic models originate from modeling large databases of text documents. When applied to images instead of documents, each topic can be thought of as a certain object type contained in an image. The topic distribution then refers to the degree to which a certain object/scene type is contained in the image. In the ideal case, this gives rise to a low dimensional description of the coarse image content and thus enables retrieval in the very large databases. Another advantage of such models is that topics are learned automatically without requiring any labeled training data. However, the performance of these models usually hinges on an inappropriate assumption[1-3], i. e., all the topics are independent of each other, which will inevitably undermine the performance of multi media processing such as object recognition, image annotation, scene classification and automatic segmentation, etc. Besides, the main drawback of these approaches is that they do not allow exploiting the huge amount of un-annotated data and consequently are affected by the small sample size problem. Except for the merits and demerits of the probabilistic topic models mentioned above, it should also be noted that most of the existing work associated with PLSA have focused on the aspects of its

[1] Supported by the National Program on Key Basic Research Project (No. 2013CB329502), the National Natural Science Foundation of China (No. 61202212), the Special Research Project of the Educational Department of Shaanxi Province of China (No. 15JK1038) and the Key Research Project of Baoji University of Arts and Sciences (No. ZK16047).

[2] To whom correspondence should be addressed. E-mail: tdp211@163.com, tiandp@ics.ict.ac.cn

Received on Sep. 30, 2016

improvement and application, whereas almost no consideration for the construction of its training set and ways of feature normalization, especially in the case of different image features with different magnitudes. Based on this recognition, a semi-supervised learning based probabilistic latent semantic analysis (abbreviated as SSPLSA) is presented for the task of automatic image annotation. First of all, TSVM, as one of the most often used semi-supervised learning methods, is exploited to boost the quality of the training image data with the help of unlabelled data in the presence of the small sample size problem. Then Gaussian normalization method (GNM) is applied to normalize different image features with different magnitudes so as to reserve intrinsic content of the images as complete as possible. Finally, a PLSA model with asymmetric modalities is constructed based on EM algorithm to predict a candidate set of annotations with confidence scores. A series of experiments on the standard Corel5k show the effectiveness and efficiency of SSPLSA.

The rest of this paper is organized as follows. Section 1 summarizes the related work, especially PLSA and several semi-supervised learning methods applied in the field of automatic image annotation. Section 2 elaborates the proposed SSPLSA model from four aspects of PLSA, transductive support vector machine, Gaussian normalization method and the implementation of the SSPLSA, respectively. Section 3 reports experimental results based on the standard Corel5k image dataset. Concluding remarks and future work are discussed in Section 4.

1 Related work

Automatic image annotation (AIA) is a promising methodology for image retrieval. However, it is still in its infancy and is not sophisticated enough to extract perfect semantic concepts according to image low-level features, often producing noisy key words irrelevant to image semantics, which significantly hinders the task of image retrieval. As one of the representative probabilistic topic models, PLSA has been extensively applied in a variety of different image annotation and retrieval tasks. Monay et al. presented a series of PLSA models for AIA[4,5], among which PLSA-MIXED[4] learned a standard PLSA on a concatenated representation of the textual and visual features while PLSA-WORDS or PLSA-FEATURES[5] allowed modeling of an image as a mixture of latent aspects that was defined either by its text captions or its visual features for which the conditional distributions over aspects were estimated from one of the two modalities only. Peng, et al.[6] employed PLSA model to discover the latent topics existing in the audio-clips and carried out the concept classification by a SVM further. In order to extract effective features to reflect the intrinsic content of images, Zhang, et al.[7] proposed a multi-feature PLSA to tackle the problem by combining low-level visual features for image region annotation in which it handled data from two different visual feature domains. In recent work of Ref.[8], a supervised PLSA model was constructed to improve image segmentation by using the classification results. Besides, the standard PLSA was extended to higher order for image indexing by treating images, visual features and tags as three observable variables of an aspect model[9]. In more recent work[10], Tian, et al. integrated PLSA with multiple Markov random fields (MRF) for AIA. Particularly, MRF was used to fuse the contextual information of images. Alternatively, as for the PLSA model itself, it could be improved from four aspects of its initialization, visual words, hidden layers and integration with other models. As the representative work, Wang, et al.[11] proposed a method to build an effective visual vocabulary by using hierarchical Gaussian mixture model instead of the traditional clustering methods to improve its visual words. Lu, et al.[12] exploited rival penalized competitive learning method to initialize the model so as to enhance the performance of PLSA. In addition, Lienhart, et al.[13] extended the standard single-layer PLSA to multiple multimodal layers that consisted of two leaf-PLSA and a single top-level PLSA node merging the two leaf-PLSA. Meanwhile, a correlated PLSA model[14] was constructed by introducing a correlation layer between images and latent topics to incorporate the image correlations.

Semi-supervised learning (SSL), which aims at learning from labeled and unlabeled data, typically a small amount of labeled data with a large amount of unlabeled data, has aroused considerable interest in the data mining and machine learning fields, since it is usually hard to collect enough labeled data in practical applications. SSL falls between unsupervised learning (without any labeled training data) and supervised learning (with completely labeled training data), whose aim is to achieve good classification performance with the help of unlabelled data in the presence of the small sample size problem. As the representative work of semi-supervised learning for AIA, Li, et al.[15] formulated automatic image annotation as a joint classification task based on 2D conditional random fields, the SSL technique was utilized to exploit the unlabeled data for improving its joint classification performance. In Ref.[16], a semi-supervised ensemble of classifiers, viz. weighted semi-supervised adaboost (WSA), was constructed for AIA. Note that WSA is able to incorporate unlabeled instances that are annotated based on the classifier from the previous stage, and then used to train the next classifier. Zhu, et al.[17] proposed to annotate images based on a progressive model to obtain the candidate annotations of unlabeled images. In addition, Yuan, et al.[18] exploited semi-supervised cross-domain learning with group sparsity to boost the performance of automatic image annotation, etc.

2 Proposed SSPLSA model

Fig. 1 illustrates the framework of the proposed SSPLSA model, which mainly includes two stages, viz. feature extraction and PLSA modeling. Details of SSPLSA will be elaborated in the following subsections.

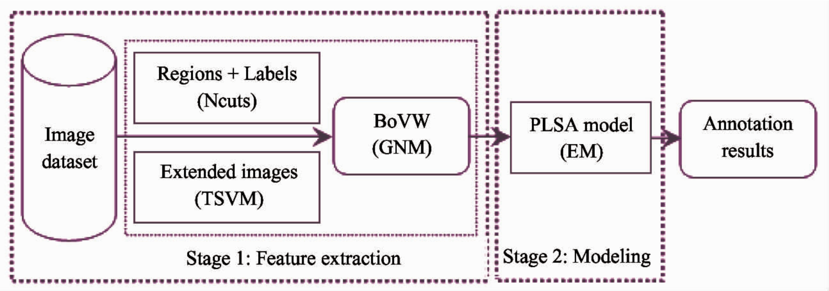


Fig. 1 Framework of the SSPLSA model

2.1 PLSA model

PLSA[1] is a statistical latent aspect model which introduces a hidden variable (latent aspect) z_k in the generative process of each element w_j in document d_i. Given the unobservable variable z_k, each occurrence w_j is independent of the document it belongs to, which corresponds to the following joint probability:

$$P(d_i,\ w_j) = P(d_i)\sum_{k=1}^{K} P(w_j \mid z_k)P(z_k \mid d_i) \tag{1}$$

The model parameters of PLSA are the two conditional distributions: $P(w_j | z_k)$ and $P(z_k | d_i)$. $P(w_j | z_k)$ characterizes each aspect and remains valid for documents out of the training set while $P(z_k | d_i)$ is only relative to the document-specific and cannot carry any prior information to an unseen document. The EM algorithm is usually utilized to estimate the parameters through maximizing the log-likelihood of the observed data.

$$L = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,\ w_j)\log P(d_i, w_j) \tag{2}$$

where $n(d_i,\ w_j)$ denotes the number of times, term w_j occurring in document d_i. The steps of the EM algorithm can be succinctly described as follows.

E-step. The conditional distribution $P(z_k | d_i,\ w_j)$ is computed from the previous estimate of the parameters:

$$P(z_k \mid d_i,\ w_j) = \frac{P(z_k \mid d_i)P(w_j \mid z_k)}{\sum_{l=1}^{K} P(z_l \mid d_i)P(w_j \mid z_l)} \tag{3}$$

M-step. The parameters $P(w_j | z_k)$ and $P(z_k | d_i)$ are updated with the new expected values $P(z_k | d_i,\ w_j)$:

$$P(w_j \mid z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)P(z_k \mid d_i, w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i, w_m)P(z_k \mid d_i, w_m)} \tag{4}$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j)P(z_k \mid d_i, w_j)}{\sum_{j=1}^{M} n(d_i, w_j)} \tag{5}$$

If one of the parameters is known, the other one can be inferred by leveraging fold-in method, which updates the unknown parameters with the known ones kept fixed so that it can maximize the likelihood with respect to the previously trained parameters. Given a new image visual features $v(d_{new})$, the conditional probability distribution $P(z_k | d_{new})$ can be inferred with the previously estimated model parameters $P(v | z_k)$, then the posterior probabilities of keywords can be computed by the following formula:

$$P(w \mid d_{new}) = \sum_{k=1}^{K} P(w \mid z_k)P(z_k \mid d_{new}) \tag{6}$$

From Eq. (6), the top n words can be selected as the annotations for the new image.

2.2 TSVM algorithm

Semi-supervised learning (SSL) problem has recently drawn large attention in the field of machine learning and pattern recognition mainly due to its significant importance in practical applications. Concretely, SSL is a family of algorithms that takes the advantage of both labeled and unlabeled data and has been extensively studied for many years. Among them, the transductive support vector machine (TSVM), also called semi-supervised support vector machine located between supervised learning with fully labeled training data and unsupervised learning without any labeled training data, is a promising way to find out the underlying relevant data from the unlabeled ones. TSVM works as follows for mining the relevant image regions: Given a keyword w, several labeled regions are taken as the relevant examples and the initial non-relevant examples are randomly sampled from the remaining regions. A two-class SVM classifier is trained firstly. Then based on the learnt SVM classifier, the most confident relevant regions and the most non-relevant ones are added into the relevant and non-relevant training set respectively. With the expanded training set, SVM classifier will be re-trained until the maximum time of iteration is reached. Finally, an expanded set of labeled regions can be obtained to benefit for modeling the visual feature distribution of the keyword w. Thus in this paper, TSVM is adopted to explore more relevant image regions to boost the performance of the PLSA model, which can be described as Algorithm 1.

Algorithm 1 Pseudocode of TSVM for mining relevant regions

```
Input: $R_L^0$ and $R_U^0$denote the sets of labeled and unlabeled regions for the keyword $w$, $S$ is a SVM classifier,
       $m$, $n$ and $K$ denote control parameters
Process:
1.  for $k = 1$ to $K$ do
2.      Learning a SVM classifier $S$ from $R_L^k$;
3.      Using $S$ to classify regions in $R_U^k$;
4.      Selecting $m$ most confidently predicted regions from $R_U^k$ which are labeled as relevant examples;
5.      Selecting $n$ most confidently predicted regions from $R_U^k$ which are labeled as non-relevant examples;
6.      Adding $m + n$ regions with their corresponding labels into $R_L^k$;
7.      Removing these $m + n$ regions from $R_U^k$;
8.  end for
Output: $R_L^k$ an expanded set of labeled regions
```

2.3 Feature normalization

During the course of image feature extraction, different kinds of features may have different magnitudes. How to appropriately normalize these features plays a crucial role in the subsequent image processing. Based on this recognition, the Gaussian normalization method (GNM) is used for image feature normalization[19]. Let $F_i = (f_{i1}, \cdots, f_{ik}, \cdots, f_{iq})$ be the feature vector representing the i-th image region. The mean μ_k and standard deviation σ_k of the k-th feature dimension can be easily calculated. Subsequently the feature vectors can be normalized to $N(0,1)$ according to:

$$\begin{aligned} F_i &= \left(\frac{f_{i1} - \mu_1}{k\sigma_1}, \cdots, \frac{f_{ik} - \mu_k}{k\sigma_k}, \cdots, \frac{f_{iq} - \mu_q}{k\sigma_q}\right) \\ &= (f'_{i1}, \cdots, f'_{ik}, \cdots, f'_{iq}) \end{aligned} \quad (7)$$

In Eq. (7), assume that each feature is normally distributed and $k = 3$. According to the 3-σ rule, the probability of an entry ' s value being in the range of $[-1, 1]$ is approximately 99%. By defining the following Eq. (8), namely, a simple additional shift embedded can guarantee that 99% of the feature values will be within $[0, 1]$.

$$F_i = \left(\frac{f'_{i1} + 1}{2}, \cdots, \frac{f'_{ik} + 1}{2}, \cdots, \frac{f'_{iq} + 1}{2}\right) \quad (8)$$

where each f'_{i1}, f'_{ik}, f'_{iq} represents a normalized feature vector within $[-1, 1]$.

2.4 Implementation of SSPLSA

Based on the contents aforementioned, the SSPLSA model can be summarized as follows. Note that Algorithm 2 is utilized to estimate the parameters of SSPLSA while Algorithm 3 is applied to calculate the annotations of the test image.

Algorithm 2 Estimation of the SSPLSA parameters $P(w|z)$

```
Input: visual features $v_n$ and textual words $w_m$ of training images
Process:
1.  Randomly initializing the probability tables $P(z_k|d_i)$ and $P(w_j|z_k)$;
```

```
while increase in the likelihood of validation data $\Delta L_s > T_s$ do
  {E step}
  for $k \in 1, \cdots, K$ and all $(d_i, w_j)$ pairs in training documents do
      Computing $P(z_k | d_i, w_j)$ with Eq. (3);
  end for
  {M step}
  for $k \in 1, \cdots, K$ and $j \in 1, \cdots, M$ do
      Computing $P(w_j | z_k)$ with Eq. (4);
  end for
  for $k \in 1, \cdots, K$ and $i \in 1, \cdots, N$ do
      Computing $P(z_k | d_i)$ with Eq. (5);
  end for
  Computing the likelihood of validation data $L_s$ with Eq. (2);
end while
Output: $\theta_k = \{P(w_1 | z_k), P(w_2 | z_k), \cdots, P(w_M | z_k)\}$, $k \in 1, \cdots, K$.
```

Algorithm 3 Annotation estimation for the testing images

```
Input: model parameters $\theta_k$, visual features $f$ of the testing image $d_{new}$
Process:
Randomly initializing the $P(v|z)$ probability table;
while increase in the likelihood of validation data $\Delta L_f > T_f$ do
  {E step}
  for $k \in 1, \cdots, K$ and all $(d_i, v_j)$ pairs in training documents do
      Computing $P(z_k | d_i, v_j)$ with Eq. (3);
  end for
  {Partial M step}
  for $k \in 1, \cdots, K$ and $j \in 1, \cdots, M$ do
      Computing $P(v_j | z_k)$ with Eq. (4);
  end for
  Computing the likelihood of validation data $L_f$ from $P(v|z)$ and $P(z|d)$
  from previous modality with Eq. (2);
end while
Saving $\eta_k = \{P(v_1 | z_k), P(v_2 | z_k), \cdots, P(v_M | z_k)\}$;
Randomly initializing the $P(z|d)$ probability table;
while increase in the likelihood of validation data $\Delta L_s > T_s$ do
  {E step}
  for $k \in 1, \cdots, K$ and all $(d_i, v_j)$ pairs in training documents do
      Computing $P(z_k | d_i, v_j)$ with Eq. (3);
  end for
  {Partial M step}
  for $k \in 1, \cdots, K$ and $i \in 1, \cdots, N$ do
      Computing $P(z_k | d_i)$ with Eq. (5);
  end for
  Computing the likelihood of validation data $L_s$ from $P(z|d)$ and $P(v|z)$
  from previous modality with Eq. (2);
end while
Computing $P(w_j | d_{new})$ with Eq. (6);
Selecting top-$n$ words with highest $P(w_j | d_{new})$ as the annotations.
Output: annotation results of $d_{new}$, $L = \{w_1, \cdots, w_l\}$.
```

3 Experimental results and analysis

3.1 Dataset and evaluation measures

To evaluate the performance of the SSPLSA model, it is tested on the Corel5k dataset[20], which consists of 5000 images from 50 Corel Stock Photo CD's. Each CD contains 100 images with a certain theme, of which 90 are designated to be in the training set and 10 in the test set, resulting in 4500 training images and a balanced 500-image test collection. Note that the dictionary contains 260 words that appear in both the training and testing sets, and the normalized cuts (Ncuts) algorithm[21] is applied to segment images into a number of meaningful regions. For each image, at most the 10 largest regions are selected and 809-dimensional visual features (color, texture, shape and saliency) are extracted for each region, which include 81-dim grid color moment features, 59-dim local binary pattern texture features, 120-dim Gabor wavelets texture features, 37-dim edge orientation histogram features and 512-dim GIST features respectively. All of the extracted features are subsequently normalized by GNM described in subsection 2.3 and further utilized to train PLSA model based on the EM algorithm. It should be noted that the keywords of the top-five ranked regions with the largest area are determined to annotate the test image. In our case, 5% regions in the entire relevant regions are labeled for each keyword and are taken as the initial positive examples for training SVM. Besides, the parameters are set as: $K = 10$, $m = 3$, $n = 5$, especially m indicates 3 relevant regions and n denotes 5 non-relevant regions to be added to the training set for re-training the SVM classifier during each round of iteration. In addition, to make a fair comparison with other AIA methods, the most commonly used metrics precision and recall of every word in the test set are calculated and the mean of these values is used to summarize the model performance.

3.2 Results of automatic image annotation

MATLAB 7.0 is applied to implement the proposed SSPLSA model. Specifically, the experiments are carried out on a 3.30GHz Intel Core i5 CPU personal computer with 4.0G memory running Win 7 Ultimate. To systematically verify the effectiveness of the SSPLSA, thorough experiments are performed on the Corel5k dataset and compared with several previous approaches[5,20,22-24]. Table 1 reports the experimental results based on two sets of words: the subset of 49 best words and the complete set of all 260 words that occur in the training set. From Table 1, it can be clearly observed that our model markedly outperforms all the others, especially the first three approaches. Meanwhile, it is also superior to MBRM, PLSA-WORDS and CRMR by the gains of 16, 20 and 6 words with non-zero recall, 12%, 7% and 1% mean per-word recall as well as 14%, 30% and 1% mean per-word precision on the set of 49 words, 25%, 25% and 9% mean per-word recall in conjunction with 21%, 64% and 5% mean per-word precision on the set of 260 words, respectively. Compared to PLSA-WORDS, the significant performance improvement is largely ascribed to the applications of the TSVM to enhance the quality of the training image data and GNM to normalize different image features with different magnitudes. Furthermore, it is argued that combining these techniques allows them to benefit from each other and yield a great deal of advantages in terms of annotation accuracy and ease-of-use of the model. Note that CRMR listed in Table 1 denotes CRM with rectangular regions as input[24].

Table 1 Performance comparison on Corel5k dataset

Models	TM	CMRM	CRM	MBRM	PLSA-WORDS	CRMR	SSPLSA
#Words with recall > 0	49	66	107	109	105	119	125
Results on 49 best words							
Mean per-word recall	0.34	0.48	0.70	0.68	0.71	0.75	0.76
Mean per-word precision	0.20	0.40	0.59	0.64	0.56	0.72	0.73
Results on all 260 words							
Mean per-word recall	0.04	0.09	0.19	0.20	0.20	0.23	0.25
Mean per-word precision	0.06	0.10	0.16	0.19	0.14	0.22	0.23

Table 2 shows some annotation results (only four cases are listed here due to the limited space) yielded by PLSA-WORDS and SSPLSA respectively. It can be clearly observed that our model can generate more accurate annotation results compared with the original annotations as well as the ones provided in the literature[5]. Note that the enriched and re-ranked annotations compared to those of the ground truth and PLSA-WORDS are underlined and italicized respectively. Taking the first image for example, there exist four tags in the original annotation list. However, after annotation by SSPLSA, its annotation is enriched by the other keyword "leaves", which is very appropriate and reasonable to describe the visual content of the image. Similarly, the enriched labels "farm" and "plants" for the second image, "trees" for the third image and "ice" for the fourth image.

Table 2 Annotation comparison between PLSA-WORDS and SSPLSA

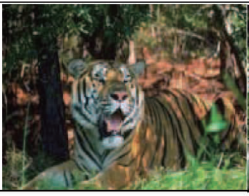
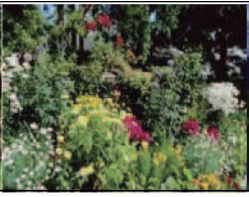
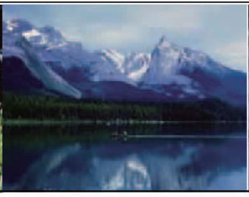
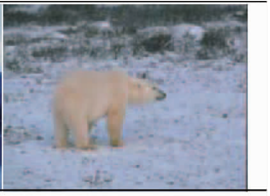
Images				
Ground Truth Annotation	tiger, forest, cat, trees	garden, flowers, landscape, trees	mountain, water, sky, clouds	polar, bear, snow, tundra
PLSA-WORDS Annotation	tiger, trees, leaves, forest, cat	flowers, trees, garden, plants, farm	sky, mountain, water, clouds, trees	snow, bear, polar, tundra, ice
SSPLSA Annotation	tiger, trees, *cat*, *forest*, leaves	flowers, *garden*, *trees*, *farm*, *plants*	*water*, *mountain*, *sky*, clouds, trees	snow, bear, polar, tundra, ice

Fig. 2 illustrates the precision-recall curves of PLSA-WORDS and SSPLSA models based on the Corel5k dataset, with the number of annotations from 2 to 10. It is easy to see that the performance of the proposed model is evidently superior to that of the PLSA-WORDS.

To further illustrate the effect of SSPLSA model for automatic image annotation, Fig. 3 displays the average annotation precision of the selected 10 words "mountain", "snow", "tree", "building", "water", "beach", "bear", "sky", "cat" and "house" based on PLSA-WORDS and SSPLSA models respectively. As shown in Fig. 3, the average precision of the model is consistently higher than that of PLSA-WORDS. In addition, as for the complexity of SSPLSA, assume that there are D training images and each image produces R visual feature vectors, then the complexity of the model is $O(DR)$.

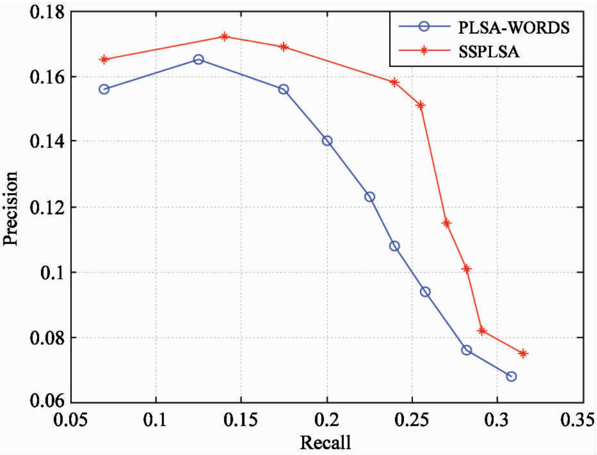


Fig. 2 Precision-recall curves of PLSA-WORDS and SSPLSA

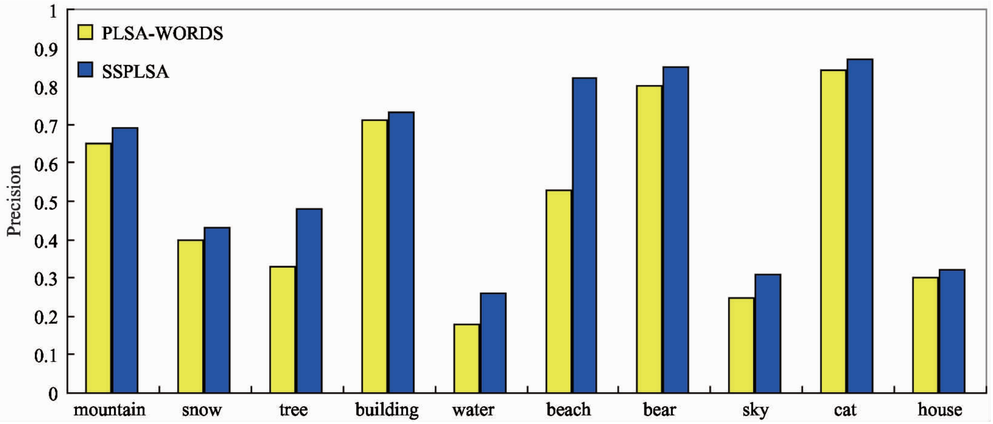


Fig. 3 Average precision based on PLSA-WORDS and SSPLSA

4 Conclusions and future work

In this paper, a semi-supervised learning based PLSA is presented for automatic image annotation.

First, the most widely used TSVM is applied to enhance the quality of the training image data with the help of unlabelled data in the presence of the small sample size problem. Second, GNM is used to normalize different image features with different magnitudes so as to reserve the intrinsic content of the images as complete as possible. Third, a PLSA model with asymmetric modalities is constructed to predict a candidate set of annotations. Extensive experiments on the Corel5k validate that the SSPLSA model outperforms peer methods in the literature in terms of accuracy, efficiency and robustness. In future work, more complicated real-world image datasets will be applied, such as NUS-WIDE and MIRFLICKR, to further evaluate the scalability of SSPLSA model. And no doubt inaccurate image segmentation will make the region based image feature representation imprecise and therefore undermine the performance of PLSA based approaches. So exploring more efficient image segmentation methods is helpful to boost the annotation performance. Furthermore, image segmentation itself is a worthy research direction in the field of computer vision and pattern recognition.

References

[1] Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001, 42(1): 177-196

[2] Blei D, Ng A, Jordan M. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 2003, 3(4): 993-1022

[3] Blei D, Lafferty J. Correlated topic models. *Annals of Applied Statistics*, 2007, l(1): 17-35

[4] Monay F, Gatica-Perez D. On image auto-annotation with latent space models. In: Proceedings of the 11th International Conference on Multimedia, Berkeley, USA, 2003. 275-278

[5] Monay F, Gatica-Perez D. Modeling semantic aspects for cross-media image indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2007, 29(10): 1802-1817

[6] Peng Y, Lu Z, Xiao J. Semantic concept annotation based on audio PLSA model. In: Proceedings of the 17th International Conference on Multimedia, Beijing, China, 2009. 841-844

[7] Zhang R, Guan L, Zhang L, et al. Multi-feature PLSA for combining visual features in image annotation. In: Proceedings of the 19th International Conference on Multimedia, Scottsdale, USA, 2011. 1513-1516

[8] Guo Q, Li N, Yang Y, et al. Integrating image segmentation and annotation using supervised PLSA. In: Proceedings of the 20th International Conference on Image Processing, Melbourne, Australia, 2013. 3800-3804

[9] Nikolopoulos S, Zafeiriou S, Patras I, et al. High order PLSA for indexing tagged images. *Signal Processing*, 2013, 93(8): 2212-2228

[10] Tian D, Zhao X, Shi Z. Fusing PLSA model and Markov random fields for automatic image annotation. *High Technology Letters*, 2014, 20(4): 409-414

[11] Wang Z, Yi H, Wang J, et al. Hierarchical Gaussian mixture model for image annotation via PLSA. In: Proceedings of the 5th International Conference on Image and Graphics, Xi'an, China, 2009. 384-389

[12] Lu Z, Peng Y, Ip H. Image categorization via robust PLSA. *Pattern Recognition Letters*, 2010, 31(1): 36-43

[13] Romberg S, Lienhart R, Horster E. Multimodal image retrieval: fusing modalities with multilayer multimodal PLSA. *International Journal of Multimedia Information Retrieval*, 2012, 1(1): 31-44

[14] Li P, Cheng J, Li Z, et al. Correlated PLSA for image clustering. In: Proceedings of the 17th International Conference on Multimedia Modeling, Taipei, China, 2011. 307-316

[15] Li W, Sun M. Semi-supervised learning for image annotation based on conditional random fields. In: Proceedings of the 5th International Conference on Image and Video Retrieval, Tempe, USA, 2006. 463-472

[16] Marin-Castro H, Sucar E, Morales E. Automatic image annotation using a semi-supervised ensemble of classifiers. In: Proceedings of the 12th Iberoamerican Conference on Progress in Pattern Recognition, Image Analysis and Applications, Valparaiso, Chile, 2007. 487-495

[17] Zhu S, Liu Y. Semi-supervised learning model based efficient image annotation. *IEEE Signal Processing Letters*, 2009, 16(11): 989-992

[18] Yuan Y, Wu F, Shao J, et al. Image annotation by semi-supervised cross-domain learning with group sparsity. *Journal of Visual Communication and Image Representation*, 2013, 24(2): 95-102

[19] Rui Y, Huang T, Ortega M, et al. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transaction on Circuits and Systems for Video Technology*, 1998, 8(5): 644-655

[20] Duygulu P, Barnard K, Freitas N De, et al. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In: Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark, 2002. 97-112

[21] Shi J, Malik J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000, 22(8): 888-905

[22] Jeon L, Lavrenko V, Manmantha R. Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of the 26th International ACM SIGIR Conference on Research and Development in Information Retrieval, Toronto, Canada, 2003. 119-126

[23] Lavrenko V, Manmatha R, Jeon J. A model for learning the semantics of pictures. In: Advances in Neural Information Processing Systems 16, Vancouver, Canada, 2003. 553-560

[24] Feng S, Manmatha R, Lavrenko V. Multiple Bernoulli relevance models for image and video annotation. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, Washington, USA, 2004. 1002-1009

Tian Dongping, born in 1981. He received his M. Sc. and Ph. D. degrees in computer science from Shanghai Normal University and Institute of Computing Technology, Chinese Academy of Sciences in 2007 and 2014, respectively. His research interests include computer vision, machine learning and evolutionary computation.