

Extended Probabilistic Latent Semantic Analysis for Automatic Image Annotation

Dongping Tian

Institute of Computer Software
Baoji University of Arts and Sciences
No.1 Hi-Tech Avenue, Hi-Tech District, Baoji, Shaanxi 721013, P.R. China

Institute of Computational Information Science
Baoji University of Arts and Sciences
No.44 Baoguang Road, Weibin District, Baoji, Shaanxi 721007, P.R. China
tiandp@ics.ict.ac.cn, tdp211@163.com

Received March 2017; revised May 2017

ABSTRACT. *Automatic image annotation is a promising solution to enable the semantic based image retrieval via keywords. 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 keywords irrelevant to image semantics, which may be an obstacle to getting high-quality image retrieval. So in this paper, we propose a novel extended probabilistic latent semantic analysis (PLSA) to improve the performance of automatic image annotation. On one hand, the traditional bag-of-visual-words model is improved by integrating the contextual semantic information among visual words based on the PLSA model. Meanwhile, the approximation strategy of pseudo-likelihood in Markov random field (MRF) is introduced to combine the feature appearance similarity in feature domain and the contextual semantic information in spatial domain. On the other hand, since the traditional expectation-maximization (EM) algorithm used to train the PLSA model is sensitive to its initialization, so a rival penalized competitive learning (RPCL) based method is employed to overcome this deficiency and to provide a good initial estimate of this model. Conducted experiments on the standard Corel5k image dataset demonstrate that the proposed method is significantly more effective than several state-of-the-art approaches regarding their effectiveness and efficiency in the task of automatic image annotation.*

Keywords: Automatic image annotation, PLSA, RPCL, MRF, Image retrieval

1. Introduction. Automatic image annotation (AIA) has been an active research topic in the recent years due to its potentially large impact on both image understanding and web/database image search. AIA aims to provide an efficient and effective searching environment for users to query their images more easily, but current image retrieval systems are still not very accurate when assigning images into a large number of keyword classes. Automatic image annotation usually includes two types of tasks. First, image annotation assigns descriptive metadata (e.g., caption) to a given (entire) image. Second, region annotation annotates each object (e.g., image region) within a given image with appropriate textual tags. From the literature, it can be clearly observed that the current techniques for AIA can be summarized into two categories. The first one poses image annotation as a supervised classification problem, which treats each semantic word or concept as an independent class and assigns each word or concept one classifier. To be more specific, such

kind of approaches predicts the annotations for a new image by computing the similarity at the visual level and propagating the corresponding words subsequently. Representative work includes the asymmetrical support vector machine based multiple-instance learning [1] and supervised formulation for semantic image annotation and retrieval [2], etc. In contrast, the second category treats the words and visual tokens in each image as equivalent features in different modalities. Followed by image annotation is formalized via modeling the joint distribution of visual and textual features on the training data and predicting the missing textual features for a new image. Representative research includes the translation model (TM) [3] which treated AIA as a process of translation from a set of blob tokens to a set of keywords, cross-media relevance model (CMRM) [4] by assuming the blobs and words were mutually independent given a specific image, continuous space relevance model (CRM) [5], multiple Bernoulli relevance model (MBRM) [6] and dual cross-media relevance model [7], etc. Particularly the latent aspect models such as probabilistic latent semantic analysis (PLSA) [8-9], latent Dirichlet allocation (LDA) [10-11] and correlated topic model (CTM) [12], etc. By comparison, the former approach is relatively direct and natural to be understood. However, its performance is limited with the increase of the number of the semantic concepts and explosive multimedia data on the web. On the other hand, the latter often requires large-scale parameters to be estimated and the accuracy is strongly affected by the quantity and quality of the training data available. This paper focuses on PLSA model to implement the task of AIA. More detailed reviews on PLSA will be summarized in the next section.

The rest of this paper is organized as follows. Section 2 summarizes the related work, particularly PLSA model applied in the field of image annotation and retrieval as well as the improvements of PLSA itself. Section 3 elaborates the proposed EPLSA from the aspects of the construction of the proposed bag-of-visual-words model and its parameter estimation. In Section 4, conducted experiments are reported and analyzed based on the Corel5k dataset. Finally, Section 5 presents some concluding remarks and future work.

2. Related Work. Probabilistic topic model (PTM) with hidden topic variables, originally developed for statistical text modeling of large document collections, has recently been an active topic of research for multimedia representation and annotation in both computer vision and pattern recognition. As a representative PTM, probabilistic latent semantic analysis has been widely applied in a variety of different image processing tasks, such as image annotation, image retrieval and image classification, etc. (1) **Image annotation.** PLSA-WORDS [9] allowed modeling of an image as a mixture of latent aspects that was defined by its text captions for which the conditional distributions over aspects were estimated only from the textual modality. In order to extract effective features to reflect the intrinsic content of images as complete as possible, Zhang et al.[13] put forward a multi-feature PLSA (MF-PLSA) to tackle the problem by combining low-level visual features for image region annotation in that it handled data from two different visual feature domains. In recent work of [14], Guo et al. constructed a supervised PLSA (S-PLSA) model to improve image segmentation by using the classification results together with an integrated framework based on PLSA and S-PLSA to accommodate segmentation and annotation procedures, etc. (2) **Image retrieval.** In [15], a multilayer PLSA was developed to eliminate the noisiest words generated by the vocabulary building process. In the meanwhile, a spatial weighting scheme was adopted to reflect the information about the spatial structure of the images. After that the authors built visual phrases from groups of visual words that were involved in strong association rules. In addition, the standard PLSA model was extended to higher order for image indexing by treating images, visual features and tags as three observable variables of an aspect model [16], whose purpose

was to learn a space of latent topics that incorporated the semantics of both visual and tag information. (3) **Image classification.** In the work of [17], Lu et al. put forward a rival penalized competitive learning based method to provide an initial estimate for PLSA model used in image categorization through directly maximizing the likelihood function based on the observed data. Subsequently a histogram was utilized to represent the spatial relationships between objects in [18], and then the PLSA was extended by considering the spatial relationships between topics (SR-PLSA) to model the image as the input for support vector machine to classify the scene. In addition, a co-regularized probabilistic latent semantic analysis (Co-PLSA)[19] was proposed for multi-view clustering, etc.

Alternatively, as far as the PLSA model itself is concerned, it can be improved from the following four aspects. (1) **Initialization.** The performance of PLSA is strongly affected by the initialization of the model. Since the expectation maximization algorithm used to train the PLSA model is sensitive to its initialization. Hence a method for identifying a good initialization or alternatively a good trained model is very important. The early notable research leveraged latent semantic analysis to better initialize the parameters of a corresponding PLSA model [20], and the EM algorithm was then employed to further refine the initial estimate. Thereafter Rodner et al.[21] proposed to use an ensemble of PLSA models that were trained using random fractions of the training data for scene recognition. Besides, Lu et al.[17] exploited rival penalized competitive learning method to initialize the PLSA model, etc. (2) **Visual words.** In [22], PLSA was employed to region-based image classification and two soft vector quantization methods were proposed to tackle the small sample problem in visual vocabulary construction. Afterwards Wang et al.[23] proposed a method to build an effective visual vocabulary by using hierarchical Gaussian mixture model instead of traditional clustering methods, etc. (3) **Adding hidden layers.** In [24], the standard single-layer PLSA model was extended to multiple multimodal layers (MM-PLSA), which consisted of two leaf-PLSA (here from two different data modalities: image tags and visual image features) and a single top-level PLSA node merging the two leaf-PLSA. Besides, a correlated probabilistic latent semantic analysis model [25] was proposed by introducing a correlation layer between images and latent topics to incorporate the image correlations. (4) **Integrating with other models.** In [17], Lu et al. integrated PLSA with ensemble-based SVM for image categorization. The work by Zhuang et al.[26] combined PLSA with visual attention model to create AM-PLSA. However, this kind of algorithm just added a preprocessing to PLSA and did not change the essence of it. In addition, the PLSA was implemented for scene classification under the framework of multi-instance multi-label learning [27]. Followed by Ergul et al.[28] fused spatial pyramid matching (SPM) and probabilistic latent semantic analysis for scene classification. In the literature [29], an image annotation system was developed by integrating the PLSA model and canonical correlation analysis. A recent work by Cheng et al.[30] integrated the unsupervised PLSA model and k -nearest neighbor classifier for automatic landslide detection. In more recent work [9], a refining image annotation method was proposed by combining PLSA and random walk model, etc.

As briefly reviewed above, most of the PLSA related models can achieve encouraging performance and motivate us to explore image annotation with the help of their excellent experiences and knowledge. So in this paper, we present a novel extended PLSA for the task of automatic image annotation (abbreviated as EPLSA). On one side, the traditional bag-of-visual-words model is improved by fusing the contextual semantic information among visual words based on PLSA. At the same time, the approximation strategy of pseudo-likelihood in Markov random field is introduced to combine the feature appearance similarity in feature domain and the contextual information in spatial domain. On the other side, since the traditional EM used to train the PLSA model is sensitive to its

initialization, a rival penalized competitive learning based method is leveraged to overcome this deficiency and to provide a good initial estimate of the EPLSA. Extensive experiments on the standard Corel5k dataset validate the effectiveness and efficiency of the proposed model.

3. The Proposed EPLSA. In this section, the proposed EPLSA will be elaborated from two aspects of the bag-of-visual-words (BoVW) model and the rival penalized competitive learning based method, respectively.

3.1. BoVW model. In the recent past, many PLSA models for automatic image annotation are limited by the scope of the representation. In particular, they failed to fully exploit the contextual information of images and words. Based on this recognition and motivated by the latest research [31], a novel bag-of-visual-words model is constructed by integrating the contextual semantic information among visual words based on the PLSA model. In the meanwhile, the approximation strategy of pseudo-likelihood in MRF is introduced to combine the feature appearance similarity in feature domain and the contextual semantic information in spatial domain. Fig. 1 illustrates the scheme of the built BoVW model. To be specific, each image is first divided into rectangular blocks in the feature domain, followed by the SIFT features of these image blocks are extracted, and the k -means algorithm is used to define visual words for image blocks, which is basically the same as the construction of the traditional bag-of-visual words model. It should be noted that, here, the Euclidean distance is utilized to measure the distance between the image blocks and visual words. On the other hand, as for the spatial correlation of image blocks, it can be estimated by the distribution of image blocks in image space. Note that both image blocks and their corresponding visual words serve as initial values of the model. Subsequently, according to Eq.(1), the contextual semantic co-occurrence relationship between the blocks and their surrounding visual words can be obtained based on the PLSA model. Finally, the Markov random field is employed to integrate the feature appearance similarity in feature domain and the contextual semantic information in spatial domain through its potential functions.

$$P(w_i|w_{N(i)}) = \frac{\exp(\beta \sum_{i \in N(i)} p(w_i, w_j))}{\sum_{w_i} \exp(\beta \sum_{j \in N(i)} p(w_i, w_j))} \quad (1)$$

where β is used to control the intensity of the neighborhood interaction, w_i denotes the image blocks. In addition, the distance function between image blocks and visual words is defined as below,

$$d_m^2(x_i, w_k) = \frac{d^2(x_i, w_k)}{P_G(w_i = k|w_{N(i)})} \quad (2)$$

where $P_G(w_i = k|w_{N(i)})$ denotes the prior probability of x_i belonging to class k under the conditions of neighborhood class label $w_{N(i)}$.

Up to this point, the complete procedure of the BoVW model can be succinctly described as follows.

- S 1. input image blocks $X = x_i$, the maximum iteration number T , threshold ε , and the initial visual words $W = w_u$.
- S 2. calculate the contextual semantic co-occurrence probability $P(w_i|w_{N(i)})$ of image blocks.
- S 3. update the distance of image block and visual word. Note that if z_i denotes the corresponding visual words after image block i updated, then

$$z_i = \underset{1 \leq k \leq M}{\operatorname{argmin}} d_m^2(x_i, w_k) \quad (3)$$

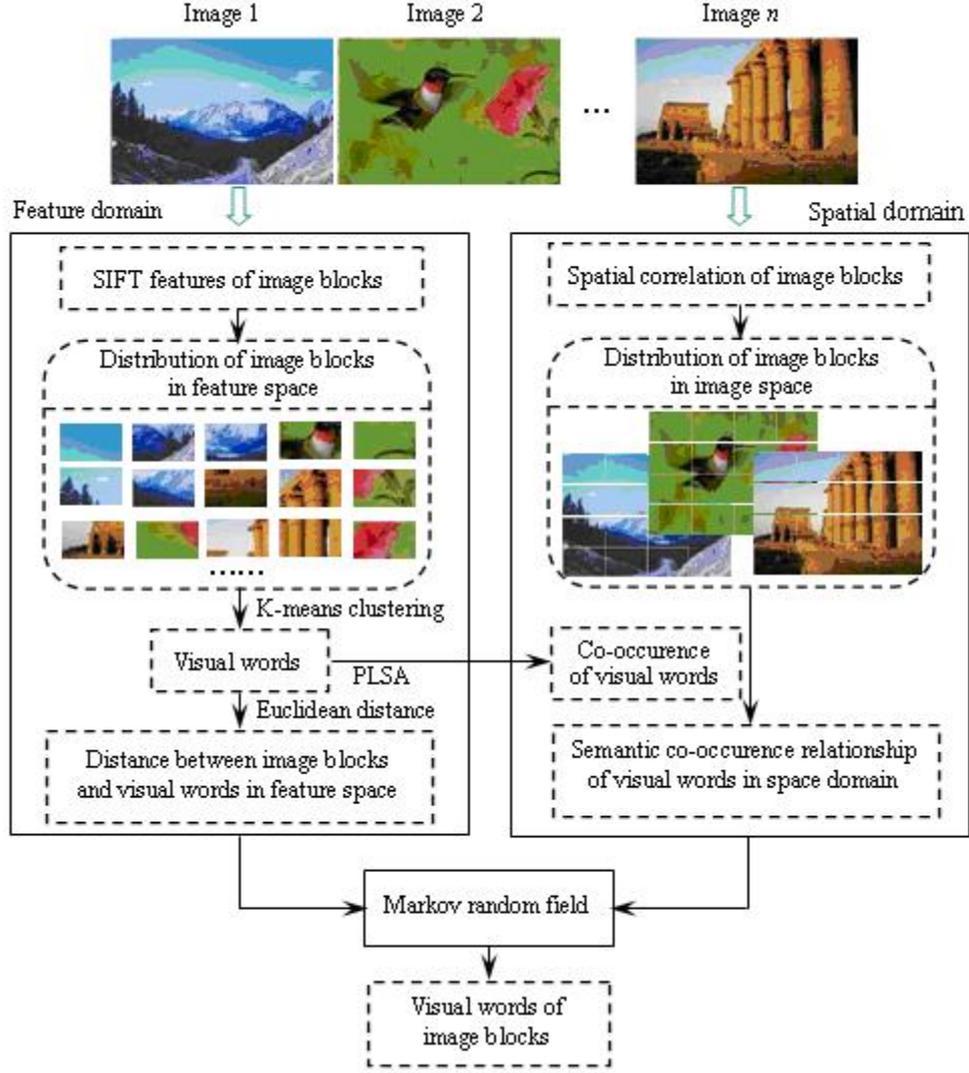


FIGURE 1. Scheme of the constructed BoVW model

It is also worth noting that N_i is set to 0 if z keeps invariant in two consecutive iterations, otherwise set to 1.

S 4. iterate S 3 until $\max_t \|N^{(t)} - N^{(t+1)}\| < \varepsilon$ or $t > T$, then the corresponding visual words of x_i is,

$$z_i = \underset{1 \leq k \leq M}{\operatorname{argmin}} d_m(x_i, w_k) \quad (4)$$

else $t = t + 1$, turn to S 2.

3.2. RPCL algorithm. Probabilistic latent semantic analysis, in brief, is a statistical latent class model that introduces a hidden variable (latent aspect) z_k in the generative process of each element w_i in a document d_j . Given that the unobservable variable z_k , each occurrence w_i is independent of the document it belongs to, which corresponds to the following joint probability based on the Bayesian transformation rule.

$$P(w_i, d_j) = \sum_{k=1}^K P(z_k) P(d_j | z_k) P(w_i | z_k) \quad (5)$$

where $P(z_k)$ is the probability of latent topic variable z_k , $P(d_j | z_k)$ and $P(w_i | z_k)$ denote the probabilities of d_j and w_i occurring in z_k , respectively.

In general, the expectation-maximization algorithm is applied to estimate the parameters of PLSA model through maximizing the log-likelihood function of the observed data. However, since the traditional EM algorithm is sensitive to the initialization, an important consideration for the PLSA model trained via EM is that its performance is strongly affected by the initialization of the model. Following the work [17,20], a rival penalized competitive learning based method is utilized to overcome this deficiency and to provide a good initial estimate of the PLSA model. The core idea behind RPCL is that for each observation, not only the winner cluster center is pulled toward the observation, but also the rival (second winner) one is pushed slightly away from it. Note that the following substitutions are considered to make sure that the model parameters $P(z_k)$, $P(d_j|z_k)$ and $P(w_i|z_k)$ satisfy the probability constraint conditions (i.e., sum to 1) during parameter learning by the RPCL:

$$P(z_k) = e^{\alpha_k} / \sum_{k'=1}^K e^{\alpha_{k'}} \quad (6)$$

$$P(w_i|z_k) = e^{\beta_{i|k}} / \sum_{i'=1}^M e^{\beta_{i'|k}} \quad (7)$$

$$P(d_j|z_k) = e^{\gamma_{j|k}} / \sum_{j'=1}^N e^{\gamma_{j'|k}} \quad (8)$$

Based on the above description, the complete RPCL algorithm for initializing the PLSA model can be summarized as follows:

- S 1. initialize the parameters $\Theta = \{\alpha_k, \beta_{i|k}, \gamma_{j|k}\}_{k=1}^K$ randomly and set $s = 0$ (s denotes the times of updating Θ so far).
- S 2. randomly select an observation pair (w_i, d_j) with $n(w_i, d_j) > 0$, and find the winner topic z_{k_c} and the rival one z_{k_r} :

$$k_c = \underset{k}{\operatorname{argmin}} f_k \operatorname{dis}_k(w_i, d_j) \quad (9)$$

$$k_r = \underset{k \neq k_c}{\operatorname{argmin}} f_k \operatorname{dis}_k(w_i, d_j) \quad (10)$$

note that where f_k denotes the cumulative times of the cluster k being winner, $\operatorname{dis}_k(w_i, d_j) = -n(w_i, d_j) \times \log(P(z_k)P(w_i|z_k)P(d_j|z_k))$ denotes the distance between the observation pair (w_i, d_j) and a latent topic z_k .

- S 3. update the model parameters Θ only for the winner and rival topics:

$$\Theta_{k_c}^{new} = \Theta_{k_c}^{old} - \eta_c \nabla_{\Theta_{k_c}} \operatorname{dis}_{k_c}(w_i, d_j) \quad (11)$$

$$\Theta_{k_r}^{new} = \Theta_{k_r}^{old} + \eta_r \nabla_{\Theta_{k_r}} \operatorname{dis}_{k_r}(w_i, d_j) \quad (12)$$

where $0 \leq \eta_c, \eta_r \leq 1$ are the learning rates for the winner and rival topics respectively, ∇_{Θ_k} denotes the derivative of $\operatorname{dis}_k(w_i, d_j)$.

- S 4. set $s = s + 1$. If $s < T$, go to S 2. Otherwise stop the algorithm. Note that T is a pre-defined maximum times of updating model parameters.

4. Experimental Results and Analysis. In this section, we will first introduce the experimental image dataset and some performance evaluation measures employed in this paper. Afterwards the results of image annotation are reported and analyzed in detail.

4.1. Dataset and evaluation measures. To validate the performance of EPLSA model proposed in this paper, we test it on the Corel5k dataset¹, which is extensively used as basic comparative data for recent research work in image annotation. Corel5k consists of 5,000 images from 50 Corel Stock Photo CD's. Each CD contains 100 images with a certain theme (e.g. polar bears), of which 90 are designated to be in the training set and 10 in the test set, resulting in 4,500 training images and a balanced 500-image test collection. Besides, for the sake of fair comparison, similar features to [6] are extracted. That is, the images are first simply divided into a set of 32×32 -sized blocks, followed by a 36-dimensional feature vector is calculated for each block, consisting of 24 color features (auto-correlogram) computed over 8 quantized colors and 3 Manhattan Distances, 12 texture features (Gabor filter) computed over 3 scales and 4 orientations. As a result, each block is represented as a 36-dimensional feature vector. Finally, each image is represented as a bag of features based on the BoVW constructed in this paper.

Without loss of generality, the most commonly used metrics precision and recall of every word in the test set are calculated and the mean of these values is applied to summarize the model performance. Similar to [7], for a given semantic word, $\text{recall} = B/C$ and $\text{precision} = B/A$, where A is the number of images automatically annotated with a given word in the top 5 returned word list, B is the number of images correctly annotated with that keyword in the top 5 returned word list, and C denotes the number of images having that word in the ground truth annotation. In addition, the top n precision and coverage rate can be formulated as follows to evaluate the performance of automatic image annotation, in which top n precision (denoted by $\text{top}_n\text{-}p(n)$) evaluates the precision of top n ranked annotations for one image whereas top n coverage rate (denoted by $\text{top}_n\text{-}c(n)$) is defined as the percentage of images that are correctly annotated by at least one word among the first n ranked annotations.

$$\text{top}_n\text{-}p(n) = \frac{1}{|T|} \sum_{i \in T} \frac{\text{precision}(i, n)}{n} \quad (13)$$

$$\text{top}_n\text{-}c(n) = \frac{1}{|T|} \sum_{i \in T} \text{coverage}(i, n) \quad (14)$$

where $\text{precision}(i, n)$ is the number of correct annotations in top n ranked annotations for image i , T is the test image set and $|T|$ denotes its size. On the contrary, $\text{coverage}(i, n)$ judges whether image i contains correct annotations in the top n ranked ones. If at least one correct annotation of image i belongs to the top n ranked annotations, then $\text{coverage}(i, n)$ is set to 1, otherwise by 0.

4.2. Results of automatic image annotation. To show the effectiveness of our model EPLSA, we performed thorough experiments on the Corel5k dataset and compared it with several previous approaches [4,5,6,8,34]. 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, in which 'Pre.' is mean precision, 'Rec.' is mean recall and 'Num.' denotes the number of words with non-zero recall values. From Table 1, it can be clearly observed that our model markedly outperforms all the others, especially the first two approaches. Meanwhile, it is also superior to PLSA-WORDS, PLSA-FUSION, CRMR and MBRM by the gains of 16, 12, 5 and 2 words with non-zero recall, 18%, 18%, 13% and 4% mean per-word recall together with 71%, 26%, 9% and 4% mean per-word precision on the set of 260 words, respectively. Similarly, our model can also achieve consistent good performance on the set of 49 best words. This validates the effectiveness

¹http://vision.sista.arizona.edu/kobus/research/data/eccv_2002/index.html

of strategies to sufficiently exploit the contextual semantic information of visual words and employ the rival penalized competitive learning method to provide good initial estimate for the PLSA model. Note that CRMR listed in Table 1 denotes CRM with rectangular regions as input. More details on it can be gleaned from reference [6].

TABLE 1. Performance comparison on Corel5k dataset

Models	CMRM	CRM	PLSA-WORDS	PLSA-FUSION	CRMR	MBRM	EPLSA
Num.	66	107	108	112	119	122	124
Results on 49 best words							
Rec.	0.48	0.70	0.76	0.76	0.75	0.75	0.77
Pre.	0.40	0.59	0.58	0.65	0.72	0.73	0.74
Results on all 260 words							
Rec.	0.09	0.19	0.22	0.22	0.23	0.25	0.26
Pre.	0.10	0.16	0.14	0.19	0.22	0.23	0.24

In addition, we compare our model with several state-of-the-art image annotation methods based on the top n precision and coverage rate respectively, including WNM [32] and RWRM [33]. As illustrated in Fig. 2 (left), it shows the precision of the top n ranked annotations for the image ($n=3,4,5,\dots,9$). Note that the precision decreases gradually with n increasing from 3 to 9, which reflects that the ranking of the words is on average consistent with the true level of accuracy. It is also worth noting that the precision of EPLSA is higher than the corresponding ones of WNM and RWRM respectively. On the other hand, all the coverage rate displayed in Fig. 2 (right) behaves uptrend, and it increases from 51% to 66% for EPLSA, 49% to 65% for RWRM and 39% to 64% for WNM, which further demonstrates the proposed image annotation model, by constructing the novel bag-of-visual-words model and introducing the rival penalized competitive learning algorithm, can efficiently boost the performance of semantic based image annotation.

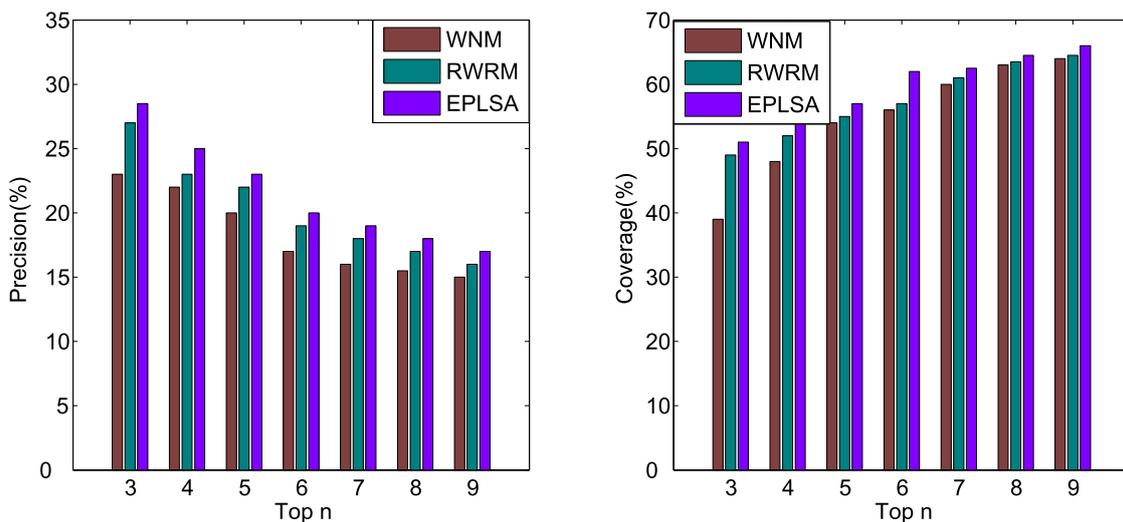


FIGURE 2. Comparison of top n precision (left) and coverage rate (right)

To better understand the effectiveness of the bag-of-visual-words model constructed in this paper and the rival penalized competitive learning based method introduced here, we report experimental results of different PLSA models to illustrate their performance by trial and error. As shown in Table 2, note that the first one is the PLSA model

with traditional bag-of-visual-words model and the traditional EM algorithm (termed as PLSA), the second one is the PLSA model without rival penalized competitive learning based method but with the proposed BoVW model (denoted as PLSA-BoVW), and the third one is just opposite to the second case, that is, PLSA without the proposed BoVW model but with rival penalized competitive learning based method (denoted as PLSA-RPCL). Similarly, the most often used metrics including the number of words with non-zero recall, precision and recall of every word in the test set are calculated to summarize their performance on the set of 260 words. From Table 2, it is clearly observed that the EPLSA model proposed in this paper can get the best annotation performance. In addition, it should be noted that the performance of PLSA-BoVW is slightly better than that of PLSA-RPCL. Of course, all of these models are markedly superior to the basic PLSA. In sum, we can see that the bag-of-visual-words model and the rival penalized competitive learning based method, to some extent, play a complementary role to each other and the combination makes them benefit from each other in the process of automatic image annotation.

TABLE 2. Performance comparison of PLSA, PLSA-RPCL, PLSA-BoVW and EPLSA on Corel5k dataset

Models	PLSA	PLSA-RPCL	PLSA-BoVW	EPLSA
#words with non-zero recall	103	116	119	124
Mean per-word recall	0.18	0.21	0.20	0.26
Mean per-word precision	0.11	0.13	0.15	0.24

To further appreciate the effectiveness of EPLSA for automatic image annotation, Fig. 3 illustrates the performance comparison of precision and recall for PLSA, PLSA-BoVW, PLSA-RPCL and EPLSA models on Corel5k dataset, respectively. Obviously the performance of EPLSA significantly outperforms those of the traditional PLSA, PLSA-BoVW and PLSA-RPCL models. The improvement is largely attributed to that EPLSA adopts the improved bag-of-visual-words model to integrate the contextual semantic information among visual words and the rival penalized competitive learning based method to provide a good initial estimate.

To further illustrate the effect of EPLSA model for automatic image annotation, Fig. 4 shows some examples of annotation (only six cases are listed here due to the limited space) produced by PLSA-WORDS and EPLSA model, 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 literature [8]. Taking the first image in the first row for example, there exist four tags in the original annotation list. However, after annotation by the EPLSA model, its annotation is enriched by the other keyword “wave”, which is very appropriate and reasonable to describe the visual content of the image. On the other side, it is important to note that the annotation ranking of the keywords compared to those generated by PLSA-WORDS is more reasonable, which plays an important role in semantic based image retrieval. This further demonstrates the effectiveness and efficiency of the proposed EPLSA model for the task of automatic image annotation.

5. Conclusions and Future Work. Automatic image annotation has attracted extensive researchers owing to its great potentials in image retrieval, whose goal is to find suitable annotation words to represent the visual content of an untagged or noisily tagged image. In this paper, we have presented an extended probabilistic latent semantic analysis

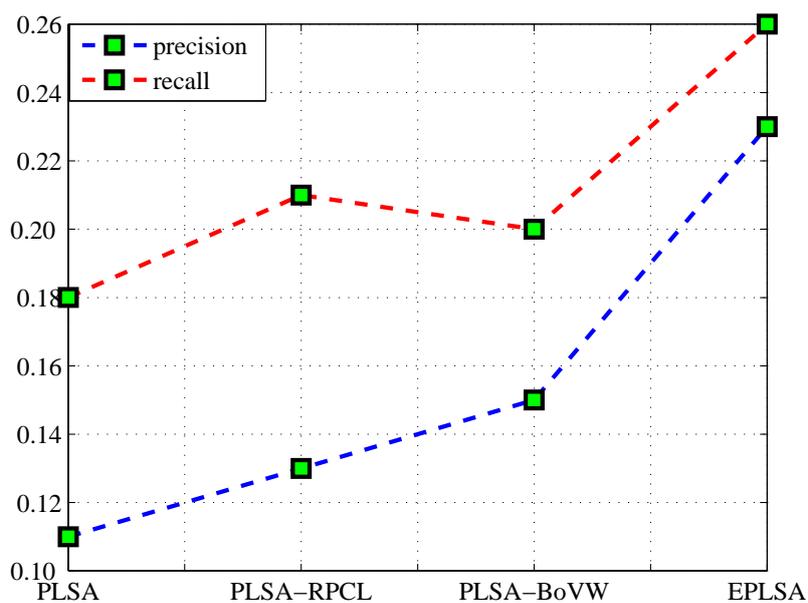


FIGURE 3. Comparison of precision and recall for PLSA, PLSA-RPCL, PLSA-BoVW and EPLSA on Corel5k dataset

Images			
Ground Truth Annotation	beach, people, water, sky	mountain, water, sky, clouds	water, boats, village, harbor
PLSA-WORDS Annotation	sky, beach, snow, sand, mountain	mountain, clouds, boat, coast, hut	water, beach, boats, harbor, skyline
EPLSA Annotation	sky, beach, water, wave, people	mountain, sky, water, clouds, boat	boats, harbor, water, sky, beach
Images			
Ground Truth Annotation	waved, albatross, flight, sky	polar, bear, snow, tundra	zebra, grass, planes, profile
PLSA-WORDS Annotation	city, flight, ceremony, pond, wallow-tailed	polar, bear, tundra, snow, ice	grass, zebra, planes, herd, cat
EPLSA Annotation	bird, flight, sky, waved, albatross	bear, polar, snow, tundra, ice	zebra, grass, planes, herd, trees

FIGURE 4. Annotation comparison with PLSA-WORDS and EPLSA

model for the task of automatic image annotation. Our main contributions are twofold. First, the traditional bag-of-visual-words model is improved by integrating the contextual semantic information among visual words based on the PLSA. At the same time, the approximation strategy of pseudo-likelihood in Markov random field is introduced to combine the feature appearance similarity in feature domain and the contextual semantic information in spatial domain. Second, since the traditional EM used to train the PLSA model is sensitive to its initialization, a rival penalized competitive learning based method is employed to overcome this deficiency and to provide a good initial estimate of the model. Extensive experiments on Corel5k dataset show that the proposed method is significantly more effective than several state-of-the-art methods regarding their effectiveness and efficiency in the task of automatic image annotation.

As for future work, we plan to introduce semi-supervised learning into our approach to leverage the labeled and unlabeled data simultaneously. In the meanwhile, we will work on web image annotation by refining more relevant semantic information from web pages and building more suitable connection between image content features and available semantic information. In addition, due to the latent topics discovered by PLSA model are just based on the regions from images while image segmentation is still an open issue. It is worth noting that inaccurate image segmentation undoubtedly makes the region-based feature representation imprecise and therefore undermine the performance of the PLSA-based approaches. So to explore efficient image segmentation methods is helpful to boost the annotation performance. Furthermore, image segmentation itself is a worthy research direction. Last but not the least, due to the lack of commonly acceptable image databases for PLSA related methods evaluation, which results in the phenomenon that different PLSA approaches make use of different image datasets for their performance evaluation and thus it is difficult to make a fair comparison with each other. So some standard image datasets are expected to be created for researches in the future.

Acknowledgment. The author would like to sincerely thank the anonymous reviewers for their valuable comments and insightful suggestions that have helped to improve the paper. Also, the author thanks Professor Zhongzhi Shi and Hong Hu for stimulating discussions and helpful hints. This work is 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).

REFERENCES

- [1] C. Yang, M. Dong and J. Hua, Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning, *Proc. of the Int'l Conf. on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 2057–2063, 2006.
- [2] G. Carneiro, A. Chan, P. Moreno, et al., Supervised learning of semantic classes for image annotation and retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.
- [3] P. Duygulu, K. Barnard, N. de Freitas, et al., Object recognition as machine translation: learning a lexicon for a fixed image vocabulary, *Proc. of the European Conf. on Computer Vision (ECCV'02)*, pp. 97–112, 2002.
- [4] L. Jeon, V. Lavrenko and R. Manmatha, Automatic image annotation and retrieval using cross-media relevance model, *Proc. of the 26th Int'l Conf. on Research and Development in Information Retrieval (SIGIR'03)*, pp. 119–126, 2003.
- [5] V. Lavrenko, R. Manmatha and J. Jeon, A model for learning the semantics of pictures, *Advances in Neural Information Processing Systems 16 (NIPS'03)*, pp. 553–560, 2003.

- [6] S. Feng, R. Manmatha and V. Lavrenko, Multiple Bernoulli relevance models for image and video annotation, *Proc. of the Int'l Conf. on Computer Vision and Pattern Recognition (CVPR'04)*, pp. 1002–1009, 2004.
- [7] J. Liu, B. Wang, M. Li, et al., Dual cross-media relevance model for image annotation, *Proc. of the 15th Int'l Conf. on Multimedia (MM'07)*, pp. 605–614, 2007.
- [8] F. Monay and D. Gatica-Perez, Modeling semantic aspects for cross-media image indexing, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 10, pp. 1802–1817, 2007.
- [9] D. Tian, X. Zhao and Z. Shi, An efficient refining image annotation technique by combining probabilistic latent semantic analysis and random walk model, *Intelligent Automation & Soft Computing*, vol. 20, no. 3, pp. 335–345, 2014.
- [10] N. Rasiwasia and N. Vasconcelos, Latent Dirichlet allocation models for image classification, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2665–2679, 2013.
- [11] Z. Qian, P. Zhong and R. Wang, Class-specific Gaussian-multinomial latent Dirichlet allocation for image annotation, *EURASIP Journal on Advances in Signal Processing*, vol. 40, no. 1, pp. 1–13, 2015.
- [12] D. Blei and J. Lafferty, Correlated topic models, *Annals of Applied Statistics*, vol. 1, no. 1, pp. 17–35, 2007.
- [13] R. Zhang, L. Guan, L. Zhang, et al., Multi-feature PLSA for combining visual features in image annotation, *Proc. of the 19th Int'l Conf. on Multimedia (MM'11)*, pp. 1513–1516, 2011.
- [14] Q. Guo, N. Li, Y. Yang, et al., Integrating image segmentation and annotation using supervised PLSA, *Proc. of the 20th Int'l Conf. on Image Processing (ICIP'13)*, pp. 3800–3804, 2013.
- [15] I. Sayad, J. Martinet, T. Urruty, et al., Toward a higher-level visual representation for content-based image retrieval, *Multimedia Tools and Applications*, vol. 60, no. 2, pp. 455–482, 2012.
- [16] S. Nikolopoulos, S. Zafeiriou, I. Patras, et al., High order PLSA for indexing tagged images, *Signal Processing*, vol. 93, no. 8, pp. 2212–2228, 2013.
- [17] Z. Lu, Y. Peng and H. Horace, Image categorization via robust PLSA, *Pattern Recognition Letters*, vol. 31, no. 1, pp. 36–43, 2010.
- [18] B. Jin, W. Hu and H. Wang, Image classification based on PLSA fusing spatial relationships between topics, *IEEE Signal Processing Letters*, vol. 19, no. 3, pp. 151–154, 2012.
- [19] Y. Jiang, J. Liu, Z. Li, et al., Co-regularized PLSA for multi-view clustering, *Proc. of the 11th Asian Conf. on Computer Vision (ACCV'12)*, pp. 202–213, 2012.
- [20] A. Farahat and F. Chen, Improving probabilistic latent semantic analysis with principal component analysis, *Proc. of the 11th Conf. of European Chapter of the Association for Computational Linguistics (EACL'06)*, pp. 105–112, 2006.
- [21] E. Rodner and J. Denzler, Randomized probabilistic latent semantic analysis for scene recognition, *Proc. of the 14th Iberoamerican Conf. on Pattern Recognition: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (CIARP'09)*, pp. 945–953, 2009.
- [22] H. Wu, Y. Liu and M. Ye, Applying PLSA to region-based image categorization with soft vector quantization, *Proc. of the 1st Int'l Conf. on Internet Multimedia Computing and Service (ICIMCS'09)*, pp. 102–106, 2009.
- [23] Z. Wang, H. Yi, J. Wang, et al., Hierarchical Gaussian mixture model for image annotation via PLSA, *Proc. of the 5th Int'l Conf. on Image and Graphics (ICIG'09)*, pp. 384–389, 2009.
- [24] S. Romberg, R. Lienhart and E. Horster, Multimodal image retrieval: fusing modalities with multilayer multimodal PLSA, *International Journal of Multimedia Information Retrieval*, vol. 1, no. 1, pp. 31–44, 2012.
- [25] P. Li, J. Cheng, Z. Li, et al., Correlated PLSA for image clustering, *Proc. of the 17th Int'l Conf. on Multimedia Modeling (MMM'11)*, pp. 307–316, 2011.
- [26] L. Zhuang, K. Tang, N. Yu, et al., Unsupervised object learning with am-PLSA, *Proc. of the WRI World Congress on Computer Science and Information Engineering (CSIE'09)*, pp. 701–704, 2009.
- [27] S. Huang and L. Jin, A PLSA-based semantic bag generator with application to natural scene classification under multi-instance multi-label learning framework, *Proc. of the 5th Int'l Conf. on Image and Graphics (ICIG'09)*, pp. 331–335, 2009.
- [28] E. Ergul and N. Arica, Scene classification using spatial pyramid of latent topics, *Proc. of the 20th Int'l Conf. on Pattern Recognition (ICPR'10)*, pp. 3603–3606, 2010.
- [29] Y. Zheng, T. Takiguchi and Y. Arikawa, Image annotation with concept level feature using PLSA + CCA, *Proc. of the 17th Int'l Conf. on Multimedia Modeling (MMM'11)*, pp. 454–464, 2011.

- [30] G. Cheng, L. Guo, T. Zhao, et al., Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and PLSA, *International Journal of Remote Sensing*, vol. 34, no. 1, pp. 45–59, 2013.
- [31] Y. Liu, D. Xu, S. Feng, et al., A novel visual words definition algorithm of image patch based on contextual semantic information, *Acta Electronic Sinica*, vol. 38, no. 5, pp. 1156–1161, 2010.
- [32] Y. Jin, L. Khan, L. Wang, et al., Image annotations by combining multiple evidence and wordnet, *Proc. of the 13th Int'l Conf. on Multimedia (MM'05)*, pp. 706–715, 2005.
- [33] C. Wang, F. Jing, L. Zhang, et al., Image annotation refinement using random walk with restarts, *Proc. of the 14th Int'l Conf. on Multimedia (MM'06)*, pp. 647–650, 2006.
- [34] Z. Li, Z. Shi, X. Liu, et al., Fusing semantic aspects for image annotation and retrieval, *Journal of Visual Communication and Image Representation*, vol. 21, no. 8, pp. 798–805, 2010.